CSC 336 EXAM SOLUTION December 2016.

1. [10 marks; 2 marks for each part]

For each of the five statements below, say whether the statement is **true** or **false** and briefly justify your answer.

(a) A problem that is highly sensitive to small changes in the problem data is poorly conditioned.

True. The definition I gave them in class is that, a problem is poorly conditioned if small changes to the input of the problem can cause large changes to the output of the problem. Saying that a problem us "highly sensitive to small changes in the problem data" is essentially the same thing.

(b) In a floating-point number system, the *underflow level* (i.e., UFL in your textbook) is the largest positive floating-point number δ such that $\mathrm{fl}(1+\delta) = 1$, where $\mathrm{fl}(1+\delta)$ is the floating-point value you get when you compute $1 + \delta$ in this floating-point number system.

False. The underflow level, UFL, is the smallest positive normalized floating-point number. The δ given above is similar to the way machine epsilon, ϵ_{mach}, is sometimes defined. Typically, $0 < \text{UFL} \ll \epsilon_{\text{mach}}$.

(c) Let A be an $n \times n$ nonsingular real matrix. If the condition number of A is very large, then the determinant of A must be close to zero.

False. Consider the example

$$A = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$$

where $a \gg 1$. Clearly

$$A^{-1} = \begin{pmatrix} 1/a & 0 \\ 0 & a \end{pmatrix}$$

So,

$$\text{cond}_\infty(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty = a \cdot a = a^2$$

However, $\det(A) = 1$. Therefore, this is an example for which condition number of A is very large, but the determinant of A is not close to zero.

(d) For a given fixed level of accuracy, a super-linearly convergent iterative method always requires fewer iterations than a linearly convergent method to find a solution to that level of accuracy.

False. Eventually a super-linearly convergent methods will converge faster than a linearly convergent method, but this may not be the case at the start of the iteration. Hence, if the level of accuracy is quite relaxed, a linearly convergent method may reach that level of accuracy before a super-linearly convergent method does.

(e) Given three pairs of points (x_1, y_1), (x_2, y_2), (x_3, y_3), where $x_i \in \mathbb{R}$ for $i = 1, 2, 3$, $y_i \in \mathbb{R}$ for $i = 1, 2, 3$ and the x_i are distinct, it is always possible to find a polynomial $p(x)$ of degree 2 or less such that $p(x_i) = y_i$ for $i = 1, 2, 3$.

True. This is a special case of the theorem I gave them in class that says:

> Given n points (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), where $n \geq 1$, the $x_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$, $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ and the x_i are distinct, there is a unique polynomial $p(x)$ of degree $n - 1$ or less that satisfies $p(x_i) = y_i$ for $i = 1, 2, \ldots, n$.

2. [10 marks: 5 marks for each part]

(a) The formula (1) produces such poor approximations to π when computed using IEEE double-precision floating-point arithmetic because if suffers from *catastrophic cancellation.* To be more specific, note that, when n is large, $p_n/2^n \ll 1$, since $p_n \in [0,4]$. So, $1-(p_n/2^n)^2 \approx 1$, whence $\sqrt{1-(p_n/2^n)^2} \approx 1$ also. Therefore, there is catastrophic cancellation when we compute $1-\sqrt{1-(p_n/2^n)^2}$.
For $n = 29$,

$$(p_{29}/2^{29})^2 \approx (\pi/2^{29})^2 \approx \pi^2/2^{58} \leq 2^4/2^{58} = 2^{-54} < \frac{1}{2}\epsilon_{\text{mach}}$$

since $\epsilon_{\text{mach}} = 2^{-52}$ in IEEE double-precision floating-point arithmetic. Therefore, $\text{fl}(1-(p_{29}/2^{29})^2) = 1$, whence $\text{fl}(1-\sqrt{1-(p_{29}/2^{29})^2}) = 0$. Therefore, from (1) and the results above, we see that $\text{fl}(p_{30}) = 0$. Moreover, it follows immediately from (1), that once $p_{\hat{n}} = 0$ for any $\hat{n}$, then $p_n = 0$ for all $n \geq \hat{n}$.

(b) To find a formula that is mathematically equivalent to formula (1) but does not suffer from the extreme loss of accuracy that we see in the numerical results for formula (1), note that

$$\begin{aligned}
1-\sqrt{1-(p_n/2^n)^2} &= \left(1-\sqrt{1-(p_n/2^n)^2}\right)\frac{1+\sqrt{1-(p_n/2^n)^2}}{1+\sqrt{1-(p_n/2^n)^2}} \\
&= \frac{1-\left(1-(p_n/2^n)^2\right)}{1+\sqrt{1-(p_n/2^n)^2}} \\
&= \frac{(p_n/2^n)^2}{1+\sqrt{1-(p_n/2^n)^2}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
p_{n+1} &= 2^n\sqrt{2\left(1-\sqrt{1-(p_n/2^n)^2}\right)} \\
&= 2^n\sqrt{2\frac{(p_n/2^n)^2}{1+\sqrt{1-(p_n/2^n)^2}}} \\
&= p_n\sqrt{\frac{2}{1+\sqrt{1-(p_n/2^n)^2}}}
\end{aligned}$$

Hence,

$$p_{n+1} = p_n\sqrt{\frac{2}{1+\sqrt{1-(p_n/2^n)^2}}} \qquad (1)$$

Note that formula (1) does not suffer from catastrophic cancellation. If you compute with it, you get very good approximations to π.

3. [20 marks: 5 marks for each part]

(a) For the matrix

$$A = \begin{pmatrix} -1 & -2 & 1 \\ 4 & 4 & -4 \\ 2 & -1 & -5 \end{pmatrix}$$

in the first stage of Gaussian elimination with partial pivoting, we first interchange rows 1 and 2. That is, we multiply A by the permutation matrix

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

to get

$$P_1 A = \begin{pmatrix} 4 & 4 & -4 \\ -1 & -2 & 1 \\ 2 & -1 & -5 \end{pmatrix}$$

Then we eliminate the elements below the main diagonal in the first column of A by multiplying A by

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

to get

$$M_1(P_1 A) = \begin{pmatrix} 4 & 4 & -4 \\ 0 & -1 & 0 \\ 0 & -3 & -3 \end{pmatrix}$$

In the second stage of Gaussian elimination with partial pivoting, we first interchange rows 2 and 3. That is, we multiply $M_1 P_1 A$ by the permutation matrix

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

to get

$$P_2(M_1 P_1 A) = \begin{pmatrix} 4 & 4 & -4 \\ 0 & -3 & -3 \\ 0 & -1 & 0 \end{pmatrix}$$

Then we eliminate the elements below the main diagonal in the second column of $P_2 M_1 P_1 A$ by multiplying $P_2 M_1 P_1 A$ by

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/3 & 1 \end{pmatrix}$$

to get

$$U = M_2(P_2M_1P_1A) = \begin{pmatrix} 4 & 4 & -4 \\ 0 & -3 & -3 \\ 0 & 0 & 1 \end{pmatrix}$$

To get the LU factorization of A from $M_2P_2M_1P_1A = U$, first define

$$\hat{M}_1 = P_2M_1P_2^T = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/4 & 0 & 1 \end{pmatrix}$$

Then note that, since $P_2M_1 = \hat{M}_1P_2$,

$$M_2\hat{M}_1P_2P_1A = M_2P_2M_1P_1A = U$$

Therefore,

$$P_2P_1A = \hat{M}_1^{-1}M_2^{-1}U$$

Therefore, if we let

$$P = P_2P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

and

$$L = \hat{M}_1^{-1}M_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/4 & 1/3 & 1 \end{pmatrix}$$

we have $PA = LU$.

(b) To use the LU factorization of A computed in part (a) above to solve the linear system $Ax = b$, where

$$b = \begin{pmatrix} -2 \\ 4 \\ -4 \end{pmatrix}$$

first let

$$\hat{b} = Pb = \begin{pmatrix} 4 \\ -4 \\ -2 \end{pmatrix}$$

Now note that

$$LUx = PAx = Pb = \hat{b}$$

Therefore, solving $LUx = \hat{b}$ is equivalent to solving $Ax = b$. To solve $LUx = \hat{b}$, first let $y = Ux$, solve $Ly = \hat{b}$ and then solve $Ux = y$. Note that $Ly = \hat{b}$ is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/4 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -4 \\ -2 \end{pmatrix}$$

So,

$$\begin{aligned} y_1 &= 4 \\ \frac{1}{2}y_1 + y_2 &= -4 \\ y_2 &= -4 - \frac{1}{2}y_1 \\ y_2 &= -4 - \frac{1}{2}4 \\ y_2 &= -6 \\ -\frac{1}{4}y_1 + \frac{1}{3}y_2 + y_3 &= -2 \\ y_3 &= -2 + \frac{1}{4}y_1 - \frac{1}{3}y_2 \\ y_3 &= -2 + \frac{1}{4}4 + \frac{1}{3}6 \\ y_3 &= 1 \end{aligned}$$

That is,

$$y = \begin{pmatrix} 4 \\ -6 \\ 1 \end{pmatrix}$$

Now note that $Ux = y$ is

$$\begin{pmatrix} 4 & 4 & -4 \\ 0 & -3 & -3 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -6 \\ 1 \end{pmatrix}$$

So,

$$\begin{aligned}
x_3 &= 1 \\
-3x_2 - 3x_3 &= -6 \\
-3x_2 &= -6 + 3x_3 \\
-3x_2 &= -6 + 3 \\
-3x_2 &= -3 \\
x_2 &= 1 \\
4x_1 + 4x_2 - 4x_3 &= 4 \\
4x_1 &= 4 - 4x_2 + 4x_3 \\
4x_1 &= 4 - 4 + 4 \\
4x_1 &= 4 \\
x_1 &= 1
\end{aligned}$$

That is,

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

(c) We want to find a u and v so that

$$A - uv^T = \hat{A} = \begin{pmatrix} -1 & -2 & 1 \\ 4 & 4 & -4 \\ 2 & -1 & 1 \end{pmatrix}$$

That is, we need

$$uv^T = A - \hat{A} = \begin{pmatrix} -1 & -2 & 1 \\ 4 & 4 & -4 \\ 2 & -1 & -5 \end{pmatrix} - \begin{pmatrix} -1 & -2 & 1 \\ 4 & 4 & -4 \\ 2 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -6 \end{pmatrix}$$

There are many vectors u and v that give the required matrix above. Give them full marks for any choice of u and v that gives the matrix above.
The choice that I will use is

$$u = \begin{pmatrix} 0 \\ 0 \\ -6 \end{pmatrix} \qquad v = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

This choice makes the arithmetic in part (d) fairly simple.

(d) Note that solving $\hat{A}\hat{x} = b$ is equivalent to computing $\hat{x} = \hat{A}^{-1}b$, where $\hat{A} = A - uv^T$. From the Sherman-Morrison formula, we have

$$\hat{A}^{-1} = (A - uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1 - v^TA^{-1}u}$$

Therefore,

$$\begin{aligned} \hat{x} = \hat{A}^{-1}b &= \left(A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1 - v^TA^{-1}u}\right) b \\ &= A^{-1}b + \frac{(A^{-1}u)\, v^T\, (A^{-1}b)}{1 - v^T\,(A^{-1}u)} \qquad (2) \end{aligned}$$

Now note that $x = A^{-1}b$ is equivalent to $Ax = b$. Moreover, we have already solved $Ax = b$ in part (b) above.

In addition, $w = A^{-1}u$ is equivalent to $Aw = u$. Moreover, we can use the LU factorization of A from part (a) to solve $Aw = u$ as follows. First compute

$$\hat{u} = Pu = \begin{pmatrix} 0 \\ -6 \\ 0 \end{pmatrix}$$

Then $Aw = u$ is equivalent to $PAw = Pu = \hat{u}$, which in turn is equivalent to $LUw = \hat{u}$. To solve $LUw = \hat{u}$, first let $z = Uw$, solve $Lz = \hat{u}$ and then solve $Uw = z$. We solve $Lz = \hat{u}$ as

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/4 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -6 \\ 0 \end{pmatrix}$$

Hence,

$$\begin{aligned} z_1 &= 0 \\ \frac{1}{2}z_1 + z_2 &= -6 \\ z_2 &= -6 \\ -\frac{1}{4}z_1 + \frac{1}{3}z_2 + z_3 &= 0 \\ z_3 &= -\frac{1}{3}z_2 \\ z_3 &= 2 \end{aligned}$$

That is,

$$z = \begin{pmatrix} 0 \\ -6 \\ 2 \end{pmatrix}$$

We solve $Uw = z$ as

$$\begin{pmatrix} 4 & 4 & -4 \\ 0 & -3 & -3 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -6 \\ 2 \end{pmatrix}$$

So,

$$\begin{aligned} w_3 &= 2 \\ -3w_2 - 3w_3 &= -6 \\ -3w_2 &= -6 + 3w_3 \\ -3w_2 &= -6 + 6 \\ -3w_2 &= 0 \\ w_2 &= 0 \\ 4w_1 + 4w_2 - 4w_3 &= 0 \\ 4w_1 &= -4w_2 + 4w_3 \\ 4w_1 &= 8 \\ w_1 &= 2 \end{aligned}$$

That is,

$$w = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}$$

Now substitute x for $A^{-1}b$ and w for $A^{-1}u$ in (2) to get

$$\begin{aligned} \hat{x} &= x + \frac{wv^T x}{1 - v^T w} \\ &= x + \frac{v^T x}{1 - v^T w} w \end{aligned}$$

Now note that

$$v^T x = (0\,0\,1) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1$$

and

$$v^T w = (0\,0\,1) \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix} = 2$$

Therefore,

$$\begin{aligned}\hat{x} &= x + \frac{v^T x}{1 - v^T w} w \\ &= x + \frac{1}{1-2} w \\ &= x - w \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}\end{aligned}$$

4. [15 marks: 5 marks for each part]

Note that $f(x) = 2 + \cos(x) - e^x$ for $x \in \mathbb{R}$ and we are given the following table.

x	$2+\cos(x)$	e^x
0.50000	2.87758	1.64872
0.60000	2.82534	1.82212
0.70000	2.76484	2.01375
0.80000	2.69671	2.22554
0.90000	2.62161	2.45960
1.00000	2.54030	2.71828
1.10000	2.45360	3.00417
1.20000	2.36236	3.32012
1.30000	2.26750	3.66930
1.40000	2.16997	4.05520
1.50000	2.07074	4.48169

(a) To find an interval of length at most 0.1 that contains a root of $f(x)$, note from the table above that

$$f(0.9) = 2 + \cos(x) - e^x = 2.62161 - 2.45960 > 0$$
$$f(1.0) = 2 + \cos(x) - e^x = 2.54030 - 2.71828 < 0$$

Since $f(x)$ is continuous and $f(0.9) > 0$ and $f(1.0) < 0$, by the intermediate value theorem, $f(x)$ must have a root in the interval $[0.9, 1.0]$. Note the length of this interval is 0.1.

(b) $f(x)$ has exactly one root.

To see that this is true, note that

$$f'(x) = -\sin(x) - \mathrm{e}^x$$

whence

$$f'(0) = -\sin(0) - \mathrm{e}^0 = 0 - 1 = -1 < 0$$

In addition, for $x > 0$, $-1 \leq \sin(x) \leq 1$ and $\mathrm{e}^x > 1$. Therefore, $-\sin(x) \leq 1$ and $-\mathrm{e}^x < -1$. Hence,

$$f'(x) = -\sin(x) - \mathrm{e}^x < 1 - 1 = 0$$

Hence, $f'(x) < 0$ for all $x \geq 0$. Therefore, $f(x)$ is strictly decreasing for all $x \geq 0$. Consequently, $f(x)$ can have at most one root $r \geq 0$. In part (a), we showed that $f(x)$ has a root $r \in [0.9, 1.0]$. Hence, $f(x)$ has exactly one root $r \geq 0$.

If $x < 0$, then $-1 \leq \cos(x) \leq 1$ and $\mathrm{e}^x < 1$. Therefore, if $x < 0$, $\cos(x) \geq -1$ and $-\mathrm{e}^x > -1$ Hence, if $x < 0$,

$$f(x) = 2 + \cos(x) - \mathrm{e}^x > 2 - 1 - 1 = 0$$

That is, if $x < 0$, then $f(x) > 0$. Hence, $f(x)$ cannot have a root $r < 0$.

So, we have shown that

- $f(x)$ has exactly one root $r \geq 0$, and
- $f(x)$ cannot have a root $r < 0$.

Hence, $f(x)$ has exactly one root r and $r \in [0.9, 1.0]$.

(c) Given a starting guess, x_0, Newton's iteration to find a root of $f(x)$ in general is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad \text{for } n = 0, 1, 2, \ldots$$

For our specific $f(x) = 2 + \cos(x) - \mathrm{e}^x$, $f'(x) = -\sin(x) - \mathrm{e}^x$. Hence, Newton's iteration to find a root of this specific function is

$$x_{n+1} = x_n - \frac{2 + \cos(x_n) - \mathrm{e}^{x_n}}{-\sin(x_n) - \mathrm{e}^{x_n}}$$

which can be simplified to

$$x_{n+1} = x_n + \frac{2 + \cos(x_n) - \mathrm{e}^{x_n}}{\sin(x_n) + \mathrm{e}^{x_n}}$$

To choose a good starting value for Newton's method above, recall that the root we are looking for is in the interval $[0.9, 1.0]$. Any value in this interval would be a good starting value for $f(x)$. Give them full marks for this part of the question if they choose any value in $[0.9, 1.0]$.

However, two particularly good choices are the midpoint of $[0.9, 1.0]$, which is $x_0 = 0.95$. This choice ensures that $|x_0 - r| \leq 0.05$, where r is the root of $f(x)$. If you were to choose an end-point of the interval instead, for example $x_0 = 0.9$, then all that you could claim is $|x_0 - r| \leq 0.1$.

A second good choice is the value that you get from the secant method. This is the same as the value you get if you interpolate the end-points $(0.9, f(0.9))$ and $(1.0, f(1.0))$ by polynomial of degree 1 (i.e., a line) and then find the root of the line. The formula for this is

$$x_0 = 0.9 - f(0.9)\frac{1.0 - 0.9}{f(1.0) - f(0.9)}$$

I wouldn't expect them to calculate the actual value for this x_0, since they don't have a calculator at the exam. However, it works out to be

$$x_0 = 0.94765$$

which is quite close to the midpoint, 0.95.

5. [5 marks]

There are several methods to find a polynomial $p(x)$ of degree 3 or less that satisfies

$$\begin{aligned} p(0) &= 1 \\ p(1) &= 0 \\ p(-1) &= 2 \\ p(2) &= 5 \end{aligned}$$

Give them full marks if use any of these methods correctly. Also, they don't need to simplify their answer. For example, they can leave it in Lagrange form or Newton form.

Note: even though there are several different forms of the interpolating polynomial, the different forms are all the same polynomial. That is, if you evaluate two different forms of the interpolating polynomial at any value x, you get the same value for $p(x)$. Equivalently, if you simplify the polynomial to the standard form

$$p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$$

you get exactly the same coefficients c_0, c_1, c_2 and c_3 not matter how you derive the polynomial.

If you use the matrix form, the interpolation conditions above are equivalent to

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 2 & 4 & 8 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 5 \end{pmatrix}$$

The students can solve this system any way they want. However, no matter how they solve it, they should get $c_0 = 1$, $c_1 = -2$, $c_2 = 0$ and $c_3 = 1$. Therefore,

$$p(x) = 1 - 2x + x^3$$

If you use the Lagrange form, you get

$$\begin{aligned} p(x) = & \frac{(x-1)(x+1)(x-2)}{2} 1 \\ & - \frac{x(x+1)(x-2)}{2} 0 \\ & - \frac{x(x-1)(x-2)}{6} 2 \\ & + \frac{x(x-1)(x+1)}{6} 5 \end{aligned}$$

If you simplify this a little, you get

$$p(x) = \frac{1}{2}(x-1)(x+1)(x-2) - \frac{1}{3}x(x-1)(x-2) + \frac{5}{6}x(x-1)(x+1)$$

As noted above, they are not required to simplify this further. However, if they do, they should get

$$p(x) = 1 - 2x + x^3$$