# LINEAR ALGEBRA: ANSWER TO HOMEWORK 4

VISHVAS VASUKI

## 1. 20.4

The Gaussian elimination algorithm:

```
Input: A
Output: L, U
U = A
L = I
foreach $k$ = 1 to m-1 do
    foreach $j$ = k+1 to m do
        $l_{j,k} = u_{j,k}/u_{k,k}$
        $u_{j,k:m} = u_{j,k:m} - l_{j,k}u_{k,k:m}$
    end
end
```

This code can be rewritten using just one explicit for loop indexed by k. Inside this loop, U will be updated at each step by a certain rank one matrix.

```
Input: A
Output: L, U
U = A
L = I
foreach $k$ = 1 to m-1 do
    $l_k = 0(u_{1:m,k})$
    $l_{k+1:m,k} = l_k + \frac{u_{k+1:m,k}}{u_{k,k}}$
    $L_k = I - l_k e_k^*$
    $U = L_k U$
end
```

## 2. 20.5

Gaussian elimination yields A=LU.

2.1. **a.** Scenario: Elimination by columns from left to right, rather than by rows from top to bottom, so that A is made lower triangular.

Yields LU factorization.

Reason: These Column operations can be represented by right multiplication by upper triangular matreces.

2.2. **b.** Scenario: Gaussian elimination applied after preliminary scaling of columns of A by multiplication with diagonal matrix D.

So, if $DA = LU$; $A = D^{-1}LU$.

So, Ax=b is $Ax = D^{-1}LUx = b$; so LUx = Db. So, b is scaled.

2.3. **c.** Scenario: Gaussian elimination carried further, so that A, assumed non singular, is reduced to upper triangular and thence to diagonal.

Yields the $A = LDU$ factorization; where L and U are unit triangular.

## 3

*Remark* 3.0.1. Below the inequalities such as the triangle inequality and the Cauchy Schwartz inequality, and the inequality due to the definition of the induced matrix norm are used without being explicitly mentioned each time.

### 3.1.

*Notation.* $\epsilon$ denotes $\epsilon_{machine}$.

**Theorem 3.1.1.** *(Forward-error analysis)*

$$|fl(x^T a) - x^T a| \leq n\epsilon|x|^T|a| + O(\epsilon^2),$$

*where $x, a$ are $n$-dimensional floating point vectors and $fl(x^T a)$ represents floating point computation of dot product between $x$ and $a$.*

*Proof.*

*Remark* 3.1.2. In the calculations below, we ignore the $\epsilon^2$ terms, as doing so does not affect the soundness of the proof.

Proof by induction.

$$
\begin{aligned}
fl(x_ia_i) &\leq x_ia_i(1+\epsilon)\\
|fl(x_ia_i)-x_ia_i| &\leq \epsilon|x_ia_i|\\
&\phantom{\leq}\ \texttt{Assume that:}\\
|fl(\sum_{i=1}^{m}x_ia_i)-(\sum_{i=1}^{m}x_ia_i)| &\leq m\epsilon\sum_{i=1}^{m}|x_ia_i|+O(\epsilon^2)\\
fl(\sum_{i=1}^{m+1}x_ia_i) &\leq (fl(\sum_{i=1}^{m}x_ia_i))+x_{i+1}a_{i+1}(1+\epsilon))(1+\epsilon)\\
&\leq fl(\sum_{i=1}^{m}x_ia_i)+x_{i+1}a_{i+1}(1+\epsilon)+\epsilon(fl(\sum_{i=1}^{m}x_ia_i))\\
&\phantom{\leq}\ +O(\epsilon^2)
\end{aligned}
$$

$$
\begin{aligned}
|fl(\sum_{i=1}^{m+1}x_ia_i)-(\sum_{i=1}^{m+1}x_ia_i)| &\leq |fl(\sum_{i=1}^{m}x_ia_i)-(\sum_{i=1}^{m}x_ia_i)|+\epsilon(|fl(\sum_{i=1}^{m}x_ia_i|)\\
&\phantom{\leq}\ +O(\epsilon^2)+\epsilon|x_{i+1}a_{i+1}|\\
&\leq m\epsilon\sum_{i=1}^{m}|x_ia_i|+O(\epsilon^2)+\epsilon(|fl(\sum_{i=1}^{m}x_ia_i)|)\\
&\phantom{\leq}\ +\epsilon|x_{i+1}a_{i+1}|\\
&\leq (m)\epsilon\sum_{i=1}^{m}|x_ia_i|+\epsilon\sum_{i=1}^{m+1}|x_ia_i|+O(\epsilon^2)\\
&\leq (m+1)\epsilon\sum_{i=1}^{m+1}|x_ia_i|+O(\epsilon^2)
\end{aligned}
$$

□

### 3.2.

**Theorem 3.2.1.** *(Forward-error analysis)*

$$\|fl(XA)-XA\|_F \leq n\epsilon_{machine}\|X\|_F\|A\|_F+O(\epsilon_{machine}^2),$$

*where $X, A$ are $n\times n$ dimensional floating point matrices and $fl(XA)$ represents floating point computation of matrix multiplication between $X$ and $A$ using dot-products.*

*Proof.*

*Notation.* Let M = fl(XA)-XA. $x_i^*$ represents ith row of X.

$$\begin{aligned}
|M_{i,j}| &= |fl(x_i^* a_j) - x_i^* a_j| \\
&\leq n\epsilon |x_i|^T |a_j| + O(\epsilon^2) \texttt{ From earlier thm} \\
M_{i,j}^2 &\leq n^2\epsilon^2 (x_i^* a_j)^2 + O(\epsilon^3) + O(\epsilon^4) \\
\sum M_{i,j}^2 &\leq \sum n^2\epsilon^2 (x_i^* a_j)^2 + O(\epsilon^3) + O(\epsilon^4) \\
\|M\|_F^2 &\leq n^2\epsilon^2 \|XA\|_F^2 + O(\epsilon^3) + O(\epsilon^4) \\
\therefore \|M\|_F &\leq n\epsilon \|XA\|_F + O(\epsilon^2) \\
\therefore \|M\|_F &\leq n\epsilon \|X\|_F \|A\|_F + O(\epsilon^2)
\end{aligned}$$

□

**3.3.**

**Theorem 3.3.1.** *(Backward-error analysis) Suppose* $fl(XA) = X(A + \delta A)$ *and* $k(X) = \|X\|_F \left\|X^{-1}\right\|_F$*. The relative backward error* $\frac{\|\delta A\|_F}{\|A\|_F} \leq k(X)O(\epsilon_{machine})$.

*Proof.*

$$\begin{aligned}
fl(XA) &= XA + X\delta A \\
\|X\delta A\|_F &= \|fl(XA) - XA\|_F \\
&\leq n\epsilon \|X\|_F \|A\|_F + O(\epsilon^2) \\
\frac{\|X\delta A\|_F}{\|A\|_F} &\leq n\epsilon \|X\|_F + O(\epsilon^2) \\
\frac{\left\|X^{-1}\right\|_F \|X\delta A\|_F}{\|A\|_F} &\leq n\epsilon \|X\|_F \left\|X^{-1}\right\|_F + O(\epsilon^2) \\
\therefore \frac{\left\|X^{-1}X\delta A\right\|_F}{\|A\|_F} &\leq n\epsilon \|X\|_F \left\|X^{-1}\right\|_F + O(\epsilon^2) \\
\therefore \frac{\|\delta A\|_F}{\|A\|_F} &\leq O(\epsilon)k(X)
\end{aligned}$$

□

## 4

Let $x$ be the solution of $Ax = b$, where $A$ is square and invertible. Carry out the perturbation analysis when *both* the matrix $A$ and the vector $b$ is perturbed.

Let $\tilde{x} = x + \delta x$ such that $(A + \delta A)\tilde{x} = b + \delta b$.

**Theorem 4.0.2.**

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right),$$

*provided that* $\delta A$ *is sufficiently small, in our case assume that* $\left\|A^{-1}\right\| \|\delta A\| < 1$. *The matrix norm is the induced norm obtained from the vector norm used and* $\kappa(A) = \|A\| \left\|A^{-1}\right\|$.

*Proof.*

*Remark* 4.0.3. Below the inequalities such as the triangle inequality and the Cauchy Schwartz inequality, and the inequality due to the definition of the induced matrix norm are used without being explicitly mentioned each time.

$$\begin{aligned}
(A+\delta A)(x+\delta x) &= b+\delta b \\
\delta x &= A^{-1}(\delta b-\delta Ax)-A^{-1}\delta A\delta x \\
\|\delta x\| &\leq \left\|A^{-1}\right\|(\|\delta b\|+\|\delta Ax\|)+\left\|A^{-1}\delta A\delta x\right\| \\
&\leq \left\|A^{-1}\right\|\|A\|\,(\frac{\|\delta b\|}{\|A\|}+\frac{\|\delta Ax\|}{\|A\|})+\left\|A^{-1}\right\|\|\delta A\|\,\|\delta x\| \\
(1-k(A)\frac{\|\delta A\|}{\|A^{-1}\|})\,\|\delta x\| &\leq k(A)(\frac{\|\delta b\|}{\|A\|}+\frac{\|\delta Ax\|}{\|A\|}) \\
\frac{\|\delta x\|}{\|x\|} &\leq \frac{k(A)}{(1-k(A)\frac{\|\delta A\|}{\|A^{-1}\|})}(\frac{\|\delta b\|}{\|A\|\,\|x\|}+\frac{\|\delta Ax\|}{\|A\|\,\|x\|}) \\
&\leq \frac{k(A)}{(1-k(A)\frac{\|\delta A\|}{\|A^{-1}\|})}(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|})
\end{aligned}$$

□