1. Zweierkomplement

a) Wie sieht die binäre Darstellung von –5 aus bei den Wortbreiten b = 4, b = 8, b = 16?

b) Berechnen Sie folgende Additionen im Format SINT(4). Geben Sie bei Überlauf auch die Ausgaben mit *wrap-around* und Sättigung an:
-5-1 -6-4 3-7 3+7-3 (in dieser Reihenfolge)

c) Geben Sie die binären Darstellungen von -0,75 in SFRAC(1.3) und SFRAC(3.5) an. Wie groß ist jeweils das LSB?

d) Eine Multiplikation liefert das Ergebnis 15/64 im Format SFRAC(1.6). Stellen Sie das Ergebnis gerundet im Format SFRAC(1.3) dar.

e) Berechnen Sie folgende Multiplikationen in SFRAC(1.3) durch binäre Verschiebung/Addition und Runden:
$\frac{5}{8}\cdot\frac{1}{8}$ $\left(-\frac{5}{8}\right)\cdot\frac{3}{8}$ $\left(-\frac{5}{8}\right)\cdot(-1)$

2. Quantisierte FIR-Filter

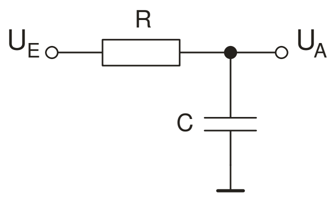
Ein FIR-Filter 255. Ordnung (256 Koeffizienten) benutzt für die Koeffizienten sowie das Eingangs- und Ausgangssignal das Format SFRAC(1.15).

a) Welches Format haben die Zwischenergebnisse der Multiplizierer-Ausgänge?

b) Bestimmen Sie den Effektivwert des Rundungsrauschens am Ausgang, wenn die Additionen mit der höheren Genauigkeit ausgeführt werden und erst am Ausgang gerundet wird.

c) Wie groß ist das Rundungsrauschen am Ausgang, wenn die Zwischenergebnisse vor der Addition gerundet werden?

3. Quantisierte IIR-Filter



Ein analoger RC-Tiefpass (R = 500 Ω, C = 1 µF) ist Vorbild für ein diskretes System. ($I_A = 0$)

a) Wie lautet die Fourier-Übertragungsfunktion $H(f)$?

b) Das diskrete System hat eine Abtastfrequenz von 9 kHz. Bestimmen Sie mit Hilfe der Bilinearen Transformation die Übertragungsfunktion $H(z)$ in der Form $H(z) = \frac{b_0+\frac{b_1}{z}}{1+\frac{a_1}{z}}$.

c) Das diskrete System wird mit der Transponierten Direktstruktur II realisiert. Das Format der Koeffizienten ist SFRAC(1.3). Wie verschieben sich der Pol und die Nullstelle dadurch? Vergleichen Sie mit Matlab die Verläufe von $H(p = j2\pi f)$ und $H\left(z = e^{j2\pi f\Delta t}\right)$ (ideal und quantisiert). Befehl: `freqz, freqs`

Ein IIR-Filter zweiter Ordnung hat die Übertragungsfunktion $H(z) = \frac{0{,}125 - \frac{0{,}177}{z} + \frac{0{,}125}{z^2}}{1 - \frac{1{,}5}{z} + \frac{0{,}6}{z^2}}$.

d) Modifizieren Sie die Transponierte Direktstruktur II so, dass das System mit Koeffizienten im Format SFRAC(1.3) darstellbar ist. Hinweis: 1,5 = 0,75 + 0,75

e) Wie verschieben sich die Pole und Nullstellen durch die Quantisierung?
Stellen Sie mit Matlab die Verläufe von H(z) des idealen und realen Systems dar.

Ein System ist durch die Differenzengleichung $y_k = x_k - 0{,}75 \cdot y_{k-1}$ gegeben.

f) Ist das System stabil? Wie sollte sich die Impulsantwort verhalten?

g) Berechnen Sie y_k im Format SFRAC(1.3) mit Runden für $x_k = \{1/8;\ 0;\ 0;\ \dots\}$.

Lösungen:

1.a) Die zu -5 gehörende UINT-Zahl ist $2^b - 5$:

Nur die letzten Bits sind charakteristisch. Bei höheren Auflösungen werden entsprechend viele „1“ vorangestellt.

b	-5 → UINT(b)	-5 binär
4	16-5 = 11	1011
8	256-5 = 251	1111 1011
16	2^{16}-5 = 65531	1111 1111 1111 1011

b)

-5	1	0	1	1	-6	1	0	1	0		3	0	0	1	1
-1	1	1	1	1	-4	1	1	0	0		-7	1	0	0	1
carry:	1	1	1		carry:						carry:		1	1	
-6	1	0	1	0	-10	0	1	1	0	= +6 (wrap)	-4	1	1	0	0
										= -8 (sat)					

wrap-around:					saturation:				
3	0	0	1	1					
+7	0	1	1	1					
carry:	1	1	1						
10	1	0	1	0	+7	0	1	1	1
-3	1	1	0	1	-3	1	1	0	1
7	0	1	1	1	4	0	1	0	0

c) Gebrochene Zahlen werden durch Ganzzahlen dargestellt: Bruch = Ganz / 2^n
SFRAC(1.3) hat 3 Nachkommastellen → Bruch = Ganz / 2^3 = Ganz / 8, das LSB beträgt 1/8
SFRAC(3.5) hat 5 Nachkommastellen → LSB = $1/2^5$ = 1/32

SFRAC(1.3): Ganzzahl = -0,75·8 = -6, binär SINT(4) 1010
SFRAC(3.5): Ganzzahl = -0,75·2^5 = -24, binär SINT(8) 11**10 10**00

Die charakteristischen Bits 1010 stehen in der Mitte, die Stellen davor werden mit dem Vorzeichen aufgefüllt (hier zweimal „1“), die Stellen danach mit „0“.

d) Beim Wechsel von SFRAC(1.6) auf SFRAC(1.3) wird die zugehörige Ganzzahl durch 2^3=8 dividiert (dadurch verschiebt sich der Radixpunkt um 3 Stellen), dann werden die Nachkommastellen abgeschnitten. Das Ergebnis ist immer die nächste negativere Ganzzahl.

Beim Runden wird vor dem Abschneiden der Nachkommastellen noch 0,5 addiert. In der Binärdarstellung entspricht das einer „1“ an der letzten abzuschneidenden Stelle.

15/64 in SFRAC(1.6) → Ganz = 15 → binär 7 Stellen:	000 1111
Runden:	+ 100
	001 0011
Abschneiden:	001 0

Die neue Ganzzahl ist 2 und bedeutet 2/8.

Kontrollrechnung: 15/64 = 0,234375 = 1,875/8 ≈ 2/8

e) Die Zwischenergebnisse liegen in SFRAC(1.6) vor. Binäre Darstellungen:
5/8 → 0000101, 1/8 →0000001, -5/8 → 1111011, 3/8 → 0000011, -1 = -8/8 → 1111000

	$\frac{5}{8}$	$\cdot\frac{1}{8}$
	000 0101	1
+	000 000	0
+	000 00	0
+	000 0	0
+	000	0
+	00	0
+	0	0 ↑
=	000 0101	
+	000 0100	
	000 1001	
	000 1	
≈	1/8	

	$\frac{-5}{8}$	$\cdot\frac{3}{8}$
	111 1011	1
+	111 011	1
+	0	0
+	0	0
+	0	0
+	0	0
+	0	0 ↑
=	111 0001	
+	000 0100	
	111 0101	
	111 0	
≈	-2/8	

	$\frac{-5}{8}$	$\cdot\frac{-8}{8}$
	000 0000	0
+	0	0
+	0	0
+	101 1	1
+	011	1
+	11	1
+	1	1 ↑
=	010 1000	
+	000 0100	
	010 1100	
	010 1	
=	+5/8	

2.a) SFRAC(1.30)

b) nur eine Quelle für Rundungsrauschen, dort ist $P_{round} = y_{noise}^2 = \frac{LSB^2}{12}$, $LSB = 2^{-15}$

$\rightarrow y_{noise} = \frac{LSB}{\sqrt{12}} = 8{,}81 \cdot 10^{-6}$; dadurch wird das unterste Bit im Mittel in 1 von 12 Fällen falsch (ein Signal, das einen Takt lang gleich LSB und dann 11mal 0 ist produziert den gleichen Effektivwert). Wenn das Eingangssignal vom AD-Wandler stammt, kommt nochmal so viel Rauschleistung dazu (unabhängige Signale/Fehler/Rauschen: Leistungen addieren sich, nicht Effektivwerte). Rundungsrauschen ist gleichverteilt, der Maximalfehler beträgt LSB/2.

c) alle Multiplizierer-Ausgänge produzieren Rundungsrauschen, das addiert wird.

$\rightarrow P_{noise} = 256 \cdot P_{round} \quad \rightarrow \quad y_{noise,gesamt} = \sqrt{256} \cdot y_{noise} = 2^4 \cdot y_{noise} = 1{,}41 \cdot 10^{-4}$

wegen 2^4 sind <u>im Mittel</u> die untersten 4 Bit verrauscht, bei b = 16 bleiben nur 12 gültige Bit.
Die Summe von 256 gleichverteilten Rundungsfehlern ist schon fast normalverteilt; tatsächlich kann ein Maximalfehler von ±128 LSB (unterste 7 Bit) auftreten, ist aber sehr unwahrscheinlich.

3.a) vgl. Übung 3: $H(f) = \frac{1/j\omega C}{R+1/j\omega C} = \frac{1}{1+j2\pi f \cdot RC}$

b) ersetze $f = \frac{2 \cdot f_a}{j2\pi} \cdot \frac{z-1}{z+1} \quad \rightarrow \quad H(z) = \frac{1}{1+j2\pi \cdot \frac{2f_a}{j2\pi} \cdot \frac{z-1}{z+1} \cdot RC}$; $\quad 2f_a \cdot RC = 9 \rightarrow$

$$H(z) = \frac{1}{1+9 \cdot \frac{z-1}{z+1}} = \frac{z+1}{z+1+9 \cdot (z-1)} = \frac{z+1}{10z-8} \cdot \frac{0{,}1/z}{0{,}1/z} = \frac{0{,}1+0{,}1z^{-1}}{1-0{,}8z^{-1}}$$

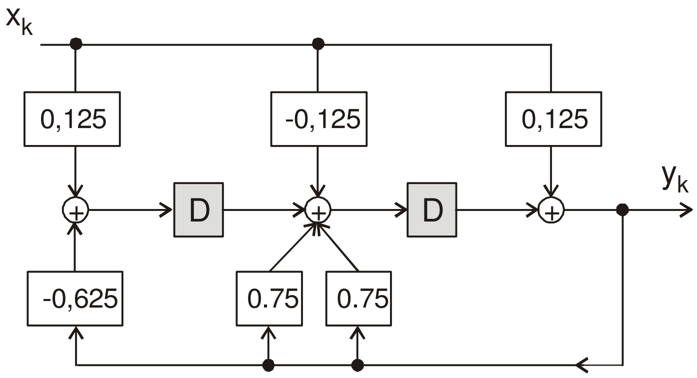
c) ideal: Nullstelle bei x_0 = -1, Polstelle bei z_∞ = 0,8

quantisiert: $H_q(z) = \frac{0{,}125+0{,}125z^{-1}}{1-0{,}75z^{-1}}$ → Nullstelle bei x_0 = -1, Polstelle bei z_∞ = 0,75

Obwohl alle Koeffizienten verändert wurden, verschiebt sich hier nur die Polstelle!

Matlab: siehe Skript DSP5_3.m

d) es können nur Koeffizienten im Bereich [-1; +1[verwendet werden. Am nächsten liegen:
-0,177 → -0,125; 0,6 → 0,625



e) ideal: Matlab oder quadratische Lösungsformel: $z_0 = 0{,}71 \pm j0{,}71$;
Pole (s. Skript) $|z_\infty|^2 = a_2 = 0{,}6$; $Re\{z_\infty\} = -\frac{a_1}{2} = +0{,}75 \rightarrow\ Im\{z_\infty\} = \pm 0{,}194$

real: $H_q(z) = \frac{0{,}125 - \frac{0{,}125}{z} + \frac{0{,}125}{z^2}}{1\ -\ 2\cdot\frac{0{,}75}{z}\ +\ \frac{0{,}625}{z^2}}$ → Nullstellen $z_0 = 0{,}5 \pm j0{,}87$;
Pole $|z_\infty|^2 = 0{,}625$; $Re\{z_\infty\} = -\frac{a_1}{2} = +0{,}75 \rightarrow\ Im\{z_\infty\} = \pm 0{,}25$

f) z-Trafo: $Y = X - 0{,}75 \cdot Y \cdot \frac{1}{z}\ \rightarrow\ H(z) = \frac{1}{1+\frac{0{,}75}{z}}$; Polstelle bei $z_\infty = -0{,}75\ \rightarrow$ stabil
Der Betrag der Impulsantwort sollte deshalb mit der Zeit immer kleiner werden.

g)

k	x_k	y_{k-1}	y_k
0	$\frac{1}{8}$	0	$\frac{1}{8}$
1	0	$\frac{1}{8}$	$\frac{-0{,}75}{8} \approx -\frac{1}{8}$
2	0	$\frac{-1}{8}$	$\frac{+0{,}75}{8} \approx +\frac{1}{8}$
3…∞	0	periodisch!	

Da hier immer zu größeren Beträgen gerundet wird, klingt y_k nicht ab sondern schwingt.