Sep 6, 2022 (Due: 08:00 Sep 13, 2022)

1. Let $\hat{x}$ be an approximation to x. In practice it is often much easier to estimate $\tilde{E}_{\text{rel}}(\hat{x}) = |x - \hat{x}|/|\hat{x}|$ compared to $E_{\text{rel}}(\hat{x}) = |x - \hat{x}|/|x|$. What is the relationship between E_{rel} and $\tilde{E}_{\text{rel}}$?

2. How to evaluate $f(x) = \tan x - \sin x$ for $x \approx 0$ such that numerical cancellation is avoided?

3. You are given $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$, both already stored in floating-point format. Show that there exists a "small" matrix $E \in \mathbb{R}^{m\times n}$ such that $\text{fl}(Ax) = (A + E)x$. Try to bound the entries of E as tight as you can. You may assume that there is no overflow or (gradual) underflow in the calculation.

4. Generate a random vector $x \in \mathbb{R}^2$. Visualize the relative error of

$$x^\top \begin{bmatrix} \cos\theta_k & \sin\theta_k \\ -\sin\theta_k & \cos\theta_k \end{bmatrix} x,$$

where $\theta_k = 2k\pi/2^n$ for $k = 0, 1, \ldots, 2^n - 1, 2^n$. What do you observe?

5. Evaluate the infinite series $\sum_{n=1}^{\infty} 1/n$ using IEEE single precision floating-point arithmetic. What do you observe?

(optional) Try different programming languages or different compiler optimization flags. Do you always obtain the same answer?