APC 523 — HW1

Eric Moseley

1. Let $x = \left[\sum_{\ell=1}^{\infty} b_{-\ell} 2^{-\ell}\right] 2^e$

If we round at the p+1 term, and we let $\ell' = \ell - p - 1$, let's consider the case where $b_{p+1} = 1$. How does this differ from truncating? we add a term of size 2^{-p-1}.

The case of truncating had

$$x - \text{trunc}(x) = \pm \left(\sum_{\ell=p+1}^{\infty} b_{-\ell} 2^{-\ell} \right) 2^e$$

Then, if $\ell' = \ell - p - 1$,

$$x - \text{trunc}(x) = \pm \left(\sum_{\ell'=0}^{\infty} b_{-\ell'-p-1} 2^{-\ell'} \right) 2^{e-p-1}$$

But we know that we've done better than this with rnd(x) by adding 2^{-p-1} to trunc(x)

$$\Rightarrow \quad x - \text{rnd}(x) = \pm \left(\sum b_{-\ell'-p-1} 2^{-\ell'} \right) 2^{e-p-1} - 2^{-p-1+e}$$

$$= \left(\underbrace{b_{-p+1)}}_{1} - 1 + \sum_{\ell'=1}^{\infty} b_{-\ell'-p-1} 2^{-\ell'} \right) 2^{e-p-1}$$

2^{-k-1}

Now shift again to some $k = \ell' + 1$

$$x - \text{rnd}(x) = \pm \left(\sum_{k=0}^{\infty} b_{-k-p-2} 2^{-k} \right) 2^{e-p-2}$$

the largest that series can be is $2. \Rightarrow$

$$\Rightarrow \max \| x - \text{rnd}(x) \| = 2^{e-p-1}$$

the smallest x can be if $x = \left[\sum_{\ell=1}^{\infty} b_{-\ell} 2^{-\ell}\right] 2^e$

is 2^{e-1}, since the leading term is always 1.

Thus, when rounding up, the error is

$$\boxed{\left| \frac{x - \text{rnd}(x)}{x} \right| \leq 2^{e-p-1} / 2^{e-1} = 2^{-p}}$$

When rounding down, we subtract a term instead that is atmost of $\mathcal{O}(2^{-p-2})$, this is the same as truncating the series in this case.

$$x - \text{trunc}(x) = x - \text{rnd}(x) = \pm \left(\sum_{\ell=p+1}^{\infty} b_{-\ell} 2^{-\ell} \right) 2^{e}$$

but for the $p+1$ term, we have $b_{-p-1} = 0$. Therefore, we have, in the worst case scenario, an effective $p \to p+1$. Thus, in our formula for the error of trunc(x), if we just make this replacement, we have our answer. That gives us:

$$\boxed{\left| \frac{x - \text{rnd}(x)}{x} \right| \leq 2 \, 2^{-p-1} = 2^{-p}}$$

And that covers both cases.