Solution to the 2018 CSC 336 Exam

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

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

   (a) A good algorithm will produce an accurate solution to a problem regardless of the conditioning of the problem being solved.

   False.

   If a problem is ill-conditioned, any small rounding that you make in solving the problem might result in a very large change in the computed solution. So, it is very likely that the computed solution will be inaccurate (at least in some cases).

   (b) In the IEEE double-precision floating-point number system, *machine epsilon*, often referred to as ϵ_{mach} in your textbook, is the smallest positive floating-point number. That is, there are no double-precision floating-point numbers between ϵ_{mach} and zero.

   False.

   The definition of *machine epsilon* that I gave them in class is that it is the distance from 1 to the next larger machine number. This is very different from the smallest positive floating-point number.

   (c) A well-conditioned matrix can have a very small determinant. That is, an $n \times n$ matrix A can have $\text{cond}(A)$ not too large (for example, $1 \le \text{cond}(A) \le 10$), but $\det(A)$ very close to 0 (i.e., $0 < \det(A) \ll 1$).

   True.

   An example of a matrix A with $\text{cond}(A)$ not too large but $\det(A)$ very close to 0 (i.e., $0 < \det(A) \ll 1$) is

$$A = \begin{pmatrix} \epsilon & 0 \\ 0 & \epsilon \end{pmatrix}$$

   where $0 < \epsilon \ll 1$. In this case,

$$\begin{aligned} \text{cond}_\infty(A) &= \|A\|_\infty \|A^{-1}\|_\infty \\ &= \epsilon \, \frac{1}{\epsilon} \\ &= 1 \end{aligned}$$

   but

$$\det(A) = \epsilon^2$$

   So, $\text{cond}_\infty(A) = 1$ but $0 < \det(A) = \epsilon^2 \ll 1$.

(d) If an iterative method for solving a nonlinear equation gains more than one bit of accuracy per iteration, then it is said to have a superlinear rate of convergence.

False.

A linearly convergent iterative method can gain more than 1 bit of accuracy per iteration. To see this, recall that a linearly convergent iterative method satisfies

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|} = C$$

for some $C < 1$, where x^* is the root. If $C < 1/2$, then this iterative method will gain more than one bit of accuracy per iteration (at least for n sufficiently large).

(e) Suppose you are given N data points, $(t_1, y_1), (t_2, y_2), \ldots, (t_N, y_N)$, where

- N is a positive integer,
- each $t_n \in \mathbb{R}$ and each $y_n \in \mathbb{R}$, for $n = 1, 2, \ldots, N$, and
- $t_1 < t_2 < \cdots < t_N$.

Then there are infinitely many polynomials of degree N that interpolate the data points $(t_1, y_1), (t_2, y_2), \ldots (t_N, y_N)$.

True.

I showed them in class that there is a unique polynomial $p_N(t)$ of degree $N - 1$ or less that interpolates the data:

$$p_N(t_i) = y_i \qquad \text{for } i = 1, 2, \ldots, N$$

Now, for any $c \in \mathbb{R}$, let

$$p_{N,c}(t) = p_N(t) + c(t - t_1) \cdots (t - t_N)$$

Note that, for any $c \neq 0$, $p_{N,c}(t)$ is a polynomial of degree N and

$$p_{N,c}(t_i) = y_i \qquad \text{for } i = 1, 2, \ldots, N$$

So, for any $c \neq 0$, $p_{N,c}(t)$ is a polynomial of degree N that interpolates the data. Since there are infinitely many nonzero $c \in \mathbb{R}$ and each of them gives rise to a different polynomial $p_{N,c}(t)$ (i.e., $p_{N,c_1}(t) \neq p_{N,c_2}(t)$ if $c_1 \neq c_2$), there are infinitely many polynomials of degree N that interpolate the data points $(t_1, y_1), (t_2, y_2), \ldots (t_N, y_N)$.

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

I told the students that the function

$$f(x) = \frac{e^x - 1}{x}$$

satisfies

$$\lim_{x \to 0} f(x) = 1 \tag{1}$$

I also told them that they don't have to prove (1); just accept it as true.

I also gave them a table on page 4 of the exam that shows the computed values of $f(x)$ for $x = 10^{-k}$ and $k = 1, 2, \ldots, 15$.

(a) I noted that the computed values for $f(x)$ first seem to be converging to 1 for $k = 1, 2, \ldots, 8$, but then diverge from 1 for $k = 11, 12, \ldots, 15$. I asked them to explain why this happens.

The students should do a little rounding error analysis to explain why the computed values for $f(x)$ in the table behave the way they do. To this end, I told them that they can assume

$$\exp(x) = e^x(1 + \delta_x)$$

where δ_x changes with x, but its magnitude is at most a few multiples of ϵ_{mach}. (I.e., $|\delta_x| \le c\,\epsilon_{\text{mach}}$ for some c that is at most 2 or 3.)

Therefore,

$$\begin{aligned} \mathrm{fl}(f(x)) &= \mathrm{fl}\left(\frac{e^x - 1}{x}\right) \\ &= \frac{(e^x(1 + \delta_x) - 1)(1 + \delta_1)}{x}(1 + \delta_2) \end{aligned} \tag{2}$$

for some δ_1 and δ_2 satisfying $|\delta_1| \le \frac{1}{2}\epsilon_{\text{mach}}$ and $|\delta_2| \le \frac{1}{2}\epsilon_{\text{mach}}$. Now we can perform standard mathematical operations on the last line of (2) to get

$$\begin{aligned} \mathrm{fl}(f(x)) &= \frac{e^x - 1 + e^x\delta_x}{x}(1 + \delta_1)(1 + \delta_2) \\ &= \left(\frac{e^x - 1}{x} + \frac{\delta_x}{x}e^x\right)(1 + \delta_1)(1 + \delta_2) \\ &= \left(\left(1 + \frac{1}{2}x + \mathcal{O}(x^2)\right) + \left(\frac{\delta_x}{x}e^x\right)\right)(1 + \delta_1)(1 + \delta_2) \\ &= \left(1 + \left(\frac{1}{2}x + \mathcal{O}(x^2)\right) + \left(\frac{\delta_x}{x}e^x\right)\right)(1 + \delta_1)(1 + \delta_2) \end{aligned} \tag{3}$$

From our assumption above, $|\delta_x| \leq c\,\epsilon_{\text{mach}} \lessapprox 10^{-15}$. So, for $k = 1, 2, \ldots, 6$ and $x = 10^{-k}$,

$$\left|\frac{\delta_x}{x}\right| \ll \frac{1}{2}x + \mathcal{O}(x^2)$$

Hence, from the last line of (3),

$$\text{fl}(f(x)) \approx 1 + \frac{1}{2}x + \mathcal{O}(x^2)$$

That is, our rounding error analysis predicts that the computed value of $f(x)$ will behave like $1 + \frac{1}{2}x + \mathcal{O}(x^2)$ for $k = 1, 2, \ldots, 6$. We see quite clearly in the table on page 4 of the exam that this is indeed the case.

For the values of k in the range $k = 7, 8, \ldots, 11$, the behaviour of $f(x)$ is not as clear. That's because, for the k in this range,

$$\left|\frac{\delta_x}{x}\right| \approx \frac{1}{2}x + \mathcal{O}(x^2)$$

Hence, from the last line of (3), we see that both

$$\frac{1}{2}x + \mathcal{O}(x^2)$$

and

$$\frac{\delta_x}{x}$$

affect the behaviour of $f(x)$. So, our rounding error analysis predicts that the behaviour of $f(x)$ is not particularly clear in this range. This prediction is supported by the data in the table on page 4 of the exam.

However, for $k = 12, 13, 14, 15$,

$$0 < \frac{1}{2}x + \mathcal{O}(x^2) \ll \left|\frac{\delta_x}{x}\right|$$

So, for this range of k, our rounding error analysis predicts that

$$\text{fl}(f(x)) \approx 1 + \frac{\delta_x}{x}$$

Since the δ_x is somewhat "random" in the range $[-c\,\epsilon_{\text{mach}}, c\,\epsilon_{\text{mach}}]$, the values of $f(x)$ for k in this range are somewhat erratic, but $|\delta_x/x|$ generally grows at x decreases (e.g., k increases). Hence, $f(x)$ diverges from 1 (in a somewhat erratic way) as k increases for k in this range. This prediction is supported by the data in the table on page 4 of the exam.

(b) The students are asked to explain why the computed values for

$$g(x) = \frac{\mathrm{e}^x - 1}{\ln(\mathrm{e}^x)}$$

shown in column four of the table on page 3 of the exam (see the file exam.2018.pdf) give much more accurate results for small x than $f(x)$ does, even though in exact arithmetic $f(x) = g(x)$ for all $x \in \mathbb{R}$ (assuming you define $f(0) = g(0) = 1$).

To see how rounding errors affect $g(x)$, we first need to see how rounding errors affect $\ln(u)$ for u close to 1. It's reasonable to assume that

$$\mathrm{fl}(\ln(u)) = \ln(u)(1 + \delta_u) \tag{4}$$

However, $|\delta_u|$ might be much larger than ϵ_{mach}, since $\ln(u)$ is ill-conditioned for u close to 1. (Note, we are assuming here that $u = \mathrm{e}^x$ and $|x|$ is small, so $u \approx 1$.) For now, let's not try to determine a bound on $|\delta_u|$. We will come back to that later. So, using (4), we can perform a rounding error analysis on $g(x)$ that is much like the one in part (a) for $f(x)$. That is,

$$\begin{aligned} \mathrm{fl}(g(x)) &= \mathrm{fl}\left(\frac{\mathrm{e}^x - 1}{\ln(\mathrm{e}^x)}\right) \\ &= \frac{(\mathrm{e}^x(1+\delta_x) - 1)(1+\delta_1)}{\big(\ln(\mathrm{e}^x(1+\delta_x))\big)(1+\delta_u)}(1+\delta_2) \end{aligned} \tag{5}$$

for some δ_1 and δ_2 satisfying $|\delta_1| \le \frac{1}{2}\epsilon_{\mathrm{mach}}$ and $|\delta_2| \le \frac{1}{2}\epsilon_{\mathrm{mach}}$. It's important to note that the rounding error that is made when computing e^x is the same for the e^x in the numerator of (5) and the e^x in the denominator of (5). More generally, the rounding error that is made when computing e^x is deterministic. So, the rounding error is the same whenever e^x computed for the same value of x. Therefore, the δ_x in the numerator of (5) is the same as the δ_x in the denominator of (5). This is very important for the analysis below.

For the analysis that follows, it is convenient to note that there is a $\hat{\delta}_x$ such that

$$\mathrm{e}^{x+\hat{\delta}_x} = \mathrm{e}^x(1 + \delta_x) \tag{6}$$

where by taking logarithms of both sides of (6), we see that

$$x + \hat{\delta}_x = x + \ln(1 + \delta_x)$$

whence

$$\hat{\delta}_x = \ln(1 + \delta_x) = \delta_x + \mathcal{O}(\delta_x^2)$$

Since we assumed in part (a) that $|\delta_x| \le c\,\epsilon_{\mathrm{mach}}$ for some c that is at most 2 or 3, it follows that $|\hat{\delta}_x| \le \hat{c}\,\epsilon_{\mathrm{mach}}$ for some $\hat{c}$ that is only slightly different from c. That is, we can also assume $\hat{c}$ is at most 2 or 3.

Therefore, we can rewrite (5) as

$$
\begin{aligned}
\mathrm{fl}(g(x)) &= \frac{(\mathrm{e}^{x+\hat{\delta}_x} - 1)(1+\delta_1)}{\big(\ln(\mathrm{e}^{x+\hat{\delta}_x})\big)(1+\delta_u)}(1+\delta_2) \\
&= \frac{\mathrm{e}^{x+\hat{\delta}_x} - 1}{\ln(\mathrm{e}^{x+\hat{\delta}_x})} \times \frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_u)} \\
&= \frac{(x+\hat{\delta}_x) + \frac{1}{2}(x+\hat{\delta}_x)^2 + \mathcal{O}((x+\hat{\delta}_x)^3)}{(x+\hat{\delta}_x)} \times \frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_u)} \\
&= \left(1 + \frac{1}{2}(x+\hat{\delta}_x) + \mathcal{O}((x+\hat{\delta}_x)^2\right) \times \frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_u)}
\end{aligned}
\tag{7}
$$

For $k = 1, 2, \ldots, 13$ and $x = 10^{-k}$,

$$|\hat{\delta}_x| \ll x$$

So,

$$\left(1 + \frac{1}{2}(x+\hat{\delta}_x) + \mathcal{O}((x+\hat{\delta}_x)^2\right) \approx 1 + \frac{1}{2}x \tag{8}$$

which agrees very well with the numerical results shown in the table on page 4 of the exam. A slightly surprising thing is that the term

$$\frac{(1+\delta_1)(1+\delta_2)}{(1+\delta_u)}$$

on the right in (7) does not disturb the result (8). Although the δ_1 and δ_2 terms would not disturb the result (8), since $|\delta_1| \le \frac{1}{2}\epsilon_{\text{mach}}$ and $|\delta_2| \le \frac{1}{2}\epsilon_{\text{mach}}$, I would have expected that the δ_u term could disturb the result (8), since I think we could have $|\delta_u| \gg \epsilon_{\text{mach}}$. However, the results in the table on page 4 of the exam do not suffer from this potentially large perturbation.

Also, for $k = 14, 15$, you might expect that

$$|\hat{\delta}_x| \not\ll x$$

This could also perturb the result (8). However, this potential perturbation does not appear to occur in the numerical results reported in the table on page 4 of the exam.

3. [15 marks: 2 marks for each of parts (a) and (c); 3 marks for each of parts (b) and (e); 5 marks for part (d)]

(a) [2 marks]

I asked the student to show that, if A is an $n \times n$ real symmetric positive-definite matrix, then $A_{i,i} > 0$ for all $i = 1, 2, \ldots, n$.

I gave them the following hint.

> Hint: for each $i = 1, 2, \ldots, n$, choose a particular $\hat{x} \in \mathbb{R}^n$ for which $\hat{x} \neq \vec{0}$ and $A_{i,i} = \hat{x}^T A \hat{x}$. Then note that $\hat{x}^T A \hat{x} > 0$, since $\hat{x} \neq \vec{0}$ and A is an $n \times n$ real symmetric positive-definite matrix.
> What is the required vector $\hat{x}$?

The required $\hat{x}$ is $\hat{x} = e_i$, where $e_i \in \mathbb{R}^n$ is the vector with all elements equal to 0 except for the i^{th} element, which is 1. (Another way of saying this is that e_i is the i^{th} column of the $n \times n$ identity matrix.) It is very easy to see from this that

$$\hat{x}^T A \hat{x} = e_i^T A e_i = A_{i,i}$$

In addition, since $e_i \neq \vec{0}$ and A is symmetric positive-definite, we must have $e_i^T A e_i > 0$. Therefore, $A_{i,i} = e_i^T A e_i > 0$.

[If they don't give the last two sentences above, don't take off any marks, since it is just repeating what is in the hint. Give them the full 2 marks if they say $\hat{x} = e_i$.]

(b) [3 marks]

I told the students to let

$$m_{i,1} = A_{i,1}/A_{1,1} \qquad \text{for } i = 2, \ldots, n$$

and form the vectors

$$m_1 = \begin{pmatrix} 0 \\ m_{2,1} \\ m_{3,1}, \\ \vdots \\ m_{n,1} \end{pmatrix} \qquad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and the matrix

$$M_1 = I - m_1\, e_1^T$$

where I is the $n \times n$ identity matrix.

Then I asked the students to show that

$$A_1 = M_1 A M_1^T = \begin{pmatrix} A_{1,1} & 0 & 0 & \cdots & 0 \\ 0 & \hat{A}_{2,2} & \hat{A}_{2,3} & \cdots & \hat{A}_{2,n} \\ 0 & \hat{A}_{3,2} & \hat{A}_{3,3} & \cdots & \hat{A}_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \hat{A}_{n,2} & \hat{A}_{n,3} & \cdots & \hat{A}_{n,n} \end{pmatrix} \tag{9}$$

where $A_{1,1}$ is the (1,1)-element of the original matrix A and the $\hat{A}_{i,j}$, for $i = 2, \ldots, n$ and $j = 2, \ldots, n$, are modified elements of A computed by multiplying A by M_1 on the left and by M_1^T on the right.

To see that (9) holds, first note that $M_1 A$ is just the matrix that we would get from the first stage of Gaussian elimination. That is,

$$\hat{A}_1 = M_1 A = \begin{pmatrix} A_{1,1} & A_{1,2} & A_{1,3} & \cdots & A_{1,n} \\ 0 & \hat{A}_{2,2} & \hat{A}_{2,3} & \cdots & \hat{A}_{2,n} \\ 0 & \hat{A}_{3,2} & \hat{A}_{3,3} & \cdots & \hat{A}_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \hat{A}_{n,2} & \hat{A}_{n,3} & \cdots & \hat{A}_{n,n} \end{pmatrix} \tag{10}$$

where the elements $A_{1,i}$ for $i = 1, 2, \ldots, n$, in the first row of $\hat{A}_1$ are the elements in the first row of A. That is, the first row of A is unchanged by the multiplication $M_1 A$.

Now, when you compute $\hat{A}_1 M_1^T$, this has the effect of multiplying column 1 of $\hat{A}_1$ by $m_{i,1}$ and subtracting it from column i of $\hat{A}_1$, for $i = 2, \ldots, n$. So, the $(1, i)$

element of $A_1 = \hat{A}_1 M_1^T$ becomes

$$\begin{aligned} A_{1,i} - m_{i,1}A_{1,1} &= A_{1,i} - m_{i,1}A_{1,1} \\ &= A_{1,i} - (A_{i,1}/A_{1,1})A_{1,1} \\ &= A_{1,i} - A_{i,1} \\ &= 0 \end{aligned}$$

where the last line follows from the symmetry of A. Therefore, the first row of $A_1 = \hat{A}_1 M_1^T = M_1 \hat{A}_1 M_1^T$ has zeros in elements $(1, i)$, for $i = 2, \ldots, n$, as shown in (9).

Note that the first column of $\hat{A}_1$ is not changed by the multiplication $\hat{A}_1 M_1^T$. Therefore, the $(1, 1)$ element of $A_1 = \hat{A}_1 M_1^T = M_1 \hat{A}_1 M_1^T$ is $A_{1,1}$ and the elements $(i, 1)$, for $i = 2, \ldots, n$, are zero, as shown in (9).

Finally, note that the multiplication $\hat{A}_1 M_1^T$ does not change the elements (i, j), for $i = 2, \ldots, n$ and $j = 2, \ldots, n$ of $\hat{A}_1$, because all the elements $(i, 1)$ for $i = 2, \ldots, n$ in the first column of $\hat{A}_1$ are zero. So, the elements $\hat{A}_{i,j}$, for $i = 2, \ldots, n$ and $j = 2, \ldots, n$, in A_1 and $\hat{A}_1$ are exactly the same.

(c) [2 marks]

Show that the matrix A_1 shown in (9) is an $n \times n$ real symmetric positive-definite matrix.

$A_1 = M_1 A M_1^T$ is obviously an $n \times n$ real matrix, because each of M_1, M_1^T and A are $n \times n$ real matrices. Hence, the product $M_1 A M_1^T$ is an $n \times n$ real matrix.
[If they do not mention this, do not take off any marks.]
To see that $A_1 = M_1 A M_1^T$ is symmetric note that

$$\begin{aligned} A_1^T &= \left(M_1 A M_1^T\right)^T \\ &= \left(M_1^T\right)^T A^T M_1^T \\ &= M_1 A M_1^T \\ &= A_1 \end{aligned}$$

where we have used the fact that A is symmetric (i.e., $A = A^T$). Since $A_1^T = A_1$, A_1 is symmetric.
To see that $A_1 = M_1 A M_1^T$ is also positive-definite, note that for any $x \neq \vec{0}$, $y = M_1^T x$ also satisfies $y \neq \vec{0}$, since M_1 is nonsingular, hence M_1^T is also nonsingular. Therefore, $y^T A y > 0$, since $y \neq \vec{0}$ and A is symmetric positive-definite. Putting these pieces together, we get that for any $x \neq \vec{0}$

$$\begin{aligned} x^T A_1 x &= x^T \left(M_1 A M_1^T\right) x \\ &= \left(x^T M_1\right) A \left(M_1^T x\right) \\ &= \left(M_1^T x\right)^T A \left(M_1^T x\right) \\ &= y^T A y \\ &> 0 \end{aligned}$$

(d) [5 marks]

Show that you can compute A_1 with $\frac{1}{2}n(n-1)$ adds and multiplications and $n-1$ divisions.

We need $n-1$ divisions to compute the multipliers

$$m_{i,1} = A_{i,1}/A_{1,1} \qquad \text{for } i = 2, \ldots, n$$

Having computed the multipliers with $n-1$ divisions, we need to show that we can compute $A_1 = M_1 A M_1^T$ with $\frac{1}{2}n(n-1)$ additional adds and multiplications.

From (9), it is clear that we only need to compute the $\hat{A}_{i,j}$ for $i = 2, \ldots, n$ and $j = 2, \ldots, n$, since $A_{1,1}$ is the $(1,1)$ element of A and so does not need to be computed, and the zeros in the first row and column of A_1 don't need to be computed either, since we chose the multipliers so that these elements would be zero.

First note that A_1 is symmetric, so we need to compute only elements (i,j) of A_1 for $i = 2, \ldots, n$ and $j = 2, \ldots, i$, since we can use the symmetry of A_1 to get the other elements. That is, you only need to compute element (i,j) of A_1 for $i = 2, \ldots, n$ and $j = 2, \ldots, i$, since elements (i,j) and (j,i) of A_1 are the same. So, you don't need to compute the elements (j,i) of A_1 — you essentially get them for free. Hence, we only need to compute $\frac{1}{2}n(n-1)$ elements of A_1.

Second, we noted in part (b) above that the elements $\hat{A}_{i,j}$ for $i = 2, \ldots, n$ and $j = 2, \ldots, n$ in (9) and (10) are the same. So, we only need to compute elements $\hat{A}_{i,j}$ in (10) for $i = 2, \ldots, n$ and $j = 2, \ldots, i$. Moreover, to compute each element $\hat{A}_{i,j}$ in (10) requires one multiplication and one subtraction (which we usually call an addition). Therefore, we can compute all the $\hat{A}_{i,j}$ for $i = 2, \ldots, n$ and $j = 2, \ldots, i$ with $\frac{1}{2}n(n-1)$ adds and multiplications and then use the symmetry of A_1 to get the other elements $\hat{A}_{i,j}$ for $i = 2, \ldots, n-1$ and $j = i+1, \ldots, n$ without any additional computational work.

Therefore, the total computational work required to compute $A_1 = M_1 A M_1^T$ is $\frac{1}{2}n(n-1)$ adds and multiplications and $n-1$ divisions.

(e) [3 marks]

I asked the students to show that they can rewrite

$$M_{n-1}M_{n-2}\cdots M_2M_1AM_1^TM_2^T\cdots M_{n-2}^TM_{n-1}^T = D \tag{11}$$

as

$$A = LDL^T \tag{12}$$

I also asked them if they can determine the L needed in (12) without any additional arithmetic work and to justify their answer.

The key here is to note that the $M_k = I - m_k e_k^T$ in (11) are the same as the M_k that we used in Gaussian elimination. Therefore, they can use without proof that $M_k^{-1} = I + m_k e_k^T$ and that

$$M_1^{-1}M_2^{-2}\cdots M_{n-1}^{-1} = I + m_1e_1^T + m_2e_2^T + \cdots + m_{n-1}e_{n-1}^T \tag{13}$$

Moreover, each of the $m_k e_k^T$ in (13) is an $n\times n$ matrix with all elements zero except for the elements in the k^{th} column below the diagonal, which are the multipliers used in the k^{th}-stage of the LDL factorization. Therefore, you can form the lower triangular matrix

$$L = M_1^{-1}M_2^{-2}\cdots M_{n-1}^{-1} = I + m_1e_1^T + m_2e_2^T + \cdots + m_{n-1}e_{n-1}^T$$

without doing any additional computational work: you just have to put 1's on the diagonal of L and copy the the multipliers used in the k^{th}-stage of the LDL factorization into the k^{th} column of L below the diagonal.

Note also that

$$L^T = (M_1^{-1}M_2^{-2}\cdots M_{n-1}^{-1})^T = M_{n-1}^{-T}\cdots M_2^{-T}M_1^{-T}$$

where I have used M_k^{-T} for $(M_k^{-1})^T$.

We also need below that $M_k^{-T} = (M_k^{-1})^T = (M_k^T)^{-1}$.

Therefore, we have from (11) and the discussion above that

$$\begin{aligned} A &= M_1^{-1}M_2^{-2}\cdots M_{n-1}^{-1}DM_{n-1}^{-T}\cdots M_2^{-T}M_1^{-T} \\ &= LDL^T \end{aligned}$$

As explained above, we don't need any additional computational work to determine the L. We just need to copy values that are already computed into the right place in L.

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

(a) To show that there is a unique point $x^* > \hat{x}$ for which $f(x^*) = 0$, we will follow the advice of the hint and first show that $f(x) \to \infty$ as $x \to \infty$.
First note that, since $f'(\hat{x}) = 0$

$$f'(x) = f'(x) - f'(\hat{x}) = \int_{\hat{x}}^{x} f''(t)\, dt > 0$$

for $x > \hat{x}$, since $f''(x) > 0$ for all $x \in \mathbb{R}$. In addition, $f'(x)$ is strictly increasing. That is, if $x_1 < x_2$, then $f'(x_1) < f'(x_2)$, since

$$f'(x_2) - f'(x_1) = \int_{x_1}^{x_2} f''(x)\, dx > 0$$

Putting these two results together, we get that $f'(x) > f'(\hat{x}+1) > 0$ for $x > \hat{x}+1$. Therefore,

$$\begin{aligned} f(x) - f(\hat{x}) &= \int_{\hat{x}}^{x} f'(t)\, dt \\ &= \int_{\hat{x}}^{\hat{x}+1} f'(t)\, dt + \int_{\hat{x}+1}^{x} f'(t)\, dt \\ &\geq 0 + f'(\hat{x}+1)(x - (\hat{x}+1)) \end{aligned}$$

since $f'(t) \geq 0$ for $t \in [\hat{x}, \hat{x}+1]$ and $f'(t) \geq f'(\hat{x}+1)$ for $t \in [\hat{x}+1, x]$. Since $f'(\hat{x}+1) > 0$, $f'(\hat{x}+1)(x - (\hat{x}+1) \to \infty$ as $x \to \infty$. Therefore, $f(x) \to \infty$ as $x \to \infty$.
Since $f(x) \to \infty$ as $x \to \infty$, there must be an $\check{x} > \hat{x}$ such $f(\check{x}) > 0$.
In addition, since $f''(x)$ exists and is continuous for all $x \in \mathbb{R}$, $f'(x)$ exists and is continuous for all $x \in \mathbb{R}$, which in turn implies that $f(x)$ exists and is continuous for all $x \in \mathbb{R}$.
Therefore, we have that

- $f(x)$ is continuous for all $x \in \mathbb{R}$,
- $f(\hat{x}) < 0$ and $f(\check{x}) > 0$.

Therefore, by the Intermediate Value Theorem, there is an $x^* \in (\hat{x}, \check{x})$ such that $f(x^*) = 0$. That is, there is an $x^* > \hat{x}$ such that $f(x^*) = 0$.
To see that x^* is the only point $> \hat{x}$ for which $f(x^*) = 0$, it is sufficient to note that $f(x)$ is a strictly increasing function of x for $x > \hat{x}$, since $f'(x) > 0$ for all $x > \hat{x}$.
Alternatively, they could prove the result by contradiction as follows. Suppose there is another point $y^* > \hat{x}$ for which $f(y^*) = 0$. If $x^* < y^*$, then

$$f(y^*) - f(x^*) = \int_{x^*}^{y^*} f'(x)\, dx > 0$$

since we showed above that $f'(x) > 0$ for all $x > \hat{x}$ and $x \in [x^*, y^*]$ implies that $x > \hat{x}$. However, this contradicts, $f(y^*) - f(x^*) = 0$, which follows from $f(x^*) = 0$ and $f(y^*) = 0$. Assuming $x^* > y^*$ leads to a similar contradiction. Therefore, we must have $x^* = y^*$. That is, there is only one point $x^* > \hat{x}$ for which $f(x^*) = 0$.

(b) I asked the students to show that, if $x_0 > \hat{x}$ and x_n, for $n = 1, 2, \ldots$, is generated by Newton's method

$$x_n = x_{n-1} - f(x_{n-1})/f'(x_{n-1}), \qquad \text{for } n = 1, 2, \ldots \tag{14}$$

then

- $x^* \leq x_n$ for $n = 1, 2, \ldots$, and
- $x_{n+1} \leq x_n$ for $n = 1, 2, \ldots$

That is, the x_n form a decreasing sequence that is bounded below by x^*.

A few people seemed to be confused by the assumption that I asked them to show $x^* \leq x_n$ for $n = 1, 2, \ldots$, but I told them to assume only $x_0 > \hat{x}$. Since $\hat{x} < x^*$, they were worried that, if $x_0 \in (\hat{x}, x^*)$, then this would violate $x^* \leq x_n$ for $n = 1, 2, \ldots$. Of course, it doesn't, since the condition $x^* \leq x_n$ for $n = 1, 2, \ldots$ starts with $n = 1$, not $n = 0$.

First suppose $x_0 = x^*$. Then $f(x_0) = f(x^*) = 0$ and, from part (a), $f'(x_0) = f'(x^*) > 0$. Therefore, (14) with $n = 1$, gives $x_1 = x_0 = x^*$. It follows immediately by induction on n that $x_n = x_0 = x^*$ for all $n = 1, 2, \ldots$. Hence,

- $x^* \leq x_n$ for $n = 1, 2, \ldots$, and
- $x_{n+1} \leq x_n$ for $n = 1, 2, \ldots$

Next assume that $x_0 \in (\hat{x}, x^*)$. Since $f(x)$ is a strictly increasing function for $x > \hat{x}$ (since $f'(x) > 0$ for $x > \hat{x}$) and $f(x^*) = 0$, we must have $f(x_0) < 0$. Also, $f'(x_0) > 0$. Therefore, from (14), $x_1 > x_0$. Hence, $x_1 > x_0 > \hat{x}$.
Now consider the line

$$l_0(x) = f(x_0) + (x - x_0) f'(x_0)$$

We showed in class that the point x_1 generated from x_0 by Newton's method (14) satisfies $l_0(x_1) = 0$. It obviously also satisfies $l_0(x_0) = f(x_0)$. Therefore,

$$\begin{aligned}
f(x_1) &= f(x_1) - l_0(x_1) \\
&= \Big(f(x_1) - l_0(x_1)\Big) - \Big(f(x_0) - l_0(x_0)\Big) \\
&= \int_{x_0}^{x_1} \big(f'(x) - l_0'(x)\big)\, dx \\
&= \int_{x_0}^{x_1} \big(f'(x) - f'(x_0)\big)\, dx \\
&> 0
\end{aligned}$$

since $f'(x) > f'(x_0)$ for $x > x_0$. Now recall that $f(x)$ is a strictly increasing function of x and $f(x^*) = 0$ and $f(x_1) > 0$. Therefore, $x_1 > x^*$.

One the other hand, if $x_0 > x^*$, then $f(x_0) > 0$, since $f(x)$ is an increasing function of x and $f(x^*) = 0$. We also have that $f'(x_0) > 0$, since we showed above that $f'(x) > 0$ for all $x > \hat{x}$ and $x_0 > x^* > \hat{x}$. Therefore, we have from (14) that $x_1 < x_0$. Now consider again the line

$$l_0(x) = f(x_0) + (x - x_0)f'(x_0)$$

As noted above, this line satisfies $l_0(x_0) = f(x_0)$. Therefore,

$$\begin{aligned}
l_0(x^*) &= l_0(x^*) - f(x^*) \\
&= \Big(f(x_0) - l_0(x_0)\Big) - \Big(f(x^*) - l_0(x^*)\Big) \\
&= \int_{x^*}^{x_0} \big(f'(x) - l_0'(x)\big)\, dx \\
&= \int_{x^*}^{x_0} \big(f'(x) - f'(x_0)\big)\, dx \\
&< 0
\end{aligned}$$

since $f'(x) < f'(x_0)$ for $x = [x^*, x_0)$, since $f'(x)$ is an increasing function of x (because $f''(x) > 0$ for all $x \in \mathbb{R}$). Now note that $l_0(x_0) = f(x_0) > 0$, $l_0(x^*) < 0$ and $l_0(x)$ is continuous. So, $l_0(x)$ has a root $x_1 \in (x^*, x_0)$. However, the root x_1 of $l_0(x)$ is the iterate x_1 of Newton's method (14). Therefore, we have shown that the iterate x_1 for Newton's method (14) satisfies $x_1 > x^*$.

Hence, whether $x_0 < x^*$ or $x_0 > x^*$, we get $x_1 > x^*$.

Now we show by induction on n that

- $x^* < x_n$ for $n = 1, 2, \ldots,$ and
- $x_{n+1} < x_n$ for $n = 1, 2, \ldots$

For the base case, $n = 1$, we have already proved $x^* < x_1$. Since $f(x)$ is an increasing function for $x > \hat{x}$ and $f(x^*) = 0$, $f(x_1) > 0$. Also, $f'(x_1) > 0$. Therefore, from (14), $x_2 < x_1$. Therefore, we have proved the two statements

- $x^* < x_n$
- $x_{n+1} < x_n$

for $n = 1$.

Moreover, the general case is essentially the same as the proof given above for $x_0 > x^*$. That is, if we assume the induction hypothesis that $x^* < x_{n-1}$, then we can prove $x^* < x_n$ using the same approach as given above for to prove $x^* < x_1$ if we start from $x^* < x_0$. Once you have proven $x^* < x_n$, it follows easily that $f(x_n) > 0$ and $f'(x_n) > 0$. Hence it follows immediately from Newton's method (14) that $x_{n+1} < x_n$.

(c) We showed in part (b) that the x_n generated by Newton's method (14) form a decreasing sequence that is bounded below by x^*. I told them that they can use without proof that a decreasing sequence that is bounded below must converge. That is, they can conclude from part (b) without proof that $x_n \to y^*$ as $n \to \infty$ and that $x^* \leq y^*$.

They are asked to show in this part that $x^* = y^*$.

We will show that $x^* = y^*$ by first showing that $f(y^*) = 0$. Then recall that $f(x)$ has a unique root $x^* > \hat{x}$. Since both $f(x^*) = 0$ and $f(y^*) = 0$ and both $x^* > \hat{x}$ and $y^* > \hat{x}$, we must have that $x^* = y^*$ (since otherwise $f(x)$ would have two roots greater than $\hat{x}$).

So all that remains is to show that $f(y^*) = 0$. To this end note we can rewrite (14) as

$$f(x_{n-1}) = -f'(x_{n-1})(x_n - x_{n-1})$$

Hence,

$$|f(x_{n-1})| = |f'(x_{n-1})||x_n - x_{n-1}| \leq |f'(x_1)||x_n - x_{n-1}| \tag{15}$$

for $n \geq 2$, since $x^* < x_{n-1} < x_1$ from part (b) and $f'(x)$ is a positive increasing function for $x > x^*$, whence $0 < f'(x_{n-1}) < f'(x_1)$. Now $x_n \to y^*$ as $n \to \infty$. So, $|x_n - x_{n-1}| \to 0$ as $n \to \infty$. Hence, it follows from (15) that

$$\lim_{n\to\infty} f(x_{n-1}) = 0$$

However, $f(x)$ is a continuous function. So,

$$\lim_{n\to\infty} f(x_{n-1}) = f(y^*)$$

Thus, $f(y^*) = 0$.

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

I told the students to assume that we are given the data

$$\begin{array}{ccc} t_1 = -1 & t_2 = 0 & t_3 = 1 \\ y_1 = 1 & y_2 = 0 & y_3 = 1 \end{array}$$

and we want to find a polynomial $p(t)$ of degree 2 or less that satisfies

$$p(t_i) = y_i \qquad \text{for } i = 1, 2, 3.$$

(a) The students are asked to use the monomial basis approach to find the polynomial $p(t)$ in the form

$$p(t) = c_1 + c_2 t + c_3 t^2 \tag{16}$$

They are also asked to give the values of the coefficients c_1, c_2, c_3.

We can convert this problem of finding the coefficients c_1, c_2, c_3 of $p(t)$ to the following linear algebra problem for the coefficients c_1, c_2, c_3.

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

The middle equation gives $c_1 = 0$. Substituting this value into the first and third equations gives the smaller system

$$\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Adding these two equations together gives

$$2c_3 = 2$$

Hence, $c_3 = 1$ from which it follows that $c_2 = 0$.
Thus, our solution is $c_1 = 0$, $c_2 = 0$ and $c_3 = 1$. Hence, the polynomial is

$$p(t) = t^2$$

(b) The students are asked to use the Lagrange basis approach to find the polynomial $p(t)$ in the form

$$p(t) = l_1(t)y_1 + l_2(t)y_2 + l_3(t)y_3 \tag{17}$$

where the $l_i(t)$, for $i = 1, 2, 3$, are the Lagrange basis functions.

They are also asked to show that the polynomial $p(t)$ written in the monomial basis form (16) is the same as the polynomial $p(t)$ written in the Lagrange basis form (17).

To begin, note that we don't need $l_2(t)$ since $y_2 = 0$. The $l_1(t)$ and $l_3(t)$ Lagrange basis functions for this example are

$$l_1(t) = \frac{(t-t_2)(t-t_3)}{(t_1-t_2)(t_1-t_3)} = \frac{t(t-1)}{(-1)(-2)} = \frac{t(t-1)}{2}$$

$$l_3(t) = \frac{(t-t_1)(t-t_2)}{(t_3-t_1)(t_3-t_2)} = \frac{(t+1)t}{(2)(1)} = \frac{(t+1)t}{2}$$

Therefore

$$\begin{aligned} p(t) &= l_1(t)y_1 + l_2(t)y_2 + l_3(t)y_3 \\ &= \frac{t(t-1)}{2} + \frac{(t+1)t}{2} \\ &= t^2 \end{aligned}$$

Note that I've shown above that the polynomial $p(t)$ written in the monomial basis form (16) is the same as the polynomial $p(t)$ written in the Lagrange basis form (17). They are both $p(t) = t^2$.