Initial each of the following six (6) pages. For the sake of partial credit, you must show work and comment your solutions as appropriate. If you need space, write on the reverse side.

1. _______ (15 pts)

2. _______ (15 pts)

3. _______ (20 pts)

4. _______ (15 pts)

5. _______ (15 pts)

6. _______ (20 pts)

_______ (100 pts)
1. Assume that $x = 64$ and $y = 1/32$ are floating point numbers in a base-2 system that has $t$-bit mantissas and rounded arithmetic. What can you say about $t$ if $\text{fl}(x + y) = x$ and $\text{fl}(x - y) = x - y$? Here, $\text{fl}(z)$ is the nearest floating point number to $z$ with rounding away from zero in the event of a tie. Hint: Look at the spacing of the floating point numbers in the vicinity of $x$.

To the right of 64 the spacing is $2^{-t} \cdot 128 = 2^{7-t}$. It follows that

$$1/32 < (1/2)2^{7-t}$$

and so $1 < 2^{11-t}$. Thus $t < 11$.

To the left of 64 the spacing is $2^{-t} \cdot 64 = 2^{6-t}$. It follows that

$$1/32 \geq (1/2)2^{6-t}$$

and so $1 \geq 2^{10-t}$. Thus $t \geq 10$.

Thus, $t = 10$. 

2. (a) Gaussian elimination with partial pivoting (GEPP) computes a matrix factorization. Explain terms. (b) Suppose $A \in \mathbb{R}^{n \times n}$ is nonsingular and $b \in \mathbb{R}^n$. How would you use GEPP to compute $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ so that $Ax = b$ and $A^T y = b$?

(a) $PA = LU$ where $L$ is unit lower triangular, $U$ is upper triangular, and $P$ is a permutation chosen so that $|L| \leq \text{ones}(n, n)$.

(b) To get $x$ so that $Pb = PAx = LUx$ we solve $Ly = Pb$ for $y$ and $Ux = y$ for $x$. Still using the same factorization we note that $A^T P^T = U^T L^T$ and so $A^T P^T (Px) = b$ transforms to $(U^T L^T)z = b$ where $z = Px$. So we solve $U^T w = b$ for $w$, $L^T z = w$ for $z$, and set $x = P^T z$. 
3. Suppose $F \in \mathbb{R}^{m \times r}$ and $G \in \mathbb{R}^{n \times r}$ each have rank $r$ and that $r < \min\{m, n\}$. Assume that the $QR$ factorizations $F = Q_F R_F$ and $G = Q_G R_G$ have been computed. Show how to compute a minimizer of $\|FG^T x - b\|_2$ where $b \in \mathbb{R}^m$ is given and $x \in \mathbb{R}^n$ is to be determined. Note that this is a rank-deficient least squares problem so that there are an infinite number of minimizers. Your method should produce a solution $x_{LS}$ that has minimal 2-norm.

\[
\|FG^T x - b\| = \|Q_F R_F R_G Q_G^T x - b\| = \| R \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \| 
\]

where

\[
R = \begin{bmatrix} R_F(1:r, 1:r) & R_G(1:r, 1:r)^T \\ 0 & 0 \end{bmatrix}
\]

\[
Q_G^k x = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}
\]

\[
\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = Q_F^T b
\]

Here, $y_1, c_1 \in \mathbb{R}^r$. It follows that we solve $R_F(1:r, 1:r)R_G(1:r, 1:r)^T y_1 = c_1$ and set $y_2 = 0$. Note that the linear system solves involves two triangular system solves.
4. Suppose $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix and that we have computed a Schur decomposition 

$$Q^T A Q = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

Assume that $\lambda_1 = \lambda_2$. Show how to compute an orthogonal $\tilde{Q} = (\tilde{q}_{ij}) \in \mathbb{R}^{n \times n}$ so that 

$$\tilde{Q}^T A \tilde{Q} = \text{diag}(\lambda_1, \ldots, \lambda_n)$$

and $\tilde{q}_{11} = 0$. Hint: Update $Q$ with a Givens rotation in the $(1,2)$ plane.

Let $\tilde{Q} = Q G(1,2)$ where $G$ is the identity except $G(1:2,1:2) = [c \ s ; -s \ c]$. Note that $\tilde{q}_{11} = cq_{11} - sq_{12}$. Setting this to zero we see that we just set 

$$c = q_{12}/r \quad s = q_{11}/r \quad r = \sqrt{q_{11}^2 + q_{12}^2}$$
5. (a) How could Newton’s method be used to evaluate \( \arcsin(x) \) given the availability of the functions \( \sin(\cdot) \) and \( \cos(\cdot) \)? Recall that if \(|x| < 1\) and \( \arcsin(x) = \theta \), then \(-\pi/2 \leq \theta \leq \pi/2\) and \( \sin(\theta) = x \). (b) How could linear interpolation be used to obtain a starting value?

Use Newton to find a zero of \( f(\theta) = \sin(\theta) - x \):

\[
\theta_{n+1} = \theta_n - (\sin(\theta_n) - x) / \cos(\theta_n)
\]

Note that \( \theta = (\pi/2)x \) interpolates the \( \arcsin \) function at \( x = -1 \) and \( x = 1 \). So set \( \theta_0 = (\pi/2)x \).
6. The Euler and Backward Euler methods for the initial value problem $\dot{y} = f(y, t)$, $y(t_0) = y_0$ are given by

$$y_{n+1} = y_n + h_n f(y_n, t_n)$$

and

$$y_{n+1} = y_n + h_n f(y_{n+1}, t_{n+1})$$

Here, $y_n \approx y(t_n)$ and $t_{n+1} = t_n + h_n$. Consider the problem

$$B\dot{y} = Ay \quad y(0) = y_0$$

where $B, A \in \mathbb{R}^{m \times m}$ and $b$ is nonsingular. (a) For either method a linear system must solved to get $y_{n+1}$ from $y_n$. Explain. (b) What overheads are associated with changing step size?

Euler:

$$y_{n+1} = y_n + h_n B^{-1} Ay_n = (I + h_n B^{-1} A)y_n$$

so

$$By_{n+1} = By_n + h_n Ay_n$$

Thus, we need to factor $PB = LU$ once at the start of the iteration. Each step is then $O(m^2)$ flops.

Backwards Euler:

$$y_{n+1} = y_n + h_n B^{-1} Ay_{n+1}$$

so

$$By_{n+1} = By_n + h_n Ay_{n+1}$$

so

$$(B - h_n A)y_{n+1} = By_n$$

Need LU of $B - h_n A$.

If $h_n = h_{n-1}$ then no big deal in Euler—we still use the LU of $B$. But in backwards Euler changing $h$ means a new LU for $B - hA$ is required.