DSP3 – Practice Homework
Solutions

Exercise 1. Interpolation.
Consider a finite-energy discrete-time sequence $x[n]$ with DTFT $X(e^{j\omega})$ and the continuous-time interpolated signal

$$x_0(t) = \sum_{n=-\infty}^{\infty} x[n]\text{rect}(t-n)$$

i.e. a signal obtained from the discrete-time sequence using a zero-centered zero-order hold with interpolation period $T_s = 1$s. Let $X_0(f)$ be the Fourier transform of $x_0(t)$.

(a) Express $X_0(f)$ in terms of $X(e^{j\omega})$.

(b) Compare $X_0(f)$ to $X(f)$, where $X(f)$ is the spectrum of the continuous-time signal obtained using an ideal sinc interpolator with $T_s = 1$:

$$x(t) = \sum_{n=-\infty}^{\infty} x[n]\text{sinc}(t-n)$$

Comment on the result: you should point out two major problems.

(c) The signal $x(t)$ can be obtained back from the zero-order hold interpolation via a continuous-time filtering operation:

$$x(t) = x_0(t) * g(t).$$

Sketch the frequency response of the filter $g(t)$.

(d) Propose two solutions (one in the continuous-time domain, and another in the discrete-time domain) to eliminate or attenuate the distortion due to the zero-order hold. Discuss the advantages and disadvantages of each.

Solution 1.

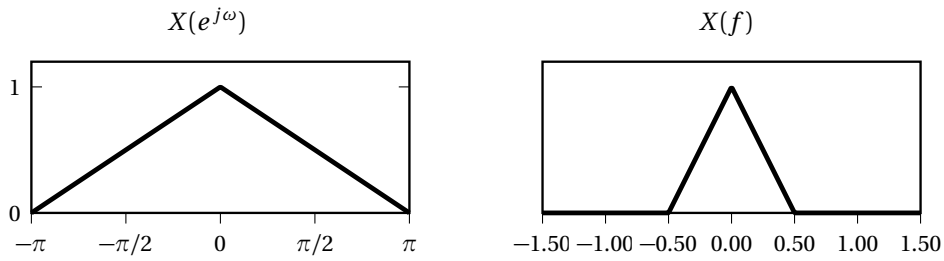
(a)

$$\begin{aligned}
X_0(f) &= \int_{-\infty}^{\infty} x_0(t) e^{-j2\pi f t}\, dt \\
&= \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[n]\text{rect}(t-n) e^{-j2\pi f t}\, dt \\
&= \sum_{n=-\infty}^{\infty} x[n] \int_{-\infty}^{\infty} \text{rect}(t-n) e^{-j2\pi f t}\, dt \\
&= \sum_{n=-\infty}^{\infty} x[n] e^{-j2\pi f n} \int_{-1/2}^{1/2} e^{-j2\pi f \tau}\, d\tau \\
&= X(e^{j2\pi f}) \frac{\sin(\pi f)}{\pi f} \\
&= \text{sinc}(f) X(e^{j2\pi f}).
\end{aligned}$$

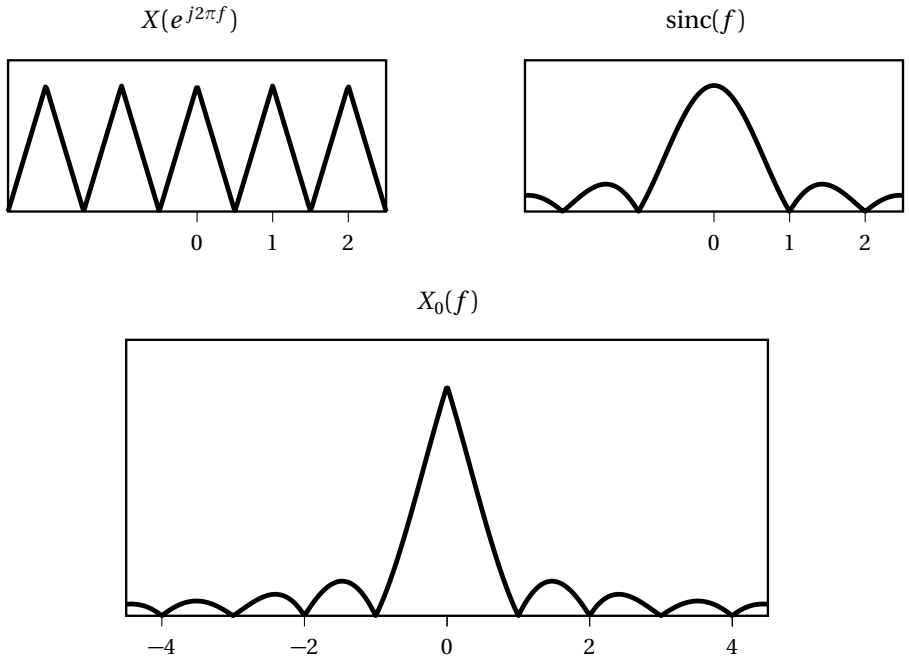
(b) To understand the effects of the zero-order hold consider for instance a discrete-time signal with a triangular spectrum, as shown in the left panel below. We know that sinc interpolation will exactly preserve the shape of the spectrum and return a continuous-time signal that is strictly bandlimited to the $[-F_s/2, F_s/2]$ interval (with $F_s = 1/T_s = 1$), that is:

$$X(f) = \begin{cases} X(e^{j2\pi f}) & |f| < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

as shown in the right panel below.

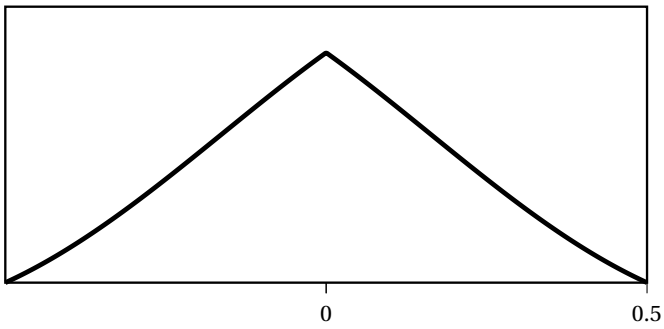


Conversely, the spectrum of the continuous-time signal interpolated by the zero-order hold is the product of $X(e^{j2\pi f})$, which is periodic with period $F_s = 1$ Hz, and the term sinc(f), whose first spectral null is for $f = 1$ Hz. Here are the two terms, and their product, in magnitude:



There are two main problems in the zero-order hold interpolation as compared to the sinc interpolation:

- The zero-order hold interpolation is NOT bandlimited: the 2π-periodic replicas of the digital spectrum leak through in the continuous-time signal as high frequency components. This is due to the sidelobes of the interpolation function in the frequency domain (rect in time $\longleftrightarrow$ sinc in frequency) and it represents an undesirable high-frequency content which is typical of all local interpolation schemes.
- There is a distortion in the main portion of the spectrum (that between $-F_s/2$ and $F_s/2 = 0.5$ Hz) due to the non-flat frequency response of the interpolation function. It can be seen if we zoom in the baseband portion:



(c) As we have seen, $X(f) = X(e^{j2\pi f})\text{rect}(f)$ while $X_0(f) = \text{sinc}(f)\,X(e^{j2\pi f})$. Therefore

$$G(f) = \begin{cases} \frac{1}{\text{sinc}(f)} & |f| < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

(d) A first solution is to compensate for the distortion introduced by $G(f)$ in the discrete-time domain. This is equivalent to pre-filtering $x[n]$ with a discrete-time filter of magnitude $1/G(e^{j2\pi f})$. The advantages of this method is that digital filters such as this one are relatively easy to design and that the filtering can be done in the discrete-time domain. The disadvantage is that this approach does not eliminate or attenuate the high frequency leakage outside the baseband.

In continuous time, one could cascade the interpolator with an analog lowpass filter to eliminate the leakage. The disadvantage is that it is hard to implement an analog lowpass which can also compensate for the in-band distortion introduced by $G(f)$; such a filter will also introduce unavoidable phase distortion (no analog filter has linear phase).

Exercise 2. A bizarre interpolator

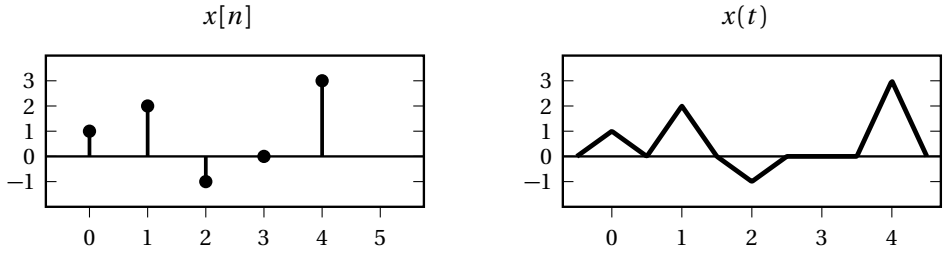
Consider a local interpolation scheme as in the previous exercise but now the characteristic of the interpolator is:

$$i(t) = \begin{cases} 1-2|t| & |t| \le 1/2 \\ 0 & \text{otherwise} \end{cases}$$

This is a triangular characteristic but the same unit support as the zero-order hold. If we pick an interpolation interval T_s and interpolate a given discrete-time signal $x[n]$ with $I(t)$, we obtain the continuous-time signal

$$x(t) = \sum_n x[n] i\left(\frac{t - nT_s}{T_s}\right)$$

an example of which is shown here



Assume that the spectrum of $x[n]$ between $-\pi$ and π is

$$X(e^{j\omega}) = \begin{cases} 1 & |\omega| \le 2\pi/3 \\ 0 & \text{otherwise} \end{cases}$$

(with the obvious 2π-periodicity over the entire frequency axis).

(a) Compute and sketch the Fourier transform $I(f)$ of the interpolating function $i(t)$. Recall that the triangular function can be expresses as the convolution of a suitably scaled rect with itself.

(b) Sketch the Fourier transform $X(f)$ of the interpolated signal $x(t)$; in particular, clearly mark the Nyquist frequency $F_s/2$

(c) The use of $i(t)$ instead of a sinc interpolator introduces two types of errors: briefly describe them.

(d) To eliminate the error in the baseband $[-F_s/2, F_s/2]$ we can pre-filter the signal $x[n]$ before interpolating with $i(t)$. Write the frequency response of the required discrete-time pre-filter $H(e^{j\omega})$.

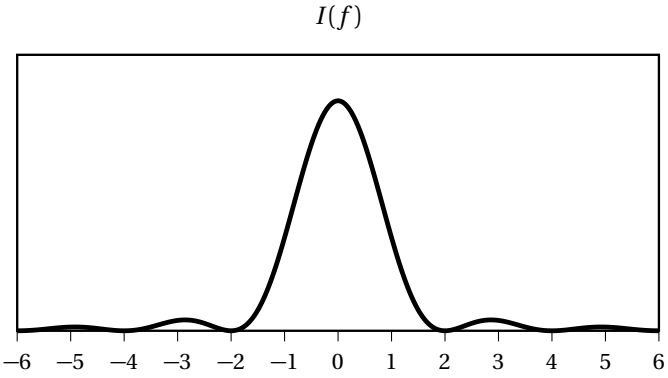
Solution 2.

(a) We know that the convolution of rect(t) with itself produces a rectangular function between -1 and 1 and with height 1. To obtain a rectangular function of support 1 and height 1, we need to shrink the rects by a factor of two and compensate for the value in $t = 0$ so that

$$i(t) = 2\,\text{rect}(2t) * \text{rect}(2t).$$

From this

$$I(f) = 2\left[\frac{1}{2}\text{sinc}\left(\frac{f}{2}\right)\right]^2 = \frac{1}{2}\text{sinc}^2\left(\frac{f}{2}\right).$$



(b) We know that the spectrum of an interpolated signal will be

$$X(f) = X(e^{j2\pi f T_s})\, T_s I\left(f T_s\right)$$

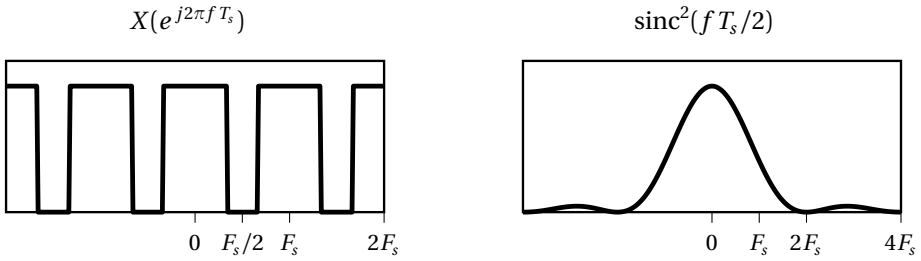
In our case

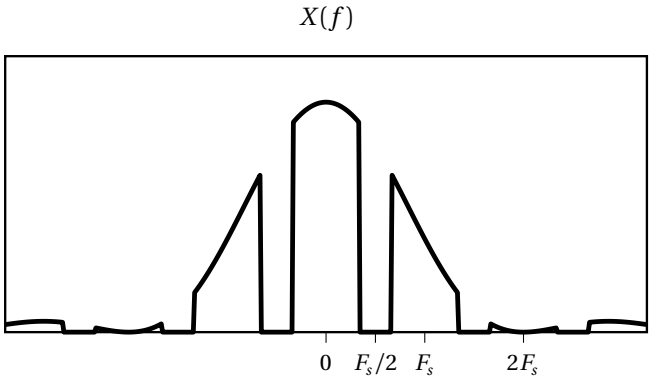
$$X(f) = X(e^{j2\pi f T_s})\, T_s I\left(f T_s\right) = \frac{T_s}{2} X(e^{j2\pi f T_s})\text{sinc}^2\left(\frac{f T_s}{2}\right)$$

So the Fourier transform of the interpolated signal is the products of two parts:

- the F_s-periodic spectrum $X(e^{j2\pi f T_s})$, with $F_s = 1/T_s$;
- the Fourier transform of the interpolating function.

The two components and their products are shown here:





(c) like before, here are two types of error, in-band and out-of-band:

- **In−band**: the spectrum between $[-F_s/2, F_s/2]$ (the baseband) is distorted by the non-flat response of the interpolating function over the baseband.
- **Out−of−band**: the periodic copies of $X(e^{j2\pi f T_s})$ outside of $[-F_s/2, F_s/2]$ are not eliminated by the interpolation filter, since it is not an ideal lowpass.

(d) As before, e need to undo the linear distortion introduced by the nonflat response of the interpolation filter in the baseband. The idea is to have a modified spectrum $H(e^{j\omega})X(e^{j\omega})$ so that, over $[-F_s/2, F_s/2]$, it is

$$X(f) = X(e^{j2\pi f T_s}).$$

If we use $H(e^{j\omega})X(e^{j\omega})$ in the interpolation formula, we have

$$X(f) = \frac{T_s}{2} H(e^{j2\pi f T_s}) X(e^{j2\pi f T_s}) \mathrm{sinc}^2\left(\frac{f T_s}{2}\right)$$

so that

$$H(e^{j2\pi f T_s}) = [\frac{T_s}{2}\mathrm{sinc}^2\left(\frac{f T_s}{2}\right)]^{-1}.$$

Therefore, the frequency response of the digital filter will be

$$H(e^{j\omega}) = \frac{2}{T_s}\mathrm{sinc}^{-2}\left(\frac{\omega}{4\pi}\right), \qquad -\pi \le \omega \le \pi$$

prolonged by 2π-periodicity over the entire frequency axis.

Exercise 3. Another view of Sampling

An alternative way of describing the sampling operation relies on the concept of *modulation by a pulse train*. Given a sampling interval T_s, a continuous-time pulse train $p(t)$ is an infinite collection of equally spaced Dirac deltas:

$$p(t) = \sum_{k=-\infty}^{\infty} \delta(t - k T_s).$$

The pulse train is the used to modulate a continuous-time signal:

$$x_s(t) = p(t)\,x(t).$$

Intuitively, $x_s(t)$ represents a "hybrid" signal where the nonzero values are only those of the discrete time samples that would be obtained by raw-sampling $x(t)$ with period T_s; however, instead of representing the samples a countable sequence (i.e. with a different mathematical object) we are still using a continuous-time signal that is nonzero only over infinitesimally short instants centered on the sampling times. Using Dirac deltas allows us to embed the instantaneous sampling values in the signal.
Note that the Fourier Transform of the pulse train is

$$P(f) = F_s \sum_{k=-\infty}^{\infty} \delta\big(f - kF_s\big)$$

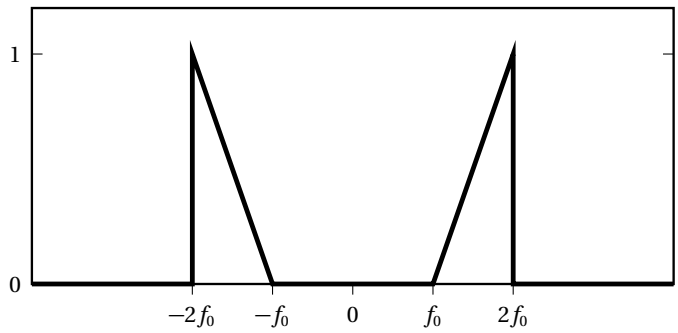
(where, as per usual, $F_s = 1/T_s$). This result is a bit tricky to show, but the intuition is that a periodic set of pulses in time produces a periodic set of pulses in frequency and that the spacing between pulses frequency is inversely proportional to the spacing between pulses in time.
Derive the Fourier transform of $x_s(t)$ and show that if $x(t)$ is bandlimited to $F_s/2$, where $F_s = 1/T_s$, then we can reconstruct $x(t)$ from $x_s(t)$.

Solution 3.
By using the modulation theorem, the product in time becomes a convolution in frequency:

$$\begin{aligned} X_s(f) &= X(f) * P(f) \\ &= \int_{\mathbb{R}} X(\varphi) P(f-\varphi) d\varphi \\ &= F_s \int_{\mathbb{R}} X(\varphi) \sum_{k\in\mathbb{Z}} \delta\big(f-\varphi-kF_s\big)\, d\varphi \\ &= F_s \sum_{k\in\mathbb{Z}} X\big(f-kF_s\big). \end{aligned}$$

In other words, the spectrum of the delta-modulated signal is the periodization (with period $F_s = 1/T_s$) of the original spectrum. If the latter is bandlimited to $F_s/2$ there will be no overlap between copies in the periodization and therefore $x(t)$ can be obtained simply by lowpass filtering $x_s(t)$ in the continuous-time domain.

Exercise 4. Bandpass sampling
Consider a real, continuous-time signal $x_c(t)$ with the following spectrum $X_c(f)$:



(a) What is the bandwidth of the signal? What is the minimum sampling frequency that satisfies the sampling theorem?

(b) If we sample the signal with a sampling frequency $F_a = 2f_0$, clearly there will be aliasing. Plot the DTFT of the resulting discrete-time signal $x_a[n] = x_c(n/F_a)$.

(c) Suggest a way to perfectly reconstruct $x_c(t)$ from $x_a[n]$.

(d) From the previous example it would appear that we can exploit "gaps" in the original spectrum to reduce the sampling frequency necessary to losslessly sample a bandpass signal. In general, what is the minimum sampling frequency that we can use to sample with no loss a real-valued signal whose frequency support on the positive axis is $[f_0, f_1]$ (with the usual symmetry around zero, of course)?

Solution 4.

(a) The highest nonzero frequency is $2f_0$ and therefore $x_c(t)$ is $4f_0$-bandlimited. The minimum sampling frequency that satisfies the sampling theorem is $F_s = 4f_0$. Note however that the support over which the (positive) spectrum is nonzero is the interval $[f_0, 2f_0]$ so that one could say that the total *effective* bandwidth of the signal is only $2f_0$.

(b) The digital spectrum will be the periodized continuous-time spectrum, rescaled to $[-\pi, \pi]$; the periodization after sampling at a frequency $F_a = 2f_0$, yields

$$\tilde{X}_c(f) = \sum_{k=-\infty}^{\infty} X_c(f - 2kf_0).$$

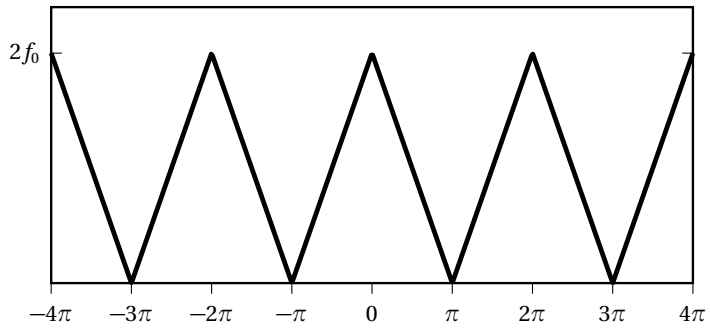
The general term $X_c(f - 2kf_0))$ is nonzero for $f_0 \le |f - 2kf_0| \le 2f_0$ for $k \in \mathbb{Z}$ or, equivalently,

$$(2k+1)f_0 \le f \le (2k+2)f_0$$
$$(2k-2)f_0 \le f \le (2k-1)f_0.$$

These are non-overlapping intervals and, therefore, no disruptive superposition will occur. The DTFT of the samples signal is

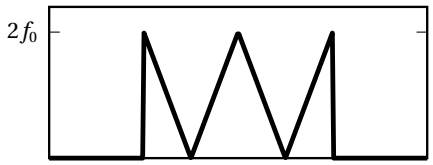
$$X_a(e^{j\omega}) = 2f_0 \sum_{k=-\infty}^{\infty} X_c(\frac{\omega}{\pi} f_0 - 2kf_0);$$

for instance, as ω goes from zero to π, the nonzero contribution to the DTFT will be the term $X_c(\frac{\omega}{\pi} f_0 - 2f_0)$ where the argument goes from $-2f_0$ to $-f_0$. The spectrum is represented here with periodicity explicit:



(c) Here's a possible scheme:

- Sinc-interpolate $x_a[n]$ with period $T_a = 1/F_a$ to obtain $x_b(t)$; the spectrum will be like so:



- filter in continuous time $x_p(t)$ with an ideal bandpass filter with (positive) passband equal to $[f_0, 2f_0]$ to obtain $x_c(t)$.

(d) The effective *positive* bandwidth of a signal whose spectrum is nonzero only over $[-f_1,-f_0]\cup[f_0,f_1]$ is $W = f_1 - f_0$. Obviously the sampling frequency must be at least equal to the *total* effective bandwidth, so a first condition on the sampling frequency is $F_s \geq 2W$. We can now distinguish two cases.

1) assume f_1 is a multiple of the positive bandwidth, i.e. $f_1 = MW$ for some integer $M > 1$ (for $x[n]$ above, it was $M = 2$). Then the argument we made before can be easily generalized: if we pick $F_s = 2W$ and sample we have that

$$\tilde{X}_c(f) = \sum_{k=-\infty}^{\infty} X_c(f - 2kW).$$

The general term $X_c(f-2kW)$ is nonzero only for

$$f_0 \leq |f - 2kW| \leq f_1 \quad \text{for } k \in \mathbb{Z}.$$

Since $f_0 = f_1 - W = (M-1)W$, this translates to

$$\begin{aligned}(2k+M-1)W \leq f \leq (2k+M)W \\ (2k-M)W \leq f \leq (2k-M+1)W\end{aligned}$$

which, once again, are non-overlapping intervals.

2) if f_1 is *not* a multiple of W the easiest thing to do is to decrease the lower frequency f_0 to a new *smaller* frequency f_0' so that the new positive bandwidth $W' = f_1 - f_0'$ divides f_1 exactly. In other words we set a new lower frequency f_0' so that it will be $f_1 = M(f_1 - f_0')$ for some integer $M > 1$; it is easy to see that

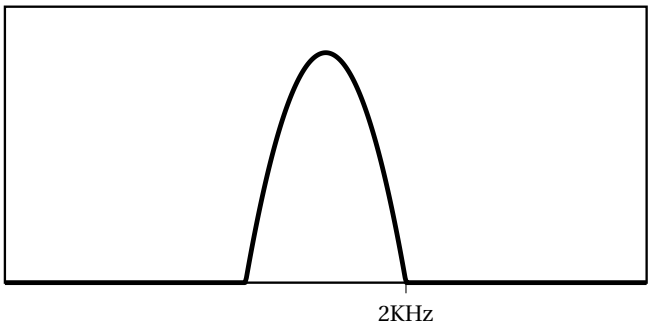
$$M = \left\lfloor \frac{f_1}{f_1 - f_0} \right\rfloor.$$

since this is the maximum number of copies of the W-wide spectrum which fit *with no overlap* in the $[0, f_0]$ interval. If $W > f_0$ obviously we cannot hope to reduce the sampling frequency and we have to use normal sampling. This artificial change of frequency will leave a small empty "gap" in the new bandwidth $[f_0', f_1]$, but that's no problem. Now we use the previous result and sample at $F_s = 2(f_1 - f_0')$ with no overlaps. Since $f_1 - f_0' = f_1/M$, we have that, in conclusion, the minimum sampling frequency is

$$F_s = 2f_1 / \left\lfloor \frac{f_1}{f_1 - f_0} \right\rfloor$$

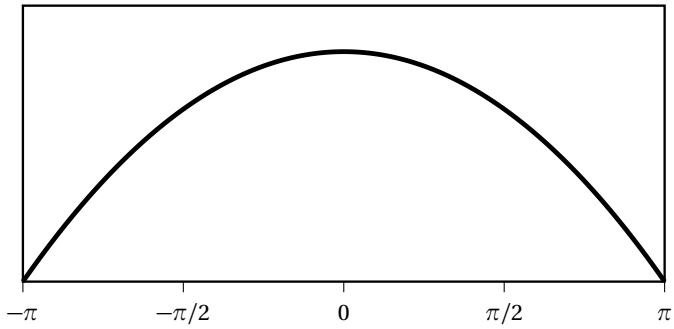
i.e. we obtain a sampling frequency reduction factor of $\lfloor f_1/(f_1 - f_0) \rfloor$.

Exercise 5. Aliasing or not.
Consider a bandlimited continuous-time signal $x(t)$ with the following spectrum $X(f)$:

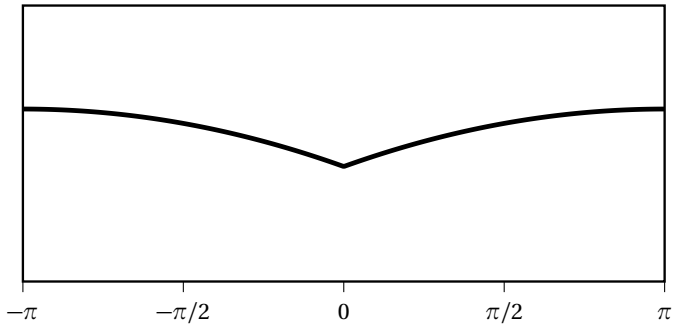


Sketch the DTFT of the discrete-time signal $x[n] = x(n/F_s)$ for the cases $F_s = 4$KHz and $F_s = 2$KHz.

Solution 5.
For $F_s = 4$KHz there is no aliasing and the spectrum is like so:



For $F_s = 2$KHz there is aliasing and we have



Exercise 6. Multirate identities
Prove the following two identities:

(a) Downsampling by 2 followed by filtering by $H(z)$ is equivalent to filtering by $H(z^2)$ followed by downsampling by 2.

(b) Filtering by $H(z)$ followed by upsampling by 2 is equivalent to upsampling by 2 followed by filtering by $H(z^2)$.

Solution 6.
We will operate in the z-domain; recall that, given the z-transform of the original sequence $X(z)$, the z-transform of a 2-upsampled sequence is $X(z^2)$ whereas the z-transform of a 2-downsampled sequence is $[X(z^{1/2})+X(-z^{1/2})]/2$. With this:

(a) Downsampling by 2 followed by filtering by $H(z)$ yields

$$Y_1(z)=\frac{1}{2}H(z)\big(X(z^{1/2})+X(-z^{1/2})\big).$$

Filtering by $H(z^2)$ followed by downsampling by 2 can be written as

$$\begin{aligned}Y_2(z)&=\frac{1}{2}\big(H((z^{1/2})^2)X(z^{1/2})+H((-z^{1/2})^2)X(-z^{1/2})\big)\\&=\frac{1}{2}H(z)\big(X(z^{1/2})+X(-z^{1/2}))\big).\end{aligned}$$

The two operations are thus equivalent. (Please note in $H(z^2)$ how it's the *argument* z of the z-transform that is replaced by $\pm z^{1/2}$, so that the minus sign goes away. This makes sense: if a sequence has been upsampled, downsampling simply returns it to its original form.)

(b) Filtering by $H(z)$ followed by upsampling by 2 can be written as

$$Y_1(z)=H(z^2)X(z^2).$$

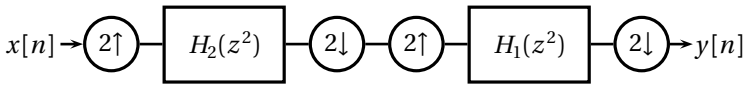
Upsampling by 2 followed by filtering by $H(z^2)$ can be written as

$$Y_2(z)=H(z^2)X(z^2).$$

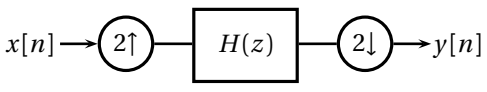
The two operations are thus equivalent.

Exercise 7. Multirate systems.
Consider the input-output characteristic of the following multirate systems. Remember that, technically, one cannot talk of transfer functions in the case of multirate systems since sampling rate changes are not time-invariant. It may happen, though, that by carefully designing the processing chain, the input-output characteristic does indeed implement a time-invariant transfer function.

(a) Find the overall transformation operated by the following system:



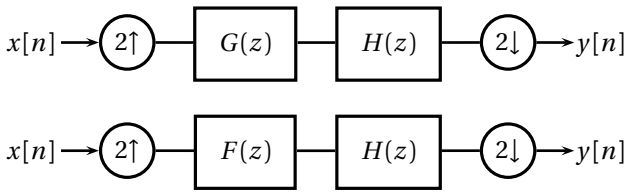
(b) Assume $H(z)=A(z^2)+z^{-1}B(z^2)$ for arbitrary $A(z)$ and $B(z)$. Show that the transfer function of the following system is equal to $A(z)$.



(c) Let $H(z)$, $F(z)$ and $G(z)$ be filters satisfying

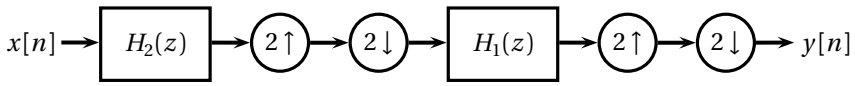
$$\begin{aligned}H(z)G(z)+H(-z)G(-z)&=2\\H(z)F(z)+H(-z)F(-z)&=0\end{aligned}$$

Prove that one of the following systems is unity and the other zero:



Solution 7.
In this solution we will be using the identities proven in the previous exercise.

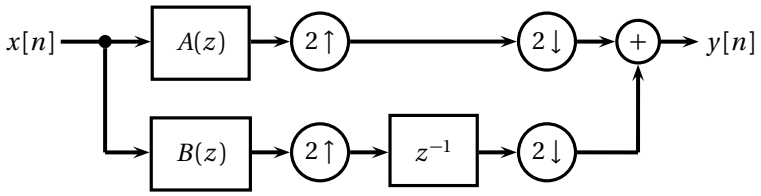
(a) The system can be redrawn as



Upsampling by N immediately followed by downsampling by N leaves the signal unchanged so the transfer function of this system is simply

$$H(z) = H_1(z)H_2(z).$$

(b) The system is equivalent to



The lower branch contains an upsampler by two followed by a delay and a downsampler by two. The output of such a processing chain is easily seen to be equal to 0 for all n. Thus only the upper branch remains and the final transfer function of the system is given by

$$H(z) = A(z).$$

(c) In both systems, call $w[n]$ the signal just before the final downsampler by two; the z-transform of the output will therefore be

$$Y(z) = [W(z^{1/2}) + W(-z^{1/2})]/2.$$

For the first system we have that $W(z) = H(z)G(z)X(z^2)$ so

$$\begin{aligned} Y(z) &= \frac{1}{2}\left(H(z^{1/2})G(z^{1/2})X(z) + H(-z^{1/2})G(-z^{1/2})X(z)\right) \\ &= \frac{1}{2}\left(H(z^{1/2})G(z^{1/2}) + H(-z^{1/2})G(-z^{1/2})\right)X(z) \\ &= X(z). \end{aligned}$$

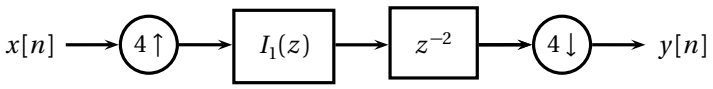
The transfer function is thus unity.

In the second system we have that $W(z) = H(z)F(z)X(z^2)$ so

$$\begin{aligned} Y(z) &= \frac{1}{2}\left(H(z^{1/2})F(z^{1/2})X(z) + H(-z^{1/2})F(-z^{1/2})X(z)\right) \\ &= \frac{1}{2}\left(H(z^{1/2})F(z^{1/2}) + H(-z^{1/2})F(-z^{1/2})\right)X(z) \\ &= 0. \end{aligned}$$

Its transfer function is thus zero.

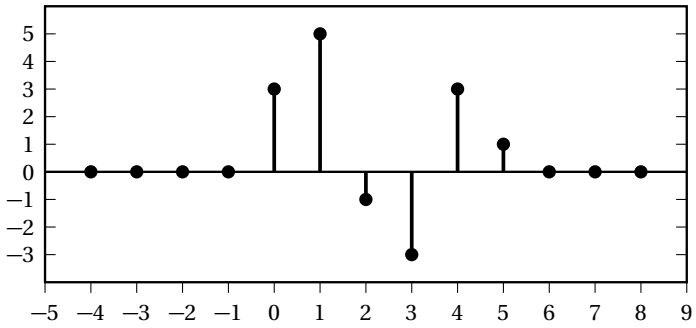
Exercise 8. Fractional resampling with multirate.
Consider the following multirate processing system:



where $I_1(z)$ is the first-order discrete-time interpolator with impulse response

$$i_1[n] = \begin{cases} 1-|n|/4 & \text{for } |n|<4 \\ 0 & \text{otherwise.} \end{cases}$$

Assume $x[n]$ is the finite-support signal shown here:



Compute the values of $y[n]$ for $0 \le n \le 6$, showing your calculation method.

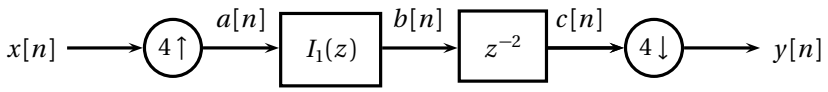
Solution 8.
Intuitively, the upsampling by four followed by $I_1(z)$ creates a linear interpolation over three "extra" samples between the original values the ("connect-the-dots" strategy). The delay by two followed by the downsampler selects the midpoint of each interpolation interval. As a whole, the chain implements a fractional delay of half a sample using a linear interpolator so that

$$y[n] = \frac{x[n] + x[n-1]}{2}$$

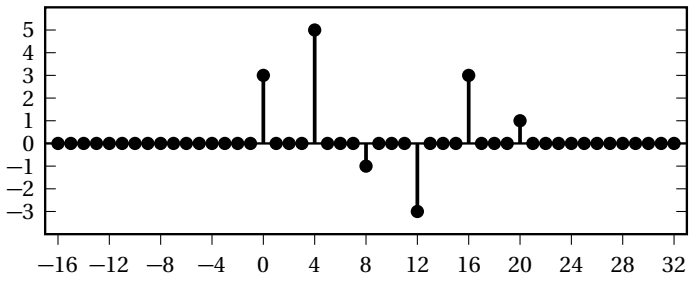
The required values are therefore:

$$y[0] = 1.5, \quad y[1] = 4, \quad y[2] = 2, \quad y[3] = -2, \quad y[4] = 0, \quad y[5] = 2, \quad y[6] = 0.5$$

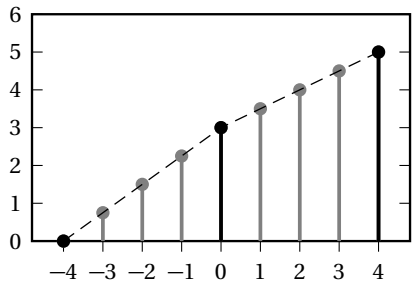
For a proof, label the intermediate signals in the processing chain like so:



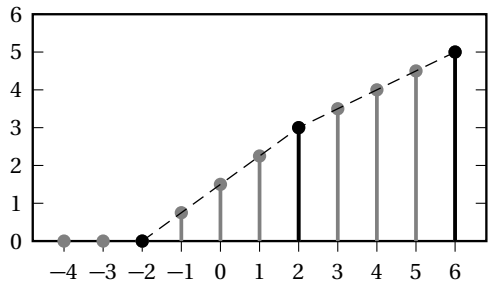
We can either proceed graphically or analytically. Graphically, which is the easiest way, we can start by plotting $a[n]$:



Since the interpolator $I(z)$ has finite support of length 7, we can concentrate on the interval $[-4,4]$ and extend the result to the other points. Linear interpolation fills in the gaps while the delay shifts the interpolated signal by two towards the right:



$b[n]$



$c[n]$

The downsampler selects the points in $c[n]$ where n is a multiple of four, which are the midpoints between original data values:

$$y[0] = c[0] = b[-2] = (x[0] + x[-1])/2 = 1.5$$
$$y[1] = c[4] = b[2] = (x[1] + x[0])/2 = 4$$
$$\cdots$$

Alternatively, we can proceed analytically as follows. The z-transform of $c[n]$ is

$$C(z) = X(z^4)z^{-2}I(z)$$

and, after the downsampler, we have

$$\begin{aligned} Y(z) &= \frac{1}{4}\sum_{m=0}^{3} C\left(e^{-j\frac{2\pi}{4}m} z^{\frac{1}{4}}\right) \\ &= \frac{1}{4}\sum_{m=0}^{3} X(z)\, e^{-j\frac{2\pi}{4}2m} z^{-\frac{1}{2}}\, I\left(e^{-j\frac{\pi}{2}m} z^{\frac{1}{4}}\right) \\ &= X(z)\frac{1}{4} z^{-\frac{1}{2}}\left[I\left(z^{\frac{1}{4}}\right) - I\left(-jz^{\frac{1}{4}}\right) + I\left(-z^{\frac{1}{4}}\right) - I\left(jz^{\frac{1}{4}}\right)\right] \end{aligned}$$

The transfer function of the interpolator is

$$I(z) = 1 + (1/4)(z + z^{-1}) + (1/2)(z^2 + z^{-2}) + (3/4)(z^3 + z^{-3})$$

and therefore

$$I\left(z^{\frac{1}{4}}\right) = 1+(1/4)(z^{1/4}+z^{-1/4})+(1/2)(z^{1/2}+z^{-1/2})+(3/4)(z^{3/4}+z^{-3/4})$$
$$I\left(-z^{\frac{1}{4}}\right) = 1-(1/4)(z^{1/4}+z^{-1/4})+(1/2)(z^{1/2}+z^{-1/2})-(3/4)(z^{3/4}+z^{-3/4})$$
$$I\left(-jz^{\frac{1}{4}}\right) = 1+(j/4)(z^{1/4}+z^{-1/4})-(1/2)(z^{1/2}+z^{-1/2})-(3j/4)(z^{3/4}+z^{-3/4})$$
$$I\left(jz^{\frac{1}{4}}\right) = 1-(j/4)(z^{1/4}+z^{-1/4})-(1/2)(z^{1/2}+z^{-1/2})+(3j/4)(z^{3/4}+z^{-3/4})$$

Finally,

$$I\left(z^{\frac{1}{4}}\right)+I\left(-z^{\frac{1}{4}}\right)-I\left(-jz^{\frac{1}{4}}\right)-I\left(jz^{\frac{1}{4}}\right) = 2(z^{1/2}+z^{-1/2})$$

so that

$$Y(z) = X(z)\frac{1}{4}z^{-\frac{1}{2}}\left[2(z^{1/2}+z^{-1/2})\right] = \frac{1+z^{-1}}{2}X(z).$$

Exercise 9. Quantization.
Consider a stationary i.i.d. random process $x[n]$ whose samples are uniformly distributed over the $[-1,1]$ interval. Consider a quantizer $\mathcal{Q}\{\cdot\}$ with the following characteristic:

$$\mathcal{Q}\{x\} = \begin{cases} -1 & \text{if } -1 \le x < -0.5 \\ 0 & \text{if } -0.5 \le x \le 0.5 \\ 1 & \text{if } 0.5 < x \le 1 \end{cases}$$

Compute the power of the quantization error.

Solution 9.
The power of the quantization error is given by

$$\begin{aligned}\sigma_e^2 &= \mathrm{E}\left[(x[n]-\mathcal{Q}\{x[n]\})^2\right] \\ &= \int_A^B f_x(\tau)(\tau-\mathcal{Q}\{\tau\})^2 d\tau\end{aligned}$$

where the integral is computed over the range of the input and $f_x(x)$ is the probability density function of the samples. We can split the integral over the non-overlapping quantization intervals i_k as

$$\sigma_e^2 = \sum_k \int_{i_k} f_x(\tau)(\tau-\mathcal{Q}\{\tau\})^2 d\tau.$$

In this quantizer there are three quantization intervals and the samples are uniformly distributed so

$$\begin{aligned}\sigma_e^2 &= \frac{1}{B-A}\left[\int_{-1}^{-0.5}(\tau+1)^2 d\tau + \int_{-0.5}^{0.5}\tau^2 d\tau + \int_{0.5}^{1}(\tau-1)^2 d\tau\right] \\ &= \dots \\ &= \frac{1}{12}\end{aligned}$$

Exercise 10. Quantization.
Consider a stationary i.i.d. random process $x[n]$ whose samples are uniformly distributed over the $[-1,2]$ interval. The process is uniformly quantized with a 1-bit quantizer with the following characteristic:

$$\mathcal{Q}\{x\} = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \ge 0 \end{cases}$$

Compute the signal to noise ratio at the output of the quantizer

---

Solution 10.
The input is uniformly distributed so its power will be

$$\sigma_x^2 = \frac{(B-A)^2}{12} = \frac{3}{4}$$

The power of the error can be computed as before:

$$\begin{aligned}\sigma_e^2 &= \frac{1}{3}\left[\int_{-1}^{0}(\tau+1)^2 d\tau + \int_{0}^{2}(\tau-1)^2 d\tau\right] \\ &= \dots \\ &= \frac{1}{3}\end{aligned}$$

so that the SNR is

$$\frac{\sigma_x^2}{\sigma_e^2} = \frac{9}{4}$$

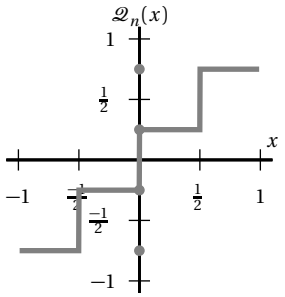
Exercise 11. Deadzone quantizers.
A *deadzone* quantizer is a quantizer that has a quantization interval centered around zero. To see the effects of the deadzone quantizer on SNR consider an i.i.d. discrete-time process $x[n]$ whose values are in the $[-1,1]$ interval. Consider the following uniform 2-bit quantizers for the interval:

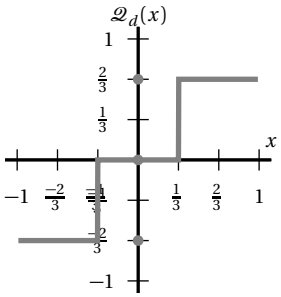
normal quantizer

$$\mathcal{Q}_n(x) = \begin{cases} 3/4 & \text{if } 1/2 \le x \le 1 \\ 1/4 & \text{if } 0 \le x < 1/2 \\ -1/4 & \text{if } -1/2 \le x < 0 \\ -3/4 & \text{if } -1 \le x < -1/2 \end{cases}$$

deadzone quantizer

$$\mathcal{Q}_n(x) = \begin{cases} 2/3 & \text{if } 1/3 \le x \le 1 \\ 0 & \text{if } |x| < 1/3 \\ -2/3 & \text{if } -1 \le x \le -1/3 \end{cases}$$





Both quantizers operate at two bits per sample but the deadzone quantizer "wastes" a fraction of a bit since it has only 3 quantization intervals instead of 4; for a uniformly distributed input, therefore, the SNR of the deadzone quantizer is smaller than the SNR of the standard quantizer. Assume now that the probability distribution for each input sample is the following:

$$P[x[n]=\alpha] = \begin{cases} 0 & \text{if } |\alpha| > 1 \\ p & \text{if } |\alpha| = 0 \\ (1-p)/2 & \text{otherwise} \end{cases}$$

In other words, each sample is either zero with probability p or drawn from a uniform distribution over the $[-1,1]$ interval; we can express this distribution as a pdf like so:

$$f(x) = \frac{1-p}{2} + p\delta(x)$$

Determine the minimum value of p for which it is better to use the deadzone quantizer, i.e. the value of p for which the SNR of the deadzone quantizer is larger that the SNR of the uniform quantizer.

Solution 11.
First of all, since the input signal is the same, in order to compare SNRs we just need to compare the mean square errors of the two quantizers and find out the value of p for which the deadzone quantizer's MSE is smaller than the normal quantizer's MSE. The formula for the MSE of a scalar quantizer over the $[-1,1]$ interval (under the usual hypotheses of iid samples with pdf $f(x)$) is

$$\sigma^2 = \int_{-1}^{1} (\mathcal{Q}(x)-x)^2 f(x)\,dx,$$

For a *uniform* quantizer with M quantization levels and uniform input distribution $f(x)=1/2$, we also know that

$$\sigma^2 = \int_{-1}^{1} (\mathcal{Q}(x)-x)^2 \frac{1}{2}\,dx = \frac{\Delta^2}{12} = \frac{(2/M)^2}{12} = \frac{1}{3M^2}.$$

The number of quantization levels in the two quantizers are $M = M_n = 4$ for the normal 2-bit quantizer and $M = M_d = 3$ for the deadzone quantizer. Let's compute the MSE for the normal quantizer using the composite pdf for the input

$$\begin{aligned}
\sigma_n^2 &= \int_{-1}^{1} (\mathcal{Q}_n(x)-x)^2 \left(\frac{1-p}{2} + p\delta(x)\right) dx \\
&= (1-p)\int_{-1}^{1} (\mathcal{Q}_n(x)-x)^2 \frac{1}{2}\,dx + p\int_{-1}^{1} (\mathcal{Q}_n(x)-x)^2 \delta(x)\,dx \\
&= (1-p)\frac{1}{3M_n^2} + p[Q_n(0)]^2 \\
&= (1-p)\frac{1}{48} + p\frac{1}{16}
\end{aligned}$$

where we have used the fact that the normal quantizer maps zero to $1/4$; similarly, for the deadzone quantizer (which maps zero to zero):

$$\begin{aligned}
\sigma_d^2 &= \int_{-1}^{1} (\mathcal{Q}_d(x)-x)^2 \left(\frac{1-p}{2} + p\delta(x)\right) dx \\
&= (1-p)\frac{1}{3M_d^2} + p[Q_d(0)]^2 \\
&= (1-p)\frac{1}{27}
\end{aligned}$$

from which we find

$$\sigma_d^2 < \sigma_n^2 \quad \text{for} \quad p > \frac{21}{102} \approx 20\%$$
