HW 4
EGEE 518 Digital Signal Processing I
Fall 2008
California State University, Fullerton

Nasser M. Abbasi

May 29, 2019 Compiled on May 29, 2019 at 5:29pm

Contents

1 my solution, First Problem

Looking at 2 floating points problems. The first to illustrate the problem when adding large number to small number. The second to illustrate the problem of subtracting 2 numbers close to each others in magnitude.

Investigate floating point errors generated by the following sum $\sum\limits_{n=1}^{N} \frac{1}{n^2}$, compare the result to that due summation in forward and in reverse directions.

1.1 Analysis

When performing the sum in the forward direction, as in $1 + \frac{1}{4} + \frac{1}{16} + \cdots + \frac{1}{N^2}$ we observe that very quickly into the sum, we will be adding relatively large quantity to a very small quantity. Adding a large number of a very small number leads to loss of digits as was discussed in last lecture. However, we adding in reverse order, as in $\frac{1}{N^2} + \frac{1}{(N-1)^2} + \frac{1}{(N-2)^2} + \cdots + 1$, we see that we will be adding, each time, 2 quantities that are relatively close to each other in magnitude. This reduces floating point errors.

The following code and results generated confirms the above. $N = 20,000$ was used. The computation was forced to be in single precision to be able to better illustrate the problem.

1.2 Computation and Results

This program prints the result of the sum in the forward direction

```
      PROGRAM main
      IMPLICIT NONE
      REAL :: s
      INTEGER :: n,MAX

      s  = 0.0;
      MAX = 20000;
      DO n = 1,MAX
         s = s + (1./n**2);
      END DO

      WRITE(*,1) s
1     format('sum = ', F8.6)
      END PROGRAM main


sum = 1.644725
```

now compare the above result with that when performing the sum in the reverse direction

```
PROGRAM main
      IMPLICIT NONE
      REAL :: s
      INTEGER :: n,MAX

      s  = 0.0;
      MAX = 20000;
      DO n = MAX,1,-1
         s = s + (1./n**2);
      END DO

      WRITE(*,1) s
1     format('sum = ', F8.6)
      END PROGRAM main

sum = 1.644884
```

The result from the reverse direction sum is the more accurate result. To proof this, we can use double precision and will see that the sum resulting from double precision agrees with the digits from the above result when using reverse direction sum

```
      PROGRAM main
      IMPLICIT NONE
      DOUBLE PRECISION :: s
      INTEGER :: n,MAX

      s  = 0.0;
      MAX = 20000;
      DO n = 1,MAX
         s = s + (1./n**2);
      END DO

      WRITE(*,1) s
1     format('sum = ', F18.16)
      END PROGRAM main

sum = 1.6448840680982091
```

1.3 Conclusion

In floating point arithmetic, avoid adding a large number to a very small number as this results in loss of digits of the small number. The above trick illustrate one way to accomplish this and still perform the required computation.

In the above, there was $1.644884 - 1.644725 = 1.\,59 \times 10^{-4}$ error in the sum when it was done in the forward direction as compared to the reverse direction (for $20,000$ steps).In relative term, this error is $\frac{1.644884-1.644725}{1.644884}100$ which is about 0.01% relative error.

2 my solution, second problem

Investigate the problem when subtracting 2 numbers which are close in magnitude. If a, b are 2 numbers close to each others, then instead of doing $a-b$ do the following $(a-b)\frac{(a+b)}{(a+b)} = \frac{a^2-b^2}{a+b}$. The following program attempts to illustrate this by comparing result from $a-b$ to that from $\frac{a^2-b^2}{a+b}$ for 2 numbers close to each others.

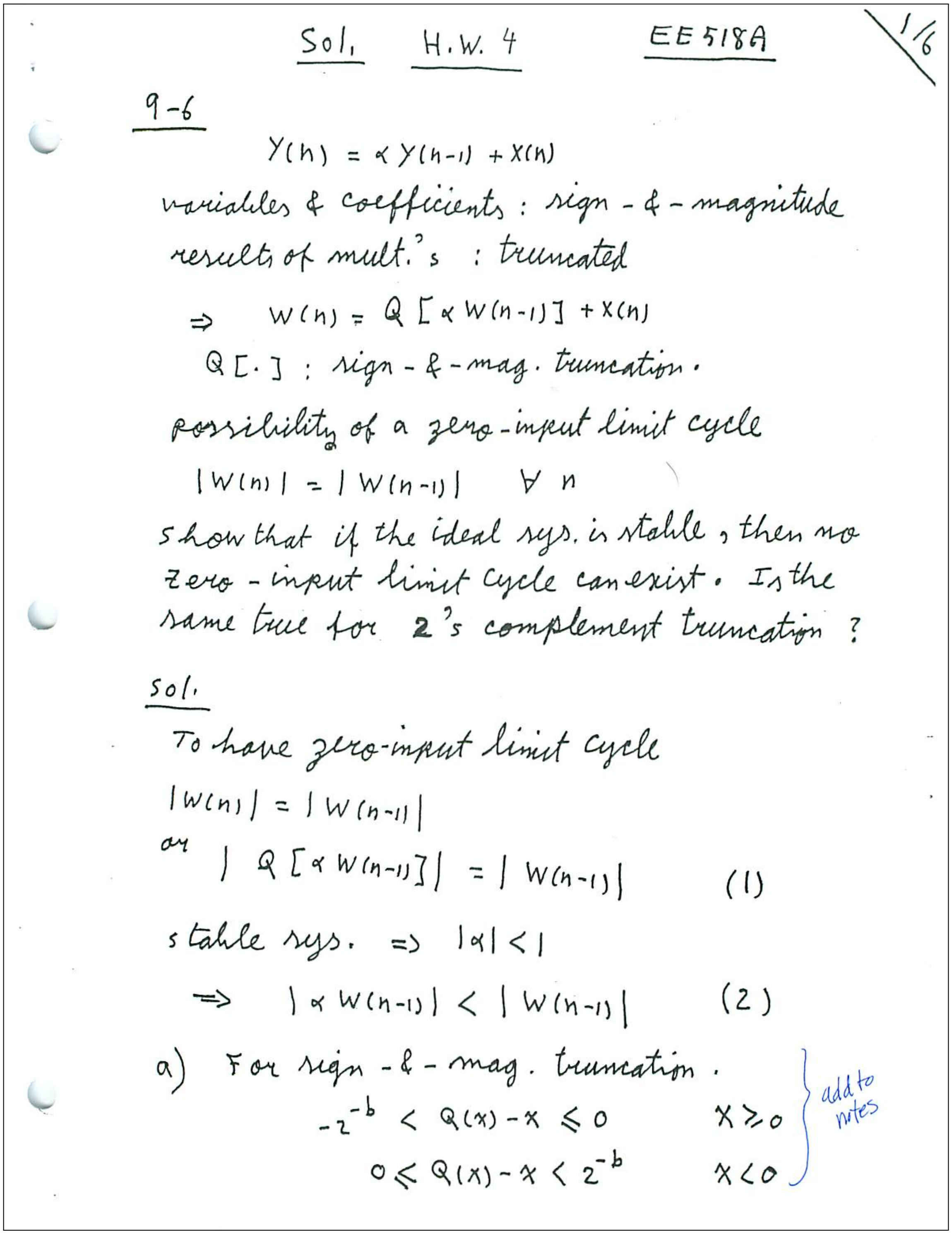
```
      PROGRAM main
      IMPLICIT NONE
      DOUBLE PRECISION :: a,b,diff

      a = 32.000008;
      b = 32.000002;
      diff = a-b;
      WRITE(*,1), diff
      diff = (a**2-b**2)/(a+b);
      WRITE(*,1), diff
1     format('diff = ', F18.16)
      END PROGRAM main

diff = 0.0000038146972656
diff = 0.0000038146972656
```

I need to look more into this as I am not getting the right 2 numbers to show this problem.

3 key solution



Sol. H.W. 4 EE518A

1/6

9-6

$$y(n) = \alpha y(n-1) + x(n)$$

variables & coefficients : sign - & - magnitude

results of mult.'s : truncated

$$\Rightarrow \quad w(n) = Q[\alpha w(n-1)] + x(n)$$

$Q[\cdot]$: sign - & - mag. truncation.

possibility of a zero-input limit cycle

$$|w(n)| = |w(n-1)| \quad \forall n$$

show that if the ideal sys. is stable, then no zero-input limit cycle can exist. Is the same true for 2's complement truncation?

Sol.

To have zero-input limit cycle

$$|w(n)| = |w(n-1)|$$

or

$$|Q[\alpha w(n-1)]| = |w(n-1)| \tag{1}$$

stable sys. $\Rightarrow |\alpha| < 1$

$$\Rightarrow \quad |\alpha w(n-1)| < |w(n-1)| \tag{2}$$

a) For sign - & - mag. truncation.

$$\left.\begin{array}{ll} -2^{-b} < Q(x) - x \leq 0 & x \geq 0 \\ 0 \leq Q(x) - x < 2^{-b} & x < 0 \end{array}\right\} \text{add to notes}$$

2/6

$\Rightarrow |Q(x)| \leqslant |x|$ for $x \geqslant 0$ or $x < 0$

Let $x = \alpha W(n-1)$

$\Rightarrow |Q[\alpha W(n-1)]| \leqslant |\alpha W(n-1)|$ (3)

(3) & (2) $\Rightarrow |Q[\alpha W(n-1)]| \leqslant |\alpha W(n-1)| < |W(n-1)|$

Since (1) is not satisfied no zero input limit cycle is possible.

b) For $Q[\cdot]$ = two's complement

$$-2^{-b} < Q(x) - x \leqslant 0 \quad \forall x$$

If $\underline{x > 0}$ $\quad x \geqslant Q[x]$ or $|x| \geqslant |Q[x]|$ (4)

If $\underline{x < 0}$ $\quad |Q[x]| \geqslant |x|$ (5)

For $\alpha W(n-1) > 0$

$$|Q[\alpha W(n-1)]| \underset{④}{\leqslant} |\alpha W(n-1)| \underset{②}{<} |W(n-1)|$$

$\Rightarrow$ no limit cycle : (1) is not satisfied

For $\alpha W(n-1) < 0$

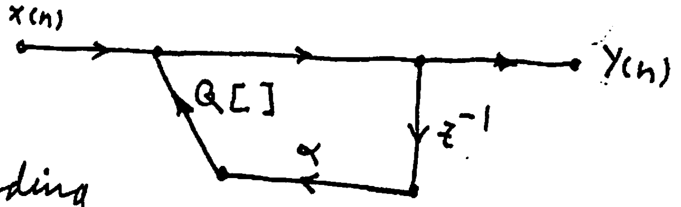
$|\alpha W(n-1)| \leqslant |Q[\alpha W(n-1)]|$ by (5)

and $|\alpha W(n-1)| < |W(n-1)|$ by (2)

Possible that $|Q[\alpha W(n-1)]| = |W(n-1)|$ for $\alpha W(n-1) < 0$ $\Rightarrow$ limit cycle

3/1

9-7



Q[ ] : rounding

Fixed-pt. fractions, b bits

zero input – $Y(-1) = A$ initial cond.

Dead band : $A \ni |Q[\alpha A]| = A$

a) dead band in terms of α and β

b) For $b=6$, $A=1/16$ sketch $Y(n)$ for $\alpha = \begin{cases} 15/16 \\ -15/16 \end{cases}$

c) For $b=6$, $A=1/2$ sketch $Y(n)$ for $\alpha = -15/16$

Sol.

$$Y(n) = Q[\alpha Y(n-1)] + X(n) \qquad (X(n)=0)$$

Rounding : $-\frac{2^{-b}}{2} < Q[\alpha W(n-1)] - \alpha W(n-1) \leq \frac{2^{-b}}{2}$

If filter is in the dead band

$$-\frac{2^{-b}}{2} < Q[\alpha A] - \alpha A \leq \frac{2^{-b}}{2}$$

or $|Q[\alpha A] - \alpha A| \leq \frac{2^{-b}}{2}$

In a limit cycle $|Q[\alpha A]| = A$

$$\Rightarrow |Q[\alpha A]| - |\alpha A| \leq |Q[\alpha A] - \alpha A| \leq \frac{1}{2} 2^{-b}$$

$$\Rightarrow |A| - |\alpha||A| \leq \frac{1}{2} 2^{-b}$$

4.1

$$\Rightarrow |A| \leqslant \frac{\frac{1}{2}\, 2^{-b}}{1-|\alpha|}$$

b) $b=6 \quad 2^{-b} = 1/64 \quad |\alpha| = 15/16 \qquad 1-|\alpha| = 1/16$

$$|A| \leqslant \frac{\frac{1}{2}\cdot\frac{1}{64}}{\frac{1}{16}} = 1/8$$ <u>dead band</u>

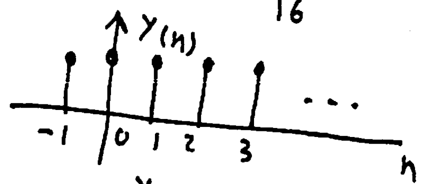
Thus for $A = 1/16$ the system starts immediately in the limit cycle.

$\alpha = \frac{15}{16} \quad Y(n) = Q\left[\alpha\, Y(n-1)\right] = Q\left[\frac{15}{16}\cdot\frac{1}{16}\right] = Q\left[\frac{15}{256}\right] = 1$

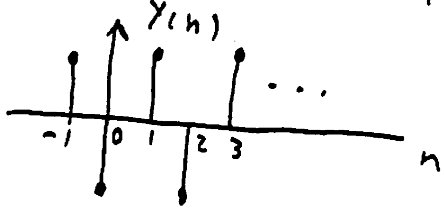
rounding!

$\alpha = \frac{-15}{16} \quad Y(n) = Q\left[-\frac{15}{16}\cdot\frac{1}{16}\right] = \begin{cases} \frac{-1}{16} & n \text{ even} \\ \frac{1}{16} & n \text{ odd} \end{cases}$

$Y(-1) = \frac{1}{16}$



$\alpha = 15/16$

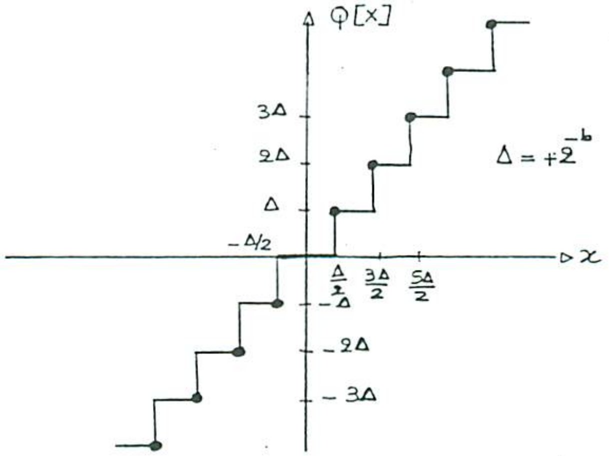


$\alpha = -15/16$

c) $b=6 \quad A = 1/2 \quad \alpha = \frac{-15}{16}$ $\Rightarrow$ same dead band

$$Y(n) = Q\left[\frac{-15}{16}\cdot Y(n-1)\right]$$

5/6



Thus we have: $W(0) = Q\left[-\frac{1}{2}\frac{15}{16}\right] = Q\left[-59\frac{\Delta}{2} - \frac{\Delta}{2}\right] = -30\Delta$

$$W(1) = Q\left[\frac{15}{16}.30\Delta\right] = Q\left[56\frac{\Delta}{2} + \frac{1}{4}\frac{\Delta}{2}\right] = 28\Delta$$

Hence we repeat the above procedure and we get:

$W(-1) = 32/64$

$W(0) = -30/64$

$W(1) = 28/64$

$W(2) = -26/64$

$W(3) = 24/64$ ← rounding up ↙ ~~$Q\left[\frac{-52.5}{128}\right] \neq$~~

$W(4) = -23/64$ → Round down!

$Q\left[\frac{24.37}{64}\right]$!

$W(5) = 22/64$

$W(6) = -21/64$

$W(7) = 20/64$

$W(8) = -19/64$

$W(9) = 18/64$

$W(10) = -17/64$

$W(11) = 16/64$

$W(12) = -15/64$

$W(13) = 14/64$

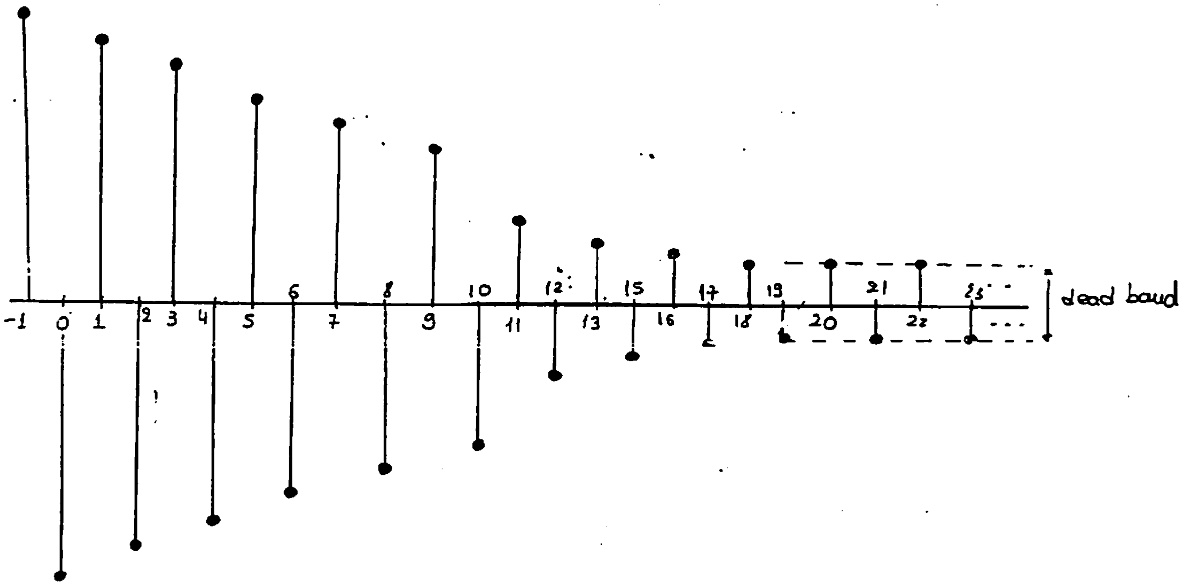
$W(14) = -13/64$

$W(15) = 12/64$

8/8

$$W(16) = -11/64$$
$$w(17) = 10/64$$
$$W(18) = -9/64$$
$$W(19) = 8/64 \quad \longleftarrow \text{ rounding up}$$
$$W(20) = -8/64$$
$$W(21) = 8/64$$
$$W(22) = -8/64$$

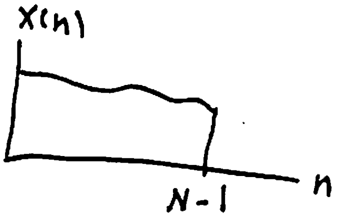
The output will be:



11.1 H.W. 5 sol. EE518A 1/2

$$C_{xx}(m) = \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)\,x(n+m) \qquad |m| \leq N-1$$

show that



$$I_N(w) = \frac{1}{N}\left|X(e^{jw})\right|^2$$

$$I_N(w) = \sum_{m=-(N-1)}^{N-1} C_{xx}(m)\, e^{-jwm}$$

sol.

$$C_{xx}(m) = \frac{1}{N}\, x(n) * x(-n)$$

$$x(-n) \xrightarrow{F.T.} X(e^{-jw}) = X^*(e^{jw})$$ For x(n) real

$$\Rightarrow I_N(e^{jw}) = \frac{1}{N}\, X(e^{jw})\, X^*(e^{jw}) = \frac{1}{N}\left|X(e^{jw})\right|^2$$

or

$$I_N(w) \triangleq \sum_{m=-(N-1)}^{N-1} C_{xx}(m)\, e^{-jwm}$$

$$= \sum_{m=-(N-1)}^{N-1}\left[\frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)\,x(n+m)\right] e^{-jwm}$$

$$= \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n) \sum_{m=-(N-1)}^{N-1} x(n+m)\, e^{-jwm}$$

$$= \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n) \sum_{\ell=n-(N-1)}^{n+(N-1)} x(\ell)\, e^{-jw\ell}\, e^{jwn}$$

$$= \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)\, e^{jwn} \sum_{\ell=n-(N-1)}^{n+(N-1)} x(\ell)\, e^{-jw\ell}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1} x(n)\, e^{jwn} \sum_{\ell=0}^{N-1} x(\ell)\, e^{-jw\ell}$$ since $x(n)=0$ for $n<0$ & $n \geq N$

2.1

$$I_N(w) = \frac{1}{N}\left[\sum_{n=0}^{N-1} x(n)\, e^{-jwn}\right]^* \sum_{\ell=0}^{N-1} x(\ell)\, e^{-jw\ell}$$

$$= \frac{1}{N}\left|X(e^{jw})\right|^2$$

$\underline{11.2}|$ $$S_{xx}(w) = \sum_{m=-(M-1)}^{M-1} C_{xx}(m)\, W(m)\, e^{-jwm}$$

$w(m)$ of length $2M-1$

show that $$E\{S_{xx}(w)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} E\{I_N(\theta)\}\, \mathcal{W}(e^{j(w-\theta)})\, d\theta$$

$$\begin{cases} W(m) = 0 & |m| \geq 2M \\ C_{xx}(m) = 0 \text{ for } & |m| \geq M \end{cases}$$

Knowing these we can say

$$S_{xx}(w) = \sum_{m=-\infty}^{\infty} C_{xx}(m)\, W(m)\, e^{-jwm}$$

$$= \mathcal{F}\{C_{xx}(m)\, W(m)\}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{F}\{C_{xx}(m)\}\, \mathcal{W}(e^{j(w-\theta)})\, d\theta \quad \text{conv}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} I_N(\theta)\, \mathcal{W}(e^{j(w-\theta)})\, d\theta$$

$$E\{S_{xx}(w)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} E\{I_N(\theta)\}\, W(e^{j(w-\theta)})\, d\theta$$