Exercise Sheet 1

IEEE-754 Floating Point Arithmetic & Linear Systems

1. IEEE-754 Properties

Answer which of the following statements unconditionally hold in IEEE-754 arithmetic, assuming that the variables a, b, c, d can take on *any finite* value.

True	**False**	
☐	☐	$(a+b)+(c+d)=(d+c)+(b+a)$.
☐	☐	$-a+a+a=a$.
☐	☐	$a \cdot a \geq 0$.
☐	☐	$(a \cdot b) \cdot c = a \cdot (b \cdot c)$
☐	☐	Denormalized (subnormal) numbers should be avoided in calculations as they cause numerical errors.
☐	☐	If $a < b$, then $a + c < b + c$.

2. Fondue Recipe

You're hosting a fondue party and want to create the perfect cheese blend. Traditional Swiss fondue uses a mix of cheeses, and you have access to three types:

- **Vacherin Fribourgeois**: 30% fat content, CHF 2.00 per 100g
- **Gruyère**: 32.5% fat content, CHF 2.10 per 100g
- **Appenzeller**: 31% fat content, CHF 2.40 per 100g

Your recipe requires:

- Exactly 600g of cheese total
- An average fat content of 31%
- A total cost of CHF 12.60

Let v, g, and a be the amounts (in grams) of Vacherin, Gruyère, and Appenzeller, respectively.

a) Explain why finding the perfect blend is equivalent to solving a linear system.

b) Write down the 3×3 matrix A and vector $\mathbf{b}$ such that $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x} = (v, g, a)^T$.

3. Exponential Function

The mathematical utility library of most programming languages (including NumPy) comes with a function named `expm1`, which computes

$$\texttt{expm1}(x) \coloneqq e^x - 1$$

in IEEE-754 floating point arithmetic. The manual of this function advises developers to

"replace expressions of the form $\texttt{exp}(x) - 1$ with $\texttt{expm1}(x)$."

What could be the purpose of such a specialized function? Is the advice always valid, or can we ignore it in some cases?