Assignment 2: Norms, Inner Products, and Linear System Solving

Gregory Smetana
ID 1917370
ACM 106a

October 23, 2013

1 Equivalence of norms

a)

To prove that all vector norms on $\mathbf{R}^n$ and $\mathbf{C}^n$ are equivalent, it is sufficient to show that all norms are equivalent to $\|\cdot\|_2$. This is true because norm equivalence is transitive

$$m_a\|x\|_2 \leq \|x\|_a \leq M_a\|x\|_2 \tag{1}$$

$$m_b\|x\|_2 \leq \|x\|_b \leq M_b\|x\|_2 \tag{2}$$

$$\frac{m_b}{M_a}\|x\|_a \leq \|x\|_b \leq \frac{M_b}{m_a}\|x\|_a \tag{3}$$

The inequality may be divided by $\|x\|_2$, so it is sufficient to consider only the case with $\|x\|_2 = 1$:

$$m\|x\|_2 \leq \|x\|_a \leq M\|x\|_2 \tag{4}$$

$$m \leq \|x'\|_a \leq M \tag{5}$$

Now, we want to show that any norm is a continuous function. This means that as $x \to y$, we need $|\|x\|_a - \|y\|_a| \to 0$. As x converges to y, we have

$$\|x - y\|_a < \epsilon \tag{6}$$

Considering a variation of the triangle inequality

$$\|x - y\|_a \geq \|x\|_a - \|y\|_a \tag{7}$$

Thus,

$$|\|x\|_a - \|y\|_a| < \epsilon \tag{8}$$

and the norm is continuous.

Therefore, by the extreme value theorem, the norm $\|x\|_a$ has a minimum and maximum on the closed disk $\|x\|_2 = 1$. We may write

$$m_a\|x\|_2 \leq \|x\|_a \leq M_a\|x\|_2 \tag{9}$$

and it follows from the transitive property of norm equivalence that there exist constants m and M independent of x, such that

$$\boxed{m\|x\|_a \leq \|x\|_b \leq M\|x\|_a} \tag{10}$$

b)

$$c_1\|x\|_\infty \leq \|x\|_2 \tag{11}$$

$$c_1 \text{max}_i|x_i| \leq \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \tag{12}$$

Setting $\|x\|_2 = 1$, and maximizing the left side with $x = (1, 0, 0, ..., 0)$, we determine $c_1 = 1$

$$\|x\|_2 \leq c_2\|x\|_\infty \tag{13}$$

$$\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \leq c_2 \text{max}_i|x_i| \tag{14}$$

Setting $\|x\|_2 = 1$, and minimizing the right side with $x = (1/\sqrt{n}, 1/\sqrt{n}, 1/\sqrt{n}, ..., 1/\sqrt{n})$, we determine $c_2 = \sqrt{n}$. Therefore,

$$\boxed{\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty} \tag{15}$$

$$c_3\|x\|_2 \leq \|x\|_1 \tag{16}$$

$$c_3\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \leq |x_1| + |x_2| + \ldots + |x_n| \tag{17}$$

Setting $\|x\|_2 = 1$, and minimizing the right side with $x = (1, 0, 0, ..., 0)$, we determine $c_3 = 1$

$$\|x\|_1 \leq c_4\|x\|_2 \tag{18}$$

$$|x_1| + |x_2| + \ldots + |x_n| \leq c_4\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \tag{19}$$

Setting $\|x\|_2 = 1$, and maximizing the left side with $x = (1/\sqrt{n}, 1/\sqrt{n}, 1/\sqrt{n}, ..., 1/\sqrt{n})$, we determine $c_4 = \sqrt{n}$. Therefore,

$$\boxed{\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty} \tag{20}$$

2 Operator norm of the inverse

The matrix operator norm induced by the vector norm $\|\cdot\|$ is given by

$$\|A\| = \max_{x\neq 0}\frac{\|Ax\|}{\|x\|} \tag{21}$$

Applying to the inverse,

$$\|A^{-1}\| = \max_{y\neq 0}\frac{\|A^{-1}y\|}{\|y\|} \tag{22}$$

Since $Ax = y$ and $x = A^{-1}y$

$$\|A^{-1}\| = \max_{x\neq 0}\frac{\|x\|}{\|Ax\|} \tag{23}$$

Therefore

$$\boxed{\|A^{-1}\|^{-1} = \frac{1}{\|A^{-1}\|} = \min_{y\neq 0}\frac{\|Ay\|}{\|y\|}} \tag{24}$$

3 Floating point dot and matrix products

Fundamental axiom of floating point arithmetic:

$$fl(x \odot y) = (x \odot y)(1+\delta) \tag{25}$$

for some $\delta \in [-\epsilon_M, \epsilon_M]$

a)

$$fl(x^T y) = fl\left(\sum_{i=1}^{d} x_i y_i\right) \tag{26}$$

Each product in Equation 26 incurs an error according to Equation 25

$$fl(x^T y) = x_1y_1(1+\delta_1) +_f x_2y_2(1+\delta_2) +_f x_3y_3(1+\delta_3) +_f \ldots +_f x_dy_d(1+\delta_d) \tag{27}$$

where $+_f$ indicates floating point addition. Applying the formula to the sum,

$$fl(x^T y) = (x_1y_1(1+\delta_1) + x_2y_2(1+\delta_2))\,(1+\delta^1) +_f x_3y_3(1+\delta_3) +_f \ldots +_f x_dy_d(1+\delta_d) \tag{28}$$

$$\begin{aligned} fl(x^T y) = {} & x_1y_1(1+\delta_1)(1+\delta^1)(1+\delta^2)\ldots(1+\delta^{d-1}) \\ & + x_2y_2(1+\delta_2)(1+\delta^1)(1+\delta^2)\ldots(1+\delta^{d-1}) \\ & + x_3y_3(1+\delta_3)(1+\delta^2)(1+\delta^3)\ldots(1+\delta^{d-1}) \\ & \ldots \\ & + x_dy_d(1+\delta_d)(1+\delta^{d-1}) \end{aligned} \tag{29}$$

Using the inequality $(1+\delta_i)(1+\delta^1)(1+\delta^2)\dots(1+\delta^{d-1}) \le (1+\delta)^d$, with $|\delta| \ge |\delta_i|$ and $|\delta| \ge |\delta^i|$, this may be simplified:

$$fl(x^T y) = \sum_{i=1}^{d} x_i y_i (1+\delta)^d \tag{30}$$

for some $\delta \in [-\epsilon, \epsilon]$. Since $(1+\delta)^d \le (1+d\delta)$, this may be further simplified to

$$\boxed{fl(x^T y) = \sum_{i=1}^{d} x_i y_i (1+\delta_i)} \tag{31}$$

for some $\delta_i \in [-d\epsilon, d\epsilon]$

b)

$$(|fl(AB) - AB|)_{ij} = (|fl(\sum_{k=1}^{n} A_{ik}B_{kj}) - \sum_{k=1}^{n} A_{ik}B_{kj}|)_{ij} \tag{32}$$

Using the earlier result,

$$(|fl(AB) - AB|)_{ij} = (|\sum_{k=1}^{n} A_{ik}B_{kj}(1+\delta_k) - \sum_{k=1}^{n} A_{ik}B_{kj}|)_{ij} = (|\sum_{k=1}^{n} A_{ik}B_{kj}\delta_k|)_{ij} \tag{33}$$

for some $\delta_k \in [-n\epsilon, n\epsilon]$ We may write this as an inequality to show that

$$\boxed{(|fl(AB) - AB|)_{ij} \le n\epsilon(|A||B|)_{ij}} \tag{34}$$

4 Linear systems and rank-one error

a)

If the columns of E span a one-dimensional vector space, its columns may be expressed as multiples of a vector u:

$$E = \begin{pmatrix} \vdots & \vdots & & \vdots \\ c_1 u & c_2 u & \cdots & c_n u \\ \vdots & \vdots & & \vdots \end{pmatrix} \tag{35}$$

The constants c may be expressed as a vector, v, so it follows that the matrix E must have the factorization

$$E = uv^T \tag{36}$$

b)

$$\left(A + uv^T\right)^{-1} = A^{-1} - \sigma A^{-1} u v^T A^{-1} \tag{37}$$

$$I = \left(A + uv^T\right)\left(A^{-1} - \sigma A^{-1} u v^T A^{-1}\right) \tag{38}$$

$$I = I - \sigma u v^T A^{-1} + u v^T A^{-1} - \sigma u v^T A^{-1} u v^T A^{-1} \tag{39}$$

$$0 = \sigma u v^T A^{-1} - u v^T A^{-1} + \sigma u v^T A^{-1} u v^T A^{-1} \tag{40}$$

Labelling scalar quantity $s = v^T A^{-1} u$ and factoring

$$u v^T A^{-1} = \sigma u v^T A^{-1} + s \sigma u v^T A^{-1} \tag{41}$$

$$1 = \sigma + s\sigma \tag{42}$$

$$\sigma = \frac{1}{1+s} \tag{43}$$

$$\boxed{\sigma = \frac{1}{1 + v^T A^{-1} u}} \tag{44}$$

So we have shown that if $v^T A^{-1} u \neq -1$,

$$\boxed{\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}} \tag{45}$$

c)

We wish to solve $\tilde{A}x = b$ using the Sherman-Morrison formula.

$$x = A^{-1} b - \frac{A^{-1} u v^T A^{-1} b}{1 + v^T A^{-1} u} \tag{46}$$

If we say that $Ay = b$ and $Az = u$, this may be written as

$$x = y - z \frac{v^T y}{1 + v^T z} \tag{47}$$

So the (pseudocode) algorithm for solving $\tilde{A}x = b$ is

1. Solve Ay =b for y
2. Solve Az = u for z
3. Compute product $r = v^T y$
4. Compute product $t = v^T z$
5. Compute $x = y - z \frac{r}{1+t}$

If $Ax = b$ can be solved in M flops, the first two steps of the algorithm will require M flops each. The dot product in steps 3 and 4 is only $O(n)$ flops each. Therefore if M satisfies $n = O(M)$ the algorithm requires only $O(M)$ operations.

d)

If A is an orthogonal matrix, in solving for $\tilde{A}x = b$, the linear systems $Ay = b$ and $Az = u$ may be solved easily by multiplying by the transpose. This was done for 25 random orthogonal matrices A, with random vectors u, v, b, for $n = 10, 100, 500$. The results of using the Sherman-Morrison formula are compared with Matlab's backslash operator in Table 1

	n=10	n=100	n=500
$\|\tilde{x} - x\|_2$	3.2406e-16(7.2439e-17)	3.07e-15(3.2974e-16)	1.8382e-14(1.0714e-15)
Time [μs]: my_alg	29.48(35.5377)	66.12(133.4302)	193.8(21.2289)
Time[μs]: $A\backslash b$	14.72(4.9034)	188.36(15.6601)	4977.48(296.7853)

Table 1

The results show that the residual $\|\tilde{x} - x\|_2$, was low for all cases but increased with n due to the round-off error associated with more floating point operations. Matlab's backslash operator was faster for small values of n, but much more rapidly with n than the Sherman-Morrison formula. This is because the Sherman-Morrison formula is a lower order method than the Gaussian elimination with partial pivoting used by the backslash operator. Gaussian elimination is $O(n^3)$ and the matrix multiplication used in orthogonal Sherman-Morrison is only $O(n^2)$

A Smetana_Gregory_1917370_A2_P4_DIARY.txt

```
run('Smetana_Gregory_1917370_A2_P4.m')
n = 10
residual = 3.2406e-16(7.2439e-17)
t_alg = 29.48(35.5377) [microsecond]
t_matlab = 14.72(4.9034) [microsecond]
n = 100
residual = 3.07e-15(3.2974e-16)
t_alg = 66.12(133.4302) [microsecond]
t_matlab = 188.36(15.6601) [microsecond]
n = 500
residual = 1.8382e-14(1.0714e-15)
t_alg = 193.8(21.2289) [microsecond]
t_matlab = 4977.48(296.7853) [microsecond]
diary off
```

B Smetana_Gregory_1917370_A2_P4.m

```
%Smetana_Gregory_191737_A2_P4
clear;
clc;

nn=[10,100,500];
t_alg = zeros(1,25);
t_matlab = zeros(1,25);
residual = zeros(1,25);
for n=nn
for(i=1:25)
%initialize
A = eye(n);
idx = randperm(n);
A=A(idx,:);
b=rand(n,1);
u = rand(n,1);
v = rand(n,1);

% solve using sherman morrison
tic;
xt= sherman_morrison(A,u,v,b);
t_alg(i) = toc;

% solve using backslash
At = A + u*v';
tic;
x = At\b;
t_matlab(i) = toc;

residual(i) = norm(xt-x);
end;
```

```
% output results
disp(['n = ',num2str(n)])
disp(['residual = ',num2str(mean(residual)),...
    '(',num2str(std(residual)),')'])
disp(['t_alg = ',num2str(mean(t_alg*10^6)),...
    '(',num2str(std(t_alg*10^6)),') [microsecond]'])
disp(['t_matlab = ',num2str(mean(t_matlab*10^6)),...
    '(',num2str(std(t_matlab*10^6)),') [microsecond]'])
end;
```

C sherman_morrison.m

```
%Smetana_Gregory_191737_A2_P4

function [ x ] = sherman_morrison( A, u, v, b )
%SHERMAN_MORRISON Solves A_t * x = b if A is an orthogonal matrix and
% A_t = A + u*v'

y = A'*b;
z = A'*u;
r = v'*y;
t = v'*z;
x = y - z*r/(1+t);
end
```