HW03 Hints

1. The problem can be re-phrased as

   > [...] Verify that the number in the computer which follows 8 is $8 + 8\,\boxed{\text{eps}}$ by numerically calculating $8 + 4\,\boxed{\text{eps}}$ and $8 + 4.01\,\boxed{\text{eps}}$. Also verify that the number in the computer which precedes 16 is $16 - 8\,\boxed{\text{eps}}$ by numerically calculating $16 - 4.01\,\boxed{\text{eps}}$ and $16 - 4\varepsilon$. Also, do the same for $2^{10} = 1024$.

   - Mimic the relevant examples shown on p. 9 of Module 2 lecture slides.
   - Be sure to read and follow the **Warning (this part only)** below the problem.
   - For 2^{10}, it is your task to determine what numbers of the form $1024 + ?\,\boxed{\text{eps}}$ are to be calculated.

2. First off, there is a typo in the problem.

   Typo: In the second line of Equation (9.25a), change "if $x = 1$" to "if $x = 0$".

   Answers to LM 9.3–10. To give a further clarification of the problem, here we present solutions to a similar problem in which the function

$$f(x) = \begin{cases} \dfrac{e^x - 1}{x} & \text{if } x \neq 0 \\ 1 & \text{if } x = 0, \end{cases}$$

   is considered for small x.

   (a) From the Maclaurin series[1] expansion of e^x

$$e^x = 1 + x + \frac{1}{2}x^2 + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!},$$

   it follows that

$$\begin{aligned} \lim_{x \to 0} f(x) &= \lim_{x \to 0} \frac{(1 + x + x^2/2 + x^3/6 + \cdots) - 1}{x} \\ &= \lim_{x \to 0} \frac{x + x^2/2 + x^3/6 + \cdots}{x} \\ &= \lim_{x \to 0} \left(1 + \frac{1}{2}x + \frac{1}{6}x^2 + \cdots\right) = 1. \end{aligned}$$

[1] That is, the Taylor series centered at 0.

(b) In what follows, note the use of the elementwise division `./`. Also recall that "log", in this class and in MATLAB, denotes the natural logarithmic[2] function. Lastly, pay attention to the vectorization; it is highly recommended that you proceed similarly.

```
k = [1:20]';
x = 10.^(-k);
fx = (exp(x) - 1)./x;              %    (i)    f(x)
f1x = (exp(x) - 1)./log(exp(x));  %   (ii)  f_1(x)
f2x = expm1(x)./x;                 %  (iii)  f_2(x)
format long e
disp([x fx f1x f2x])               % or use fprintf to suit your taste
```

Note. The explanation below frequently uses Calc 2 stuffs (power series). Be sure to brush up on those prerequisites!

Explanation. For small x, the evaluation of the expression $y = e^x - 1$ suffers from catastrophic cancellation because $e^x \approx 1$. This explains why the numerical evaluation of $f(x)$ is inaccurate for small x.

To understand why $f_1(x)$ is doing better, denote by $\hat{y}$ the floating-point representation of the numerator $y = e^x - 1$. As mentioned above, due to catastrophic cancellation, many significant digits are lost in $\hat{y}$ for small x.

Now, the Taylor expansion for the denominator $\log e^x = \log(1 + (e^x - 1)) = \log(1 + y)$, written in terms of y, is

$$\log e^x = y - \frac{1}{2}y^2 + \frac{1}{3}y^3 - \cdots ,$$

and so its numerical evaluation can be approximated by

$$\widehat{\log e^x} = \hat{y} - \frac{1}{2}\hat{y}^2 + \frac{1}{3}\hat{y}^3 - \cdots ,$$

which involves an error as well. Nonetheless, when both are put together,

$$\begin{aligned}\underbrace{\frac{e^x - 1}{\log e^x}}_{\text{analytical expression}} &\approx \frac{\hat{y}}{\hat{y} - \hat{y}^2/2 + \hat{y}^3/3 - \cdots} \\ &= \underbrace{1 + (\text{tiny higher-order terms})}_{\text{numerical evaluation}} .\end{aligned}$$

yielding the correct asymptotic behavior for small y, in turn, for small x. This explains $f_1(x)$ results in plausible results until when the evaluation of y loses all the significant digits, *i.e.*, when $\hat{y} = 0$. In the script above, this happens when $k = 16$, at which point $\widehat{\log e^x} = 0$, *i.e.*, the denominator also evaluates to 0, resulting in `NaN`.

The `expm1` function[3] was designed to avoid catastrophic cancellation in the calculation of $e^x - 1$ for small x; type `help expm1`. Hence, $f_2(x)$ is evaluated to the full double-precision for all x values used.

3. This is not a hint *per se*, but an explanation, for those curious, of why the proposed formula ($\star$) for $\operatorname{acosh}(x)$

$$\log(x - \sqrt{x^2 - 1})$$

[2] The natural logarithmic function is commonly denoted by "ln", *e.g.*, in calculus.

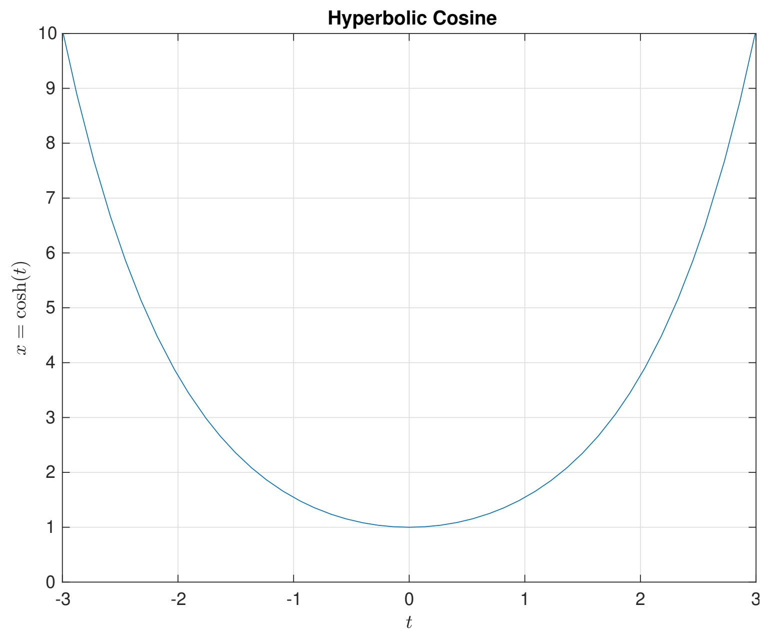
[3] The name comes from $e^x - 1$ (`exp` of x minus `1`).

is different from what is presented in typical calculus textbooks or online sources[4].

One can readily confirm from the definition

$$\cosh(t) = \frac{e^t + e^{-t}}{2}$$

that the hyperbolic cosine function is an even function. Consequently, it is not *one-to-one* over its domain, $\mathbb{R}$, and so cannot be inverted entirely.



The usual workaround is to invert $\cosh(t)$ only on $[0, \infty)$ over which $\cosh(t)$ is one-to-one, and the resulting formula is conventionally regarded as *the* inverse hyperbolic cosine function. However, to handle cases where $t < 0$ such as ours (because we set `t=-4:-4:-16`), $\cosh(t)$ must be inverted over $(-\infty, 0]$ and the result is the formula given in the problem.

Exercise. Confirm that the inverse of the function $x = \cosh(t) = (e^t + e^{-t})/2$ on $(-\infty, 0]$ is

$$t = \log(x - \sqrt{x^2 - 1}), \quad x \in [1, \infty).$$

[4] For example, see
https://en.wikipedia.org/wiki/Inverse_hyperbolic_functions#Inverse_hyperbolic_cosine