APC523 - Homework #1

1)

If you do symmetric rounding

case 1, round down (ie $b_{p+1}$ is zero)

$$x = rd(x) + \varepsilon_{rd} = \sum_{i=1}^{p} b_i 2^{-i} 2^e + \sum_{i=p+2}^{\infty} b_i 2^{-i} 2^e$$

then the absolute error $|x - \varepsilon_{rd}| = \left| \sum_{i=p+2}^{\infty} b_i 2^{-i} 2^e \right|$

the max absolute error would occur if all the rest of the bits are 1 (except the one immediately following where we rounded

$$\max\left| \sum_{i=p+2} 2^{-i} 2^e b_i \right| = \max\left| \sum_{i=p+2} 2^{-i} 2^e \right|$$

$$= \left|(2^{-p-2} + 2^{-p-3} + \dots) 2^e\right| = \left|2^{-p-1} \underbrace{(2^{-1} + 2^{-2} + \dots)}_{\text{infinite series}} 2^e\right|$$

$$= \frac{a_1}{1-r} = \frac{1/2}{1-1/2} = 1$$

$\max |x - \varepsilon_{rd}| = |2^{-p-1} 2^e|$

now the relative error $= \dfrac{\max |x - rd(x)|}{\min |x|}$

$\min |x| = 2^{-1} 2^e$

then max rel error $= \dfrac{2^{-p-1} 2^e}{2^{-1} 2^e} = 2^{-p}$

case 2, you round up

$$x = rd(x) + \varepsilon$$

$$= \sum_{i=1}^{p} b_i 2^{-i} 2^e - \underbrace{2^{-p} 2^e}_{\text{added due to rounding}} + \sum_{i=p+1}^{\infty} b_i 2^{-i} 2^e$$

$$|x - rd(x)| = \left| \sum_{i=p+1}^{\infty} b_i 2^{-i} 2^e - 2^{-p} 2^e \right| = \left| 2^{-p-1} 2^e + \sum_{i=p+2}^{\infty} b_i 2^{-i} 2^e - 2^{-p} 2^e \right|$$

since the first digit of the residual is 1

$$\left| \left( \frac{2^{-p}}{2} + \sum_{i=p+2}^{\infty} b_i 2^{-i} - 2^{-p} \right) 2^e \right| = \left| \left( \frac{-2^{-p}}{2} + \underbrace{\sum_{i=p+2}^{\infty} b_i 2^{-i}}_{0 \le \sum b_i 2^{-i} \le 1} \right) 2^e \right|$$

to get the max value we set the sum to zero

$$\left| \frac{x - rd(x)}{x} \right| = \frac{2^{-p-1} 2^e}{2^{-1} 2^e} = 2^{-p}$$

then we get a max rel error of $2^{-p}$ in both cases