1.15. (a) We expand the product:

$$\begin{aligned}[B(B^\top B)^{-1}B^\top]^2 &= [B(B^\top B)^{-1}B^\top]\cdot[B(B^\top B)^{-1}B^\top] \\ &= B[(B^\top B)^{-1}(B^\top B)(B^\top B)^{-1}]B^\top \text{ by associativity} \\ &= B(B^\top B)^{-1}B^\top \text{ by canceling out inverses.}\end{aligned}$$

(b) Again expanding the product,

$$\begin{aligned}(I_{n\times n} - A)^2 &= I_{n\times n} - 2A + A^2 \\ &= I_{n\times n} - 2A + A \text{ since } A \text{ is idempotent} \\ &= I_{n\times n} - A\end{aligned}$$

(c) The inverse of $\frac{1}{2}I_{n\times n} - A$ is $2I_{n\times n} - 4A$, since:

$$\begin{aligned}\left(\frac{1}{2}I_{n\times n} - A\right)(2I_{n\times n} - 4A) &= I_{n\times n} - 2A - 2A + 4A^2 \\ &= I_{n\times n} - 2A - 2A + 4A \text{ since } A \text{ is idempotent} \\ &= I_{n\times n}\end{aligned}$$

(d) $A\vec{x} = \lambda\vec{x} \implies A^2\vec{x} = \lambda A\vec{x} = \lambda^2\vec{x}$. But $A^2 = A$, so $A^2\vec{x} = A\vec{x} = \lambda\vec{x}$. Thus, $\lambda\vec{x} = \lambda^2\vec{x} \implies \lambda = \lambda^2 \implies \lambda \in \{0, 1\}$.

1.16. The product AB is of size n^2. Obviously finding AB requires considering each element of AB at least once (if nothing else, to write the result in memory!), already requiring $O(n^2)$ time even if each element of AB is computed in $O(1)$ time. The algorithms in the figure take $O(n^3)$ time to run due to the nested loops. Hence, there is room for improvement, and indeed Strassen's algorithm and several others achieve faster than $O(n^3)$ asymptotic runtime, at least for large n.

1.17. Define $\ell(\vec{x}) \equiv -\ln p(\vec{x})$. Since ln is monotonic, any local maximum of $p(\vec{x})$ is also a local maximum of $\ell(\vec{x})$. Hence, $\vec{x}^*$ is a critical point of $\ell(\vec{x})$, implying $\nabla\ell(\vec{x}^*) = \vec{0}$. Let H be the Hessian of ℓ at $\vec{x}^*$. Then, near $\vec{x}^*$ we can approximate:

$$-\ln p(\vec{x}) = \ell(\vec{x}) \approx \ell(\vec{x}^*) + \frac{1}{2}(\vec{x} - \vec{x}^*)^\top H(\vec{x} - \vec{x}^*) = -\ln p(\vec{x}^*) + \frac{1}{2}(\vec{x} - \vec{x}^*)^\top H(\vec{x} - \vec{x}^*).$$

The first derivative term of the expansion vanishes since $\nabla\ell(\vec{x}^*) = \vec{0}$. Exponentiating both sides shows

$$p(\vec{x}) \approx \text{const.} \cdot e^{-\frac{1}{2}(\vec{x}-\vec{x}^*)^\top(-H)(\vec{x}-\vec{x}^*)}.$$

Hence, a reasonable Gaussian approximation of $p(\vec{x})$ near $\vec{x}^*$ takes $\Sigma = -H^{-1}$ and $\vec{\mu} = \vec{x}^*$.

2.1. Depending on the processor, fixed-point arithmetic can be faster than floating-point since it can be carried out on the ALU with integer-type operations without the need for dealing with an exponent. Fixed-point arithmetic also can be applicable when the scale of numbers under consideration is known ahead of time, e.g., in financial software. Floating-point representations are more accurate, especially when values care on many scales.

2.2. (a) (answers may vary) Rounding error can come from multiplication and division to find the value n from the other variables. Discretization error can come from the representations of values from the sensors. Modeling error can result from inaccuracies of the Ideal Gas Law and/or failure to account for secondary factors like sensor noise or pollutants. Input error can result from using an inaccurate value of the constant R.

(b) From the ideal gas law, we can write

$$n = \frac{PV}{RT}.$$

Then, if we measure $\bar{P}$ and $\bar{T}$ rather than the ground-truth values, we can write the forward error as:

$$\left|\frac{\bar{P}V}{R\bar{T}} - \frac{PV}{RT}\right| = \frac{V}{R}\left|\frac{\bar{P}}{\bar{T}} - \frac{P}{T}\right|$$

$$
\begin{aligned}
&= \frac{V}{R}\left|\frac{P+\delta_P}{T+\delta_T} - \frac{P}{T}\right| \text{ for } |\delta_P| \le \varepsilon_P, |\delta_T| \le \varepsilon_T \\
&= \frac{V}{R}\left|\frac{(P+\delta_P)T - P(T+\delta_T)}{T(T+\delta_T)}\right| \\
&= \frac{V}{RT}\left|\frac{PT+\delta_P T - PT - \delta_T P}{T+\delta_T}\right| \\
&= n\left|\frac{\delta_P T - P\delta_T}{P(T+\delta_T)}\right|
\end{aligned}
$$

Hence, the relative forward error can be bounded as follows:

$$\left|\frac{\delta_P T - P\delta_T}{P(T+\delta_T)}\right| \le \frac{T\varepsilon_P + P\varepsilon_T}{P(T-\varepsilon_T)}$$

(c) In this case,

$$n = \frac{PV}{RT} = \frac{(100\text{ Pa})(0.5\text{ m}^3)}{(8.31\text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1})(300\text{ K})} = 0.0201\text{ mol}$$

With the given measurement bounds, the largest possible value is

$$\frac{(101\text{ Pa})(0.5\text{ m}^3)}{(8.31\text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1})(299.5\text{ K})} = 0.0203\text{ mol} = n + 0.000234\text{ mol} = n + 1.17\%.$$

The smallest possible value is

$$\frac{(99\text{ Pa})(0.5\text{ m}^3)}{(8.31\text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1})(300.5\text{ K})} = 0.0198\text{ mol} = n - 0.000234\text{ mol} = n - 1.17\%.$$

Hence, the absolute error is bounded by 0.0198 mol and the relative error is bounded by 1.17%.

(d) At the range indicated by the problem, it is relatively well-conditioned. When the scale of ε_T is commensurate with that of T, the problem becomes ill-conditioned.

2.3. We can understand the relative error as the fraction

$$\kappa_{\text{rel}} = \frac{|\Delta y|/|y|}{|\Delta x|/|x|} = \left|\frac{x\Delta y}{y\Delta x}\right|,$$

where $y + \Delta y = f(x + \Delta x)$ and $y = f(x)$. By Taylor's theorem, $f(x + \Delta x) = y + f'(x)\Delta x + O(\Delta x^2)$. Hence, $\Delta y = f'(x)\Delta x + O(\Delta x^2)$, so for small Δx,

$$\kappa_{\text{rel}} \approx \left|\frac{x \cdot f'(x)\Delta x}{f(x)\cdot\Delta x}\right| = \left|\frac{xf'(x)}{f(x)}\right|.$$

The absolute condition number of this problem is:

$$\left|\frac{\Delta y}{\Delta x}\right| \approx |f'(x)|.$$

The function $f(x) = \ln x$ has a large relative condition number near $x = 1$, since $\kappa_{\text{rel}} = 1/\ln x$, which blows up near $x = 1$. Contrastingly, the function $f(x) = x$ has relative condition number 1 for all x.

2.4. Since minima are roots of f', we can use the conditioning for root-finding, but with an extra derivative:

(a) $|x_{est} - x^*|$

(b) $|f'(x_{est}) - f'(x^*)| \approx \delta x|f''(x^*)|$

(c) $1/|f''(x^*)|$

2.5. (a) The range is $(-\infty, 0]$ since $\lim_{t\to 0}\log t = -\infty$ and $\log 1 = 0$.

(b) If the x_k is very negative, then e^{x_k} is exponentially close to zero. This near-zero value may not be representable, and regardless a single slightly larger value will dominate the sum.

(c) We simplify directly:

$$\begin{aligned}
\ell(x_1,\ldots,x_n) &= \ln\left[\sum_k e^{x_k}\right] \text{ by definition} \\
&= \ln\left[\sum_k e^{x_k-a+a}\right] \\
&= \ln\left[e^a \sum_k e^{x_k-a}\right] \\
&= \ln e^a + \ln\left[\sum_k e^{x_k-a}\right] \\
&= a + \ln\left[\sum_k e^{x_k-a}\right]
\end{aligned}$$

Suppose we take $a = \min_k x_k$. Then, rather than adding together tiny values we have moved the scale to be around $e^0 = 1$. (Other heuristics for choosing a are possible)

2.6. There are rendering artifacts because the two surfaces overlap and hence have the same depth values; rounding during depth computation can make one surface appear on top of the other. Possible resolutions include slightly offsetting one surface, adding a tie-breaking rule when depths are within some tolerance of each other, or merging the geometry before rendering to avoid overlap altogether.

2.7. (a) Recall that floating point arithmetic changes spacing as the order of magnitude of the value changes. Thus, it makes sense to have multiplicative error that is relative to the scale of x and y.

(b) (adapted from course notes by D. Bindel, Cornell CS) The recurrence for the ground-truth sum is simply $s_k = s_{k-1} + x_k y_k$. Error terms for the addition and multiplication steps show

$$\hat{s}_k = (\hat{s}_{k-1} + x_k y_k (1+\varepsilon_k^\times))(1+\varepsilon_k^+).$$

Subtracting the two shows:

$$\begin{aligned}
\hat{s}_k - s_k &= [(\hat{s}_{k-1} + x_k y_k (1+\varepsilon_k^\times))(1+\varepsilon_k^+)] - [s_{k-1} + x_k y_k] \text{ by the recurrences above} \\
&= [\hat{s}_{k-1} + \varepsilon_k^+ \hat{s}_{k-1} + x_k y_k (1+\varepsilon_k^\times) + x_k y_k \varepsilon_k^+ (1+\varepsilon_k^\times)] - [s_{k-1} + x_k y_k] \\
&= [\hat{s}_{k-1} - s_{k-1}] + \varepsilon_k^+ \hat{s}_{k-1} + x_k y_k (\varepsilon_k^\times + \varepsilon_k^+ + \varepsilon_k^+ \varepsilon_k^\times) \\
&= [\hat{s}_{k-1} - s_{k-1}](1+\varepsilon_k^+) + \varepsilon_k^+ s_{k-1} + x_k y_k (\varepsilon_k^\times + \varepsilon_k^+ + \varepsilon_k^+ \varepsilon_k^\times) \\
&= [\hat{s}_{k-1} - s_{k-1}](1+\varepsilon_k^+) + \varepsilon_k^+ s_k + x_k y_k (\varepsilon_k^\times + \varepsilon_k^+ \varepsilon_k^\times) \text{ since } s_k = s_{k-1} + x_k y_k \\
&= [\hat{s}_{k-1} - s_{k-1}](1+\varepsilon_k^+) + \varepsilon_k^+ s_k + x_k y_k \varepsilon_k^\times + x_k y_k \varepsilon_k^+ \varepsilon_k^\times
\end{aligned}$$

We can expand this inductively:

$$\begin{aligned}
\hat{s}_0 - s_0 &= 0 \\
\hat{s}_1 - s_1 &= [\hat{s}_0 - s_0](1+\varepsilon_1^+) + \varepsilon_1^+ x_1 y_1 + x_1 y_1 \varepsilon_1^\times + x_1 y_1 \varepsilon_1^+ \varepsilon_1^\times \\
&= x_1 y_1 (\varepsilon_1^+ + \varepsilon_1^\times) + x_1 y_1 \varepsilon_1^+ \varepsilon_1^\times \\
\hat{s}_2 - s_2 &= [\hat{s}_1 - s_1](1+\varepsilon_2^+) + \varepsilon_2^+ (x_1 y_1 + x_2 y_2) + x_2 y_2 \varepsilon_2^\times + x_2 y_2 \varepsilon_2^+ \varepsilon_2^\times \\
&= [x_1 y_1 (\varepsilon_1^+ + \varepsilon_1^\times) + x_1 y_1 \varepsilon_1^+ \varepsilon_1^\times](1+\varepsilon_2^+) + \varepsilon_2^+ (x_1 y_1 + x_2 y_2) + x_2 y_2 \varepsilon_2^\times + x_2 y_2 \varepsilon_2^+ \varepsilon_2^\times \\
&= x_1 y_1 (\varepsilon_1^+ + \varepsilon_1^\times + \varepsilon_1^+ \varepsilon_2^+ + \varepsilon_1^\times \varepsilon_2^+ + \varepsilon_2^+) + x_2 y_2 (\varepsilon_2^+ + \varepsilon_2^\times) + [x_1 y_1 \varepsilon_1^+ \varepsilon_1^\times + x_2 y_2 \varepsilon_2^+ \varepsilon_2^\times] + O(\varepsilon_{\max}^3) \\
&\;\;\vdots
\end{aligned}$$

Applying induction, this recurrence shows

$$\begin{aligned}
\hat{s}_k - s_k &= \sum_{i=1}^{k} \left[x_i y_i \left(\varepsilon_i^\times + \sum_{j=i}^{k} \varepsilon_j^+ \right) \right] + O(k\varepsilon_{\max}^2) \\
\implies e_n &\le n\varepsilon_{\max} \sum_k |x_k||y_k| + O(n\varepsilon_{\max}^2), \text{ as desired.}
\end{aligned}$$

2.8. For convenience, define $d \equiv x - y$. We'll start by simplifying the numerator of relative error and then substitute:

$$\begin{aligned}
(1+\varepsilon_x)x - (1+\varepsilon_y)y &= (x-y) + (\varepsilon_x x - \varepsilon_y y) \\
&= d + \varepsilon_x d + (\varepsilon_x - \varepsilon_y)y \\
\implies (1+\varepsilon_-)((1+\varepsilon_x)x - (1+\varepsilon_y)y) &= (1+\varepsilon_-)(d + \varepsilon_x d + (\varepsilon_x - \varepsilon_y)y) \\
&= (1+\varepsilon_-)d + \varepsilon_x(1+\varepsilon_-)d + (1+\varepsilon_-)(\varepsilon_x - \varepsilon_y)y \\
\implies E &= \left| \frac{(1+\varepsilon_-)((1+\varepsilon_x)x - (1+\varepsilon_y)y) - (x-y)}{x-y} \right| \\
&= \left| \frac{\varepsilon_- d + \varepsilon_x(1+\varepsilon_-)d + (1+\varepsilon_-)(\varepsilon_x - \varepsilon_y)y}{d} \right| \\
&= \left| \varepsilon_- + \varepsilon_x(1+\varepsilon_-) + (1+\varepsilon_-)(\varepsilon_x - \varepsilon_y)\frac{y}{d} \right|
\end{aligned}$$

This can be unbounded as $d \to 0$.

2.9. (a) Implicitly differentiating the relationship $0 = f(x(\varepsilon)) + \varepsilon p(x(\varepsilon))$ with respect to ε shows

$$\begin{aligned}
0 &= \frac{d}{d\varepsilon}[f(x(\varepsilon)) + \varepsilon p(x(\varepsilon))] \\
&= f'(x(\varepsilon))x'(\varepsilon) + p(x(\varepsilon)) + \varepsilon p'(x(\varepsilon))x'(\varepsilon) \text{ by the chain rule.}
\end{aligned}$$

Substituting $\varepsilon = 0$ and using $x^* = x(0)$ shows

$$0 = f(x^*)x'(0) + p(x^*) \implies x'(0) = -\frac{p(x^*)}{f(x^*)}.$$

(b) We differentiate

$$\begin{aligned}
f'(x) &= \frac{d}{dx}(x-1)\cdot(x-2)\cdot\cdots\cdot(x-20) \\
&= (x-2)\cdot\cdots\cdot(x-20) + (x-1)\cdot(x-3)\cdot\cdots\cdot(x-20) \\
&\quad + \cdots + (x-1)\cdot\cdots\cdot(x-19) \text{ by the product rule}
\end{aligned}$$

Substituting $x = j$ shows

$$f'(j) = (j-1)\cdot(j-2)\cdot\cdots\cdot(j-(j-1))\cdot(j-(j+1))\cdot\cdots\cdot(j-20)$$

For $p(x) = x^{19}$, from the previous part we have

$$x'(j) = -\frac{j^{19}}{(j-1)\cdot(j-2)\cdot\cdots\cdot(j-(j-1))\cdot(j-(j+1))\cdot\cdots\cdot(j-20)} = -\prod_{k\neq j}\frac{j}{j-k}.$$

(c) $x'(1) \approx 8.2 \times 10^{-18}$ and $x'(20) \approx -4.3 \times 10^{7}$; hence, the root $x^* = 1$ is far more stable.

2.10. (a) The alternative formula can be obtained by scaling the numerator and denominator of the quadratic equation:

$$\begin{aligned}
\frac{-b \pm \sqrt{b^2-4ac}}{2a} &= \frac{-b \pm \sqrt{b^2-4ac}}{2a} \cdot \frac{-b \mp \sqrt{b^2-4ac}}{-b \mp \sqrt{b^2-4ac}} \\
&= \frac{b^2 - (b^2-4ac)}{-2ab \mp 2a\sqrt{b^2-4ac}} \\
&= \frac{4ac}{-2ab \mp 2a\sqrt{b^2-4ac}} \\
&= \frac{-2c}{b \pm \sqrt{b^2-4ac}}
\end{aligned}$$

(b) When $b \leq 0$, take

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, x_2 = \frac{c}{ax_1},$$

and otherwise take

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, x_2 = \frac{c}{ax_2}.$$

This way, there never can be cancellation because we always move b farther from the origin in the numerator.

2.11. The bounds are worked out below:

$$[x] + [y] = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$$

$$[x] - [y] = [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$$

$$[x] \times [y] = \begin{cases} \text{value} & \text{sign}(\underline{x}) & \text{sign}(\overline{x}) & \text{sign}(\underline{y}) & \text{sign}(\overline{y}) \\ \hline \left[\underline{x}\underline{y}, \overline{x}\overline{y}\right] & + & + & + & + \\ \left[\overline{x}\underline{y}, \overline{x}\overline{y}\right] & + & + & - & + \\ \left[\underline{y}\overline{x}, \overline{y}\underline{x}\right] & + & + & - & - \\ \left[\underline{x}\overline{y}, \overline{x}\overline{y}\right] & - & + & + & + \\ \left[\min(\underline{x}\overline{y}, \underline{y}\overline{x}), \max(\underline{x}\underline{y}, \overline{x}\overline{y})\right] & - & + & - & + \\ \left[\overline{x}\underline{y}, \underline{x}\underline{y}\right] & - & + & - & - \\ \left[\underline{x}\overline{y}, \overline{x}\overline{y}\right] & - & - & + & + \\ \left[\underline{x}\overline{y}, \underline{x}\overline{y}\right] & - & - & - & + \\ \left[\overline{x}\overline{y}, \underline{x}\underline{y}\right] & - & - & - & - \end{cases}$$

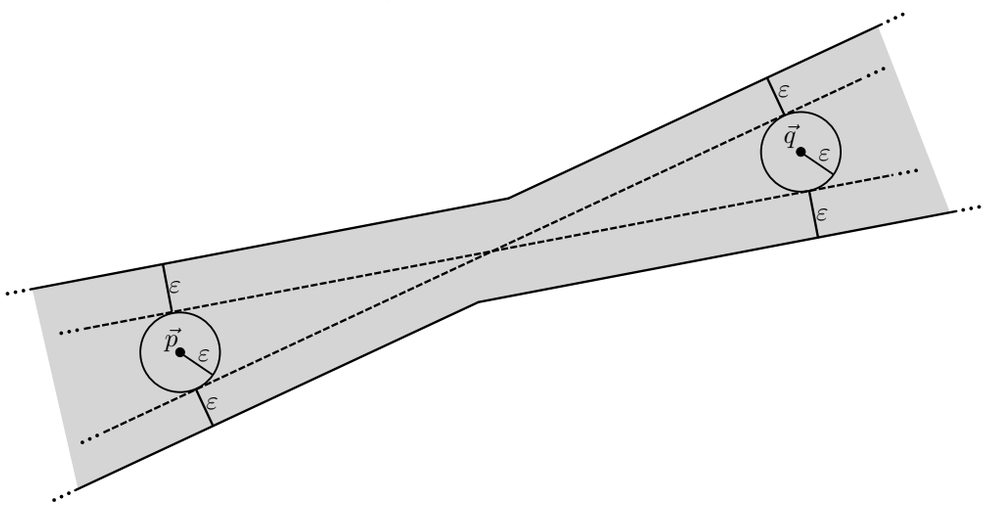
$$[x] \div [y] = [x] \times \left[\frac{1}{\overline{y}}, \frac{1}{\underline{y}}\right]$$

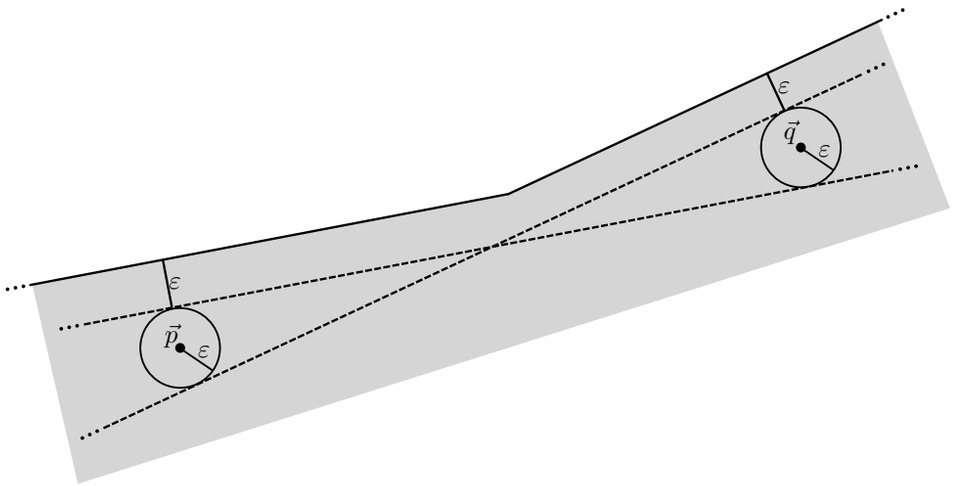
$$[x]^{1/2} = [\underline{x}^{1/2}, \overline{x}^{1/2}]$$

In finite-precision arithmetic, always round down the lower bounds and round up the upper bounds.

2.12. (a) Perturbing any of three collinear points slightly makes them not collinear. Furthermore, points may appear collinear if you zoom out far enough but appear less so as you zoom in.

(b)





(c)

(d) Obvious from drawings above; ε-collinear points form the intersection of four half-planes, two of which come from the ε-clockwise condition and two of which come from the ε-counterclockwise condition.

(e) No. See §3.1 of [55] for an example.

3.1. No; LU may not be possible for matrices requiring pivoting.

3.2. The steps of Gaussian elimination are below:

$$\left(\begin{array}{cc|c} 2 & 4 & 2 \\ 3 & 5 & 4 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 3 & 5 & 4 \end{array}\right), \text{ with elimination matrix } \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$$
$$\sim \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 1 & -1 \end{array}\right), \text{ with elimination matrix } \begin{pmatrix} 1 & 0 \\ 3 & -1 \end{pmatrix}$$
$$\sim \left(\begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & -1 \end{array}\right), \text{ with elimination matrix } \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$$

So, $x = 3$ and $y = -1$.

From the steps above, we know

$$U = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

and

$$L = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 3 & -1 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 3 & -1 \end{pmatrix}.$$

3.3. Computed using Gaussian elimination:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 6 & 11 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 1 & 2 & 7 \\ 0 & -1 & -22 \\ 0 & 0 & 204 \end{pmatrix}$$

3.4. Where it states "optionally insert pivoting code here," find row r with largest value in column p; then swap row r and row p of both A and $\vec{b}$.

3.5. No. Full pivoting can be preferable numerically but technically does not make a difference. The only way partial pivoting would fail is if there is an all-zero column, which would indicate that A is not invertible.

3.6. Write $A = A_1 + A_2 i$, $\vec{b} = \vec{b}_1 + \vec{b}_2 i$, and $\vec{x} = \vec{x}_1 + \vec{x}_2 i$. Then, $A\vec{x} = \vec{b} \implies (A_1 + A_2 i)(\vec{x}_1 + \vec{x}_2 i) = \vec{b}_1 + \vec{b}_2 i \implies (A_1\vec{x}_1 - A_2\vec{x}_2) + (A_2\vec{x}_1 + A_1\vec{x}_2)i = \vec{b}_1 + \vec{b}_2 i$. So, we can solve the block system

$$\begin{pmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix} = \begin{pmatrix} \vec{b}_1 \\ \vec{b}_2 \end{pmatrix}.$$

3.7. Carrying out Gaussian elimination is the same as pre-multiplying by the inverse of the leftmost $n \times n$ block. Hence, the output is $A^{-1}(A|I_{n\times n}) = (A^{-1}A|A^{-1}) = (I_{n\times n}|A^{-1})$.