1 Problem 2-1

Give the binary format of $-528$ as a double floating-point number.

2 Problem 2-2

Answer the following questions, and show your experiments with C language and GCC compiler to check your arguments.

(a) In a regular program, which is the representation of $0.0$? $+0.0$ or $-0.0$ .

Please find the statement in the manual


that supports your answer.

(b) How to specifically assign $+0.0$ and $-0.0$ ?

(c) How to use subroutines to get the sign of a number?

(d) Suppose

$a < 0$ and $|b| < B$, 

where $B$ is a positive number that is not too large or small, how do we guarantee

$a / \max(b, 0.0) < 0$

always holds? Hint: $\max(b, 0.0)$ may return $-0.0$ .

(e) What formulation should we use for the “max” function?

$(x > y) ? x : y$

or

$(x < y) ? y : x$

Hint: Your “max” function need to return $+0.0$ but not $-0.0$ .