Final project for the Course High Performance Programming, Philipp Noel von Bachmann

March 20, 2022

1 Introduction

Machine Learning and Data analysis are some of the most influencial fields in our current society. One of the main task in Machine Learning is to give a prediction based on some input variables. A common algorithm to use is linear-least-square, which finds the best fit for a linear model. Here, we need an efficient way to calculate this fit. In Data analysis, a common tool is the Principal Component analysis, which tries to reduce the dimensionality of the data in a way, that the reconstruction error gets minimized. PCA relies on an eigenvalue decomposition. For both methods, we can use the QR decomposition too speed up the computation.

2 Problem description

Suppose $A \in \mathbb{R}^{m \times n}$ is a real, square matrix. Then it can be shown that A can be decomposed into

$$A = QR \tag{1}$$

where Q is orthogonal and R is upper triangular. Finding $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{m \times n}$ is the task of the algorithm.

3 Solution

We will implement an algorithm known as **Givens Rotation**. This algorithm relies on construction a sequences of matrices G_i, such that when multiplying A with G_i, we get a new matrix with a zero at a predefined place. Choosing G_i such that we eliminate the lower diagonal of A, we end up with a R and by multiplying all G_i with Q. We will first show how to eliminate one value at the time by constructing G_i and then how to combine these.

3.1 Givens rotation

First, we define a Givens rotation matrix as

$$G(i,j,\theta) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c_{ii} & \cdots & -s_{ij} & & \\ & & \vdots & \ddots & \vdots & & \\ & & s_{ji} & \cdots & c_{jj} & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \tag{2}$$

where $c = \cos(\theta)$, $s = \sin(\theta)$ and any not filled out values are 0. Note: Therefore we can equally represent G by $G(i,j,c,s)$

3.2 Eliminating one value

We will how to find θ to solve

$$G(i,j,\theta)^T \begin{bmatrix} \times \\ \vdots \\ a \\ \vdots \\ b \\ \vdots \\ \times \end{bmatrix} = \begin{bmatrix} \times \\ \vdots \\ r \\ \vdots \\ 0 \\ \vdots \\ \times \end{bmatrix} \tag{3}$$

where $\times$ are arbitrary numbers $a, b \in \mathbb{R}$, $r = \sqrt{a^2 + b^2}$. A trivial solution would be

$$c = \frac{a}{r} \tag{4}$$

$$s = \frac{b}{r} \tag{5}$$

However, r is prone to overflow, so we can instead store it in a different way. If

- $|b| \geq |a|$:
  $t = \frac{a}{b}$, $s = \frac{sgn(b)}{\sqrt{1+t^2}}, c = st$
- else:
  $t = \frac{b}{a}$, $c = \frac{sgn(a)}{\sqrt{1+t^2}}, s = ct$

3.3 Order of elimination

Let us assume that we have $R = \begin{bmatrix} x & x & x \\ a = 0 & x & x \\ b = 0 & c & x \end{bmatrix}$, which we get after finishing the first column in algorithm 1. The next steps would be to set $c = 0$. This in turn modifies row 1,2 in R afterwards, which means a and b necessarily need to be 0 in order to not destroy the solution. Therefore, if we want to set $R_{i,j}$ to 0, then we need to make sure that $R_{i,:j}$ and $R_{i-1,:j}$ are already 0.

3.4 Final algorithm

Algorithm 1 QR factorization with Givens rotation

```
procedure FACTORIZE(A)
    set R = A, Q = I
    for j in 1 to n do
        for i in m down to j + 1 do
            compute the Givens rotation G(i, i−1, c, s) eliminating R_ij with a = R_{i−1,j},
    b = R_{i−1,j}.
            set R = G(i, j, c, s)R, Q = QG(i, j, c, s)
```

Note that we have to do it in this order, since each update affects row and row above.

4 Experiments

4.1 Check for correctness

4.2 Results

4.2.1 Format

In the following section, we will show timing results for different matrix sizes. We denote the sizes by i, j, which we defined as decomposing the matrix $A \in \mathbb{R}^{i \times j}$.

4.2.2 Baseline Results

i \ j	10	25	50
10	0.0005	0.0008	0.001
25	0.018	0.034	0.049
50	0.235	0.611	1.076

We see that the if we vary i, we get a superlinear increase of runtime. If we increase j instead, we see that we get a nearly linear increase in runtime.

(We should only get a minor increase in runtime for j, since we dont need to calculate new 0s). (For i this makes sense, since increasing i by one adds i new 0s to compute.)

4.3 Optimizing the matmul

If we take a closer look at the Givens rotation matrices, we see that they are very sparse. Additionally, we know in advance which entries of the givens rotation wil be nonzero. Thus we can optimize our matmul. First of all, we can only recalculate columns that will get modified by the multiplication. If we multiply GR, where $G = G(i, j, c, s)$, then we only need to modify rows i, j in R. Secondly, we know that for calculating those rows we only need to muliply by the c, s in the corresponding row of G, since all other values are 0.

We need further distinguish between left- (GR) and right-hand (QG) multiplication. We show the pseudo-code for left-hand, but right-hand is very similar.

Algorithm 2 optimized left-hand Matmul with Givens

1: **procedure** MATMUL($G = G(i, j, c, s)$, R)
2: copy the two rows i,j of R into temporary array $T = \begin{bmatrix} R_{i,:} \\ R_{j,:} \end{bmatrix}$
3: **for** k in columns of R **do**
4: $R_{i,k} = c \cdot T_{1,k} - s \cdot T_{2,k}$
5: $R_{j,k} = s \cdot T_{1,k} + c \cdot T_{2,k}$
return R

This leads to the following results:

i \ j	10	25	50
50	0.0003	0.0007	0.0011
500	0.019	0.040	0.079

We see that for $i = 50, j = 50$ we get a runtime of roughly $\frac{1}{1000}$ of the original runtime. However these new runtimes vary quite a bit and we would need to scale them up to get more stable, which is not feasible for the baseline.

4.4 Compiler flags

For i,j=500	Baseline	-O1	-O2	-O3	-03 -march=native -ffast-math
	0.747	0.167	0.137	0.134	0.119

4.5 Further serial optimizations

Tried those further optimizations

- Write pow calls out (compiler flag probably does it already)
- Exchanged div in computation of c, s by multiplication with $\frac{1}{r}$
- Declare the array sizes as constant

Those ones helped:

- Inline functions, took time down from 1.09 to 0.99 for i,j=1000

4.6 Cache blocking

Doesnt really make sense, since we dont have any matmul anymore. For loading of q and r, they are already optimal since we take the two closest cchelines for our mul

- optimize the representation of G, thus also optimizing matmul of G
- cache usage (mayby useful?)

4.7 serial code optimization

- maybe profile code
- DONE use compiler flags, best gcc -O3 -march=native -ffast-math

- DONE look if we can reduce arithmetic, eg no div in loop
- DONE inline functions
- loop unrolling
- no if statements
- change pow calls
- cache blocking
- make constants, especially i and j
- pure functions
- uautovectorization (included in 03)

4.8 Parallelization

4.8.1 Theory and Implementation

In order to parallelize our code, we first need to check which parts can be done in parallel. First, note that the computation of the Givens rotation in line 5 of algorithm 1 is a pure function, eg just depends on the inputs and doesn't have any side effects. Thus, we only need to focus on updating Q, R in line 6. Here, it is necessary to distinguish in two cases:

- For left-hand Givens rotation GR, if $G = G(i, j, c, s)$, only rows i, j of R get modified
-
- For righ-hand Givens rotation QG, if $G = G(i, j, c, s)$, only columns i, j of Q get modified

Consequently, two update steps rotating around $G(i_1, j_1, c_1, s_1)$ and $G(i_2, j_2, c_2, s_2)$ are independent, if i_1, i_2, j_1, j_2 are pairwise distinct. This means we need to make sure that if we set $R_{i,j}$ to 0 in one iteration, we can set any other elements in Row i, j to 0 in the same iteration, in order to not interfere between the computations.

In combination with the constrain that if we want to set $R_{i,j}$ to 0, $R_{i,:j}$ and $R_{i-1,:j}$ already need to be 0, (see Section 3.3) we arrive at the following scheme, shown for a 5×5 matrix:

$$\begin{bmatrix} x & x & x & x & x \\ 3 & x & x & x & x \\ 2 & 4 & x & x & x \\ 1 & 3 & 5 & x & x \\ 0 & 2 & 4 & 6 & x \end{bmatrix} \tag{6}$$

where the values denote in which iteration we can set the element to 0 without cross-interference. This shows that we can set $R_{2,0}$ and $R_{4,1}$ to 0 in parallel for example.

We arrive at the parallel algorithm:

Algorithm 3 parallel QR factorization

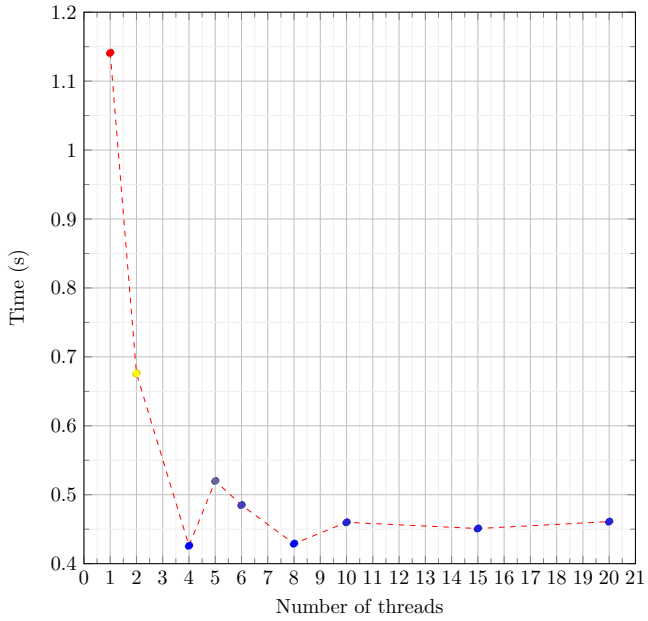
1: **procedure** FACTORIZE_PARALLEL($A \in \mathbb{R}^{m \times n}$)
2: set $R = A$, $Q = I$
3: **for** it in 1 to $(n-1) * 2 + m - n$ **do**
4: In parallel:
5: **for** i, j given by the iteration scheme 6 **do**
6: compute the Givens rotation $G(i, i-1, c, s)$ eliminating R_{ij}
7: set $R = G(i, j, c, s)R$, $Q = QG(i, j, c, s)$

The code was getting parallelized with **OpenMP**. Since we have equal load in line 6+7 of algorithm 3, we don't need to use load balancing.

Looking at 6, we cannot expect a linear scaling with the number of threads. For example in iteration 0 and 1, we only have 1 element so there is no speedup by using more threads. More formally if $A \in \mathbb{R}^{m \times n}$, then we can have a maximum of $\min(m/2, n)$ computations running in parallel, achieved in the itertation where we set $R_{1,0}$ to 0.

4.8.2 Results

Because the computation of i, j is a bit more involved now, we get a slightly higher baseline with 1.137s compared to 0.986s for our serial code and $i, j = 1000$. However when increasing the number of threads:



	10	100	1000	2000	4000
$j = 1000, i =$	0.031	0.023	0.425	4.125	58.866
$i = 1000, j =$	0.031	0.087	0.430	1.084	2.247

5 Memory optimization

Maybe try to store Q in column format and R in row format??

6 Random

- we can order them by affecting row and above
- best is probably to create tasks because upper rows require less work
- maybe need memory optimizations, but most of the operations seem to be near anyway

TODO:

- implement checks
- implement good storage of G
- implement efficient matmul of G
- implement one givens rotation
- implement outer loop
- serial optimizations
- parallelizations, probably with openmp

7 Experiments

8 Conclusion

9 References

$https://www.math.usm.edu/lambers/mat610/sum10/lecture9.pdf https://www.math.usm.edu/lambers$
$//en.wikipedia.org/wiki/Givens_rotation$