EX.0.3.10, Sauer

Decide whether $\mathrm{fl}(1 + x) > 1$ in double precision floating point arithmetic, with Rounding to Nearest.

a. $x = 2^{-53}$

b. $x = 2^{-53} + 2^{-60}$

EX.0.3.10, Sauer, solution, Langou

For double precision floating point, the machine epsilon is $\epsilon = 2^{-52}$. In (a), $x = 2^{-53}$, and thus $x + 1$ is an exceptional case for IEEE floating point rounding. $1 + 2^{-53}$ is in the exact middle of its two nearest floating point numbers $x_{\min} = 1$ and $x_{\max} = 1 + 2^{-52}$. In this case, this is a draw and $x_{\min} = 1$ and $x_{\max} = 1 + 2^{-52}$ are equally near. The rule tells us that we should pick the one which 52nd bit is a 0. Since

$$\begin{aligned} x_{\min} = 1 \quad &= \quad (1.0000000000000000000000000000000000000000000000000000)_2 \\ x_{\max} = 1 + 2^{-52} \quad &= \quad (1.0000000000000000000000000000000000000000000000000001)_2 \end{aligned}$$

we have

$$\mathrm{fl}(1 + 2^{-53}) = 1,$$

thus $\mathrm{fl}(1 + x) > 1$ is false for (a).
In (b),

$$x_{\min} = 1 < 1 + 2^{-53} + 2^{-60} < x_{\max} = 1 + 2^{-52}.$$

However $x_{\max}$ is closer from $1 + x$ than $x_{\min}$ so

$$\mathrm{fl}(1 + x) = x_{\max} = 1 + 2^{-52},$$

and thus $\mathrm{fl}(1 + x) > 1$ is true for part (b).