Notes:

- Read Course Information: Section 6 (Miscellaneous) and Section 8 (Academic Dishonesty or Misconduct).
- When you are giving a construction, example, etc., provide a justification with your argument. Your solutions to numerical problems must contain the derivation of your answers. In all of your presentations, strive for correctness, completeness, and clarity. When in doubt about the assumptions of problems, the interpretations of wording, etc., consult the instructor.
- You should strive to complete all problems assigned, and a subset of them will be graded.

1. Read the notes above carefully.
2. Do [Cha10] Chapter 2, Section 2.1, exercises 1 and 4.
3. Do [Cha10] Chapter 2, Section 2.2, exercises 5 and 9.
5. For this problem, we represent real numbers by a modified floating point number system with the parameters (radix $b = 2$, precision $s$, range of exponent $[m, M]$), in which a modified floating point is of the form:

   \[ (-1)^{\text{sign}} \times 1.f \times 2^e \]

   where sign is either 0 or 1, $f$ is an $s$-bit binary integer (note that the normalization is achieved at “1.”), and $e \in [n, M]$. For example, the real number (in decimal) $(-4.125)_{10} = -(2^2 + 2^{-3})$ is represented by the modified floating point number (with $s = 6$) $(-1)^1 \times 1.000010 \times 2^2$.

   (a) Assume that the truncation supporting the modified floating point system are “round down” and “round up”. For a real number $q$ in the form of $(-1)^{\text{sign}} \times 1.a_1a_2\cdots a_s a_{s+1}\cdots \times 2^e$, (1) the round-down of $q$, denoted by $q_-$, is the modified floating point number $(-1)^{\text{sign}} \times 1.a_1a_2\cdots a_s \times 2^e$ (by simply chopping off all the bits after the $s$-th one), and (2) the round-up of $q$, denoted by $q_+$, is the modified floating point number $(-1)^{\text{sign}} \times 1.a_1a_2\cdots a_s \times 2^e + 2^{-s}$ (by effectively adding the last/$s$-th bit of $q_-$ by 1).

   Assume that the modified floating point system is intelligent enough to decide which rounding, $q_-$ or $q_+$, is closest to $q$ — denote it by $q^*$.

   Give the best/tightest “error bound” for $q^*$ (that is, for $|q - q^*|$). Also, give the corresponding “absolute-error bound” and “relative-error bound”. Justify/explain your answers.

   (b) …

6. … More problem(s) will be given in later version.