

Dither

Marcel van de Gevel, second attempt of 30 August 2021

1. Introduction

This document is an attempt to make the most important results of nonsubtractive dither theory plausible while using only very basic mathematics. In section 2, the most important results of nonsubtractive dither theory are summarized. In section 3, they are more or less explained. This document is definitely not meant as a proof or derivation of anything, see references [1] up to and including [6] if you are interested in the real mathematical proof.

The document is only about nonsubtractive dither, as subtractive dither is almost never used in audio. Subtractive dither means that the dither is subtracted again somewhere after quantization. For example, you could make a recording with an ADC that gets dithered with pseudorandom noise, generate the same pseudorandom sequence at playback, convert it to analogue and subtract it from the main DAC's output. It has its theoretical advantages, no noise penalty and no noise modulation, but it's just too inconvenient to be used a lot. An exception will be made for Anagram Technologies 'Sonic Scrambling', this will be discussed in appendix A with lots of guesswork as the details of their technique are not known to me.

Other techniques that will not be discussed are dithering with coloured noise or with low-discrepancy sequences. These are very useful for image processing, but again not used a lot for audio, as far as I know. Noise shaping a dithered quantizer is usually more effective, see [7] and [8], but it is a completely separate topic that is outside the scope of this document.

2. Summary of some results of dither theory

When a signal needs to be rounded off to integer multiples of some quantization step, as happens in analogue to digital converters or when digital signals need to be rounded off to some smaller wordlength, this results in distortion, particularly for small signals. For example, suppose the signal is $0.5 + 0.5 \sin(2\pi ft)$, rounding it off to integers will change the sine wave to a square wave, as illustrated in figure 1.

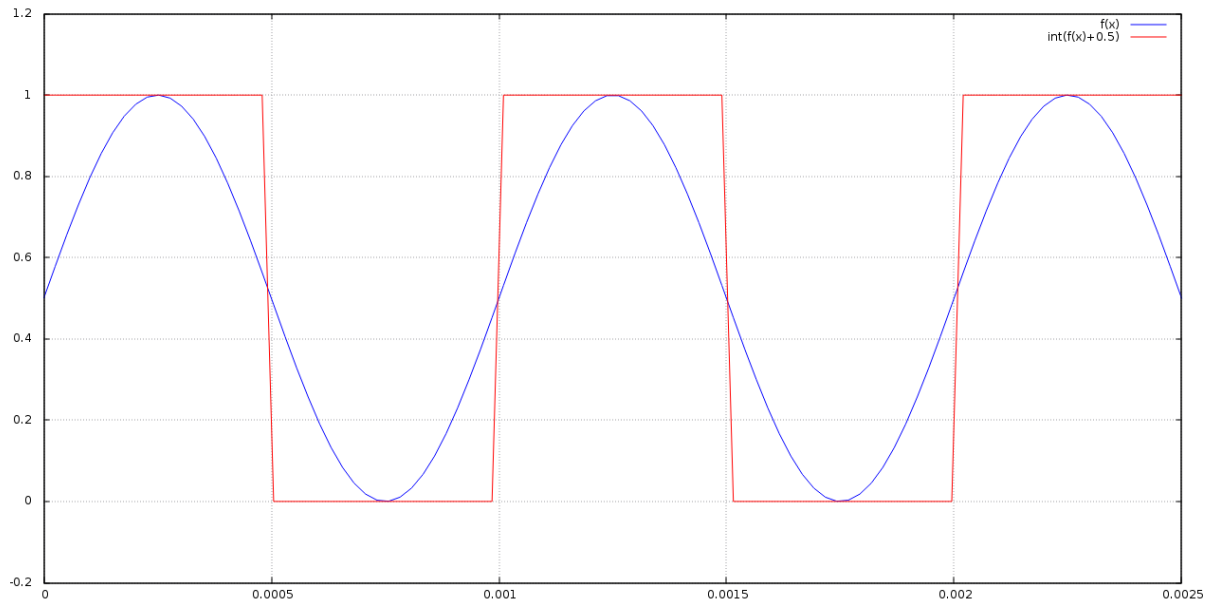


Figure 1: Small 1 kHz sine wave ($0.5 + 0.5 \sin(2 \pi 1000 t)$) and the same rounded to integers

A way to get rid of the most annoying artefacts of quantization or requantization is dithering [1], [2], [3], [4], adding a small noise signal at the input of the quantizer¹. The total error of a dithered quantizer has a mathematical expectation (ensemble average) and a standard deviation that are independent of the signal when the dither has the following characteristics [1]:

A. The dither must have a two-LSB peak-to-peak value and a triangular probability density function²

B. It must be independent of the signal to be quantized.

For a quantizer that is used inside a feedback loop, requirement B implies that:

C. The dither samples must be independent of each other.

There are some special cases where requirement C does not apply, such as loops with integer coefficients. See references [3] and [4] for details.

With dither according to requirements A, B and C, the power spectral density of the rounding error becomes white and independent of the signal. When the quantizer is embedded in a noise shaping loop, the power spectral density at the noise shaper output is shaped according to the noise transfer function, but it is still continuous, without any tones, and independent of the signal. This makes the round-off error sound like (white or coloured) background noise rather than like a very unpleasant kind of distortion.

The efficacy of dither has been grossly over- and understated in some literature, ranging from 'dither makes quantization essentially linear' to 'dither only works on steady-state signals such

¹ Scientists believe that dither-like phenomena play a role in processes as diverse as the periodicity of ice ages and the operation of the mammalian auditory system, the inner hair cells being dithered by the Brownian motion of the fluid in the cochlea. See reference [9] for more information.

² At least this is the option with least noise; in general you have to add two or more random signals with uniform distribution of 1 LSB peak-peak each. Adding precisely two of these signals results in triangular dither.

as sine waves and not on musical transients'. A non-subtractively dithered quantizer is neither linear nor affine, because the quantization error is statistically dependent on the signal. For example, when the input to the quantizer is 12.25 LSB, for any kind of non-subtractive dither, the quantization error will always be $n - 0.25$ LSB with integer n . When the input to the quantizer is 13.31 LSB, the quantization error will be $n - 0.31$ LSB with integer n . Hence, the probability distribution of the quantization error is always dependent on the input signal, even though the average and the standard deviation can be made independent of it.

Claiming that dither only works on steady-state signals is equally incorrect. This impression may have been left by some early articles on dither that for simplicity showed the effect on sine waves or on constant input signals; the later, more general, articles are often difficult to read because of their advanced mathematical content (such as large numbers of high- or even infinite-dimensional Fourier transforms). In any case, it is perfectly possible to calculate the effect of dithering over an ensemble of independently dithered quantizers that all quantize the same musical transient.

3. Heuristic story to make it plausible

3.1. Correcting the expectation (ensemble average) with rectangular dither

Suppose you want to round some real number to an integer. A fractional number can always be written as the sum of an integer part and a fractional part, $n + f$ where n is an integer and $0 \leq f < 1$. The same holds for irrational real numbers, except that you can't call f the fractional part anymore.

When $f \geq 0.5$, the number gets rounded up to $n + 1$, so the round-off error is $1 - f$. When $f < 0.5$, the number gets rounded down to n , so the error is $-f$.

Suppose you add a random variable d (as in dither) that's uniformly distributed on $-0.5 \leq d < 0.5$ to the number $n + f$ before rounding. As $0 \leq f < 1$ and $-0.5 \leq d < 0.5$, the sum of f and d will be in the range $-0.5 \leq f + d < 1.5$. When $-0.5 \leq f + d < 0.5$, then $n + f + d$ will be rounded down to n and when $0.5 \leq f + d < 1.5$, then $n + f + d$ will be rounded up to $n + 1$.

Hence, when $-0.5 \leq f + d < 0.5$, the difference between the dithered and rounded number n and the original number $n + f$ will be $-f$ and when $0.5 \leq f + d < 1.5$, the difference between the dithered and rounded number $n + 1$ and the original number $n + f$ will be $1 - f$.

As d is uniformly distributed with $-0.5 \leq d < 0.5$ and $0 \leq f < 1$, the probability of $f + d$ being greater than or equal to 0.5 is precisely f . For example, when f is 0, $f + d$ must always be less than 0.5. When f is almost 1, $f + d$ is almost always at least 0.5. When f is 0.43, $f + d \geq 0.5$ when $d \geq 0.07$. Hence, $f + d \geq 0.5$ when d is in the range from 0.07 to 0.5, which is 43 % of the interval from -0.5 to 0.5.

As a result, when you have a large number of quantizers (ensemble of quantizers) with independent dither generators all rounding off the number $n + f$, there will be a fraction f of them rounding off the number to $n + 1$ and the rest to n . The ensemble average, also known as the mathematical expectation, will therefore be $n + f$. This means that there is no systematic error anymore, the error due to rounding off has changed into a random error.

When there is a discrete-time signal (that is, a series of sample values) to be rounded off and each sample gets its own d , the error is also random over time and will sound like noise. Hence, the distortion due to round-off errors has been changed into noise.

The exact same result can be obtained with rounding down and dither with $0 \leq d < 1$. This may sometimes be easier to implement.

A uniform probability distribution on some interval can also be called a rectangular probability distribution, hence the title of this section.

3.2. Getting rid of noise modulation with triangular dither

The dither of section 3.1 changes distortion into noise, but the noise gets modulated by the original signal. For example, suppose the signal consists of a constant number with a fractional part $f = 0$, that is, an integer that gets repeated all the time. Adding dither with $-0.5 \leq d < 0.5$ will then always result in a sum $n + f + d$ that gets rounded to n . The output signal will therefore be constant at n and have no noise.

When $f = 0.5$, half the samples will be rounded to $n + 1$ and half to n . When at random half the samples are rounded to $n + 1$ and half to n , there is noise.

Like the round-off error in section 3.1, the noise level is dependent only on the fractional part f of the number, not on the integer part n . This fractional part can be randomized by adding yet another uniformly distributed signal $-0.5 \leq d_2 < 0.5$ to $n + f$ before doing all the things that were described in section 3.1. That is, $n + f + d_2$ can again be written as the sum of an integer part $\text{int}(n + f + d_2)$ and a fractional part $n + f + d_2 - \text{int}(n + f + d_2)$, which is $f + d_2 \bmod 1$. This fractional part is uniformly distributed over the interval from 0 to 1.

For example, when $f = 0$, $n + f + d_2$ will be uniformly distributed between $n - 0.5$ and $n + 0.5$. If $n - 0.5 \leq n + f + d_2 < n$, the integer part will be $n - 1$ and the fractional part will be $n + f + d_2 - (n - 1) = f + d_2 + 1$ with $0.5 \leq f + d_2 + 1 < 1$. If $n \leq n + f + d_2 < n + 0.5$, the integer part will be n and the fractional part will be $n + f + d_2 - n = f + d_2$ with $0 \leq f + d_2 < 0.5$. All in all, the fractional part of $n + f + d_2$ can be anywhere between 0 and 1 and all values are equally probable.

Similarly, when $f = 0.6$, $n + f + d_2$ will be uniformly distributed between $n + 0.1$ and $n + 1.1$. If $n + 0.1 \leq n + f + d_2 < n + 1$, the integer part will be n and the fractional part will be $n + f + d_2 - n = f + d_2$ with $0.1 \leq f + d_2 < 1$. If $n + 1 \leq n + f + d_2 < n + 1.1$, the integer part will be $n + 1$ and the fractional part will be $n + f + d_2 - (n + 1) = f + d_2 - 1$ with $0 \leq f + d_2 - 1 < 0.1$. All in all, the fractional part of $n + f + d_2$ can again be anywhere between 0 and 1 and all values are again equally probable.

As the noise level depends on the fractional part of the input signal of the dithered quantizer of section 3.1, randomizing that fractional part will remove the signal dependence of the noise level.

All in all, when you add two independent random signals uniformly distributed from -0.5 to 0.5 to the signal $n + f$ before rounding it, you get rid of distortion and noise modulation. The probability distribution of the sum of two uniformly distributed random signals is triangular, hence this is known as triangular dithering.

3.3. High-order dither

It is shown in reference [1] that the so-called third and higher moments of the total error are still dependent on the signal with triangular dither. Adding k independent uniformly distributed random signals to the signal before rounding it makes the first k moments of the total error independent of the signal, but the noise level gets higher and higher as k is increased. (Assuming that the quantization error without dither is uniformly distributed over one quantization step, which is not necessarily true, the RMS value of the total error increases with a factor of $\sqrt{k+1}$ with k th order dither.)

It is often claimed that only the first two moments matter for audio and that making the first two moments independent of the signal makes the dithered quantization error indistinguishable from real additive noise. I'm not so sure about that anymore since conducting a listening test on diyaudio.com, see [10]. Still, the participants, Mooly and PMA, did not have a *preference* for real additive noise over triangularly dithered quantization errors, so the conclusion that triangular dither suffices for audio still stands.

According to the central limit theorem, the probability density function of the sum of k independent uniformly distributed random variables approaches a Gaussian distribution as k increases, and it actually does so pretty fast. The standard deviation of the sum increases with the square root of k . Hence, Gaussian noise with an RMS level greater than a few quantization steps is imperfect but pretty good dither, as its distribution is quite close to that of k independent uniformly distributed random variables. Normal analogue circuit noise is usually Gaussian, so analogue circuit noise can be used as imperfect but pretty good dither for an analogue to digital converter if it is large enough compared to the quantization steps (which can be the case for high-resolution SAR ADCs, but is almost never the case for the coarse quantizers of sigma-delta ADCs).

3.4. Statistical independence

There is no way to make the quantization error completely statistically independent of the signal with nonsubtractive dither, as was explained in section 2. For example, when the input signal is 12.25 LSB, for any kind of non-subtractive dither, the total error will always be $n - 0.25$ LSB with integer n . When the input signal is 13.31 LSB, the total quantization error will be $n - 0.31$ LSB with integer n . Hence, the probability distribution of the error is always dependent on the input signal, even though the average and the standard deviation (and any number of higher moments) can be made independent of it.

3.5. Example: 24 bit signal to 16 bit signal

Rounding (requantizing) a 24 bit signal to a 16 bit signal with triangular probability density function dither can be done as follows:

- Generate two independent 8 bit random numbers per sample (8 being 24 - 16)
- Add them to the 24 bit sample value, and correct for offsets if needed. (When you use a signed and an unsigned random number and later round downwards (throw away bits), there is no offset to correct.)
- Take some precautions to clip everything properly if the added dither should cause an over- or underflow
- Throw away the lower eight bits of the sum (8 still being 24 - 16)

4. References

- [1] Robert A. Wannamaker, Stanley P. Lipshitz, John Vanderkooy and J. Nelson Wright, "A theory of nonsubtractive dither", *IEEE Transactions on Signal Processing*, vol. 48, no. 2, February 2000, pages 499...516
- [2] Stanley P. Lipshitz, Robert A. Wannamaker and John Vanderkooy, "Quantization and dither: a theoretical survey", *Journal of the Audio Engineering Society*, vol. 40, no. 5, May 1992, pages 355...375
- [3] Michael A. Gerzon, Peter G. Craven, J. Robert Stuart and Rhonda J. Wilson, "Psychoacoustic noise shaped improvements in CD and other linear digital media", Audio Engineering Society preprint 3501, presented at the 94th convention, March 1993
- [4] Stanley P. Lipshitz, Robert A. Wannamaker and John Vanderkooy, "Dithered noise shapers and recursive digital filters", Audio Engineering Society preprint 3515, presented at the 94th convention, March 1993
- [5] Bernard Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory", *IRE Transactions on Circuit Theory*, December 1956, pages 266...276
- [6] Bernard Widrow, "Statistical analysis of amplitude-quantized sampled data systems", *American Institute of Electrical Engineers, Applications and industry*, vol. 79 part II (1960-1961), January 1961, pages 555...568
- [7] Michael A. Gerzon and Peter G. Craven, "Optimal noise shaping and dither of digital signals", Audio Engineering Society preprint 2822, presented at the 87th convention, October 1989
- [8] Marcel van de Gevel, "The valve DAC - Submicron silicon meets submillimetre vacuum", *Linear Audio*, vol. 13, April 2017, pages 23...77,
<https://linearaudio.net/sites/linearaudio.net/files/03%20Didden%20LA%20V13%20mvdg.pdf>
- [9] Kurt Wiesenfeld and Fernan Jaramillo, "Minireview of stochastic resonance", *Chaos*, vol. 8, no. 3, September 1998, pages 539...548
- [10] High-order dither listening test, <https://www.diyaudio.com/forums/everything-else/313257-dither-listening-test.html>

Appendix A. Sonic scrambling

Anagram Technologies used a technique called sonic scrambling where a differential signal is made by driving two DACs with opposite signals, but with the dither in phase. The idea is that the dither then largely cancels in the differential output, as it mainly causes a common-mode disturbance.

That is, an interpolating filter produces an interpolated large-wordlength signal. Its opposite is calculated, the very same dither signal is added to the interpolated signal and its opposite, both are requantized, and one is sent to the positive DAC and the other to the negative DAC.

A.1. Half dither with ideal DACs

Assuming ideal DACs and ideal subtraction, an interesting property of this technique is that it can work with half the normal dither level.

For example, suppose we use a rectangular dither signal d that is uniformly distributed on $0 \leq d < 0.5$, add it to the signal $n + f$ and its opposite $-n - f$, and round down (floor) the results.

When $0 \leq f < 0.5$, the sum $n + f + d$ will always be greater than or equal to n and smaller than $n + 1$ and will therefore be rounded down to n . However, $-n - f + d$ will be greater than $-n - 0.5$ and smaller than $-n + 0.5$ and can therefore be rounded down to $-n - 1$ or to $-n$. When f is close to 0.5, it will almost certainly be rounded down to $-n - 1$ and when f is 0, it will always be rounded down to $-n$. When f is 0.23, there is a 46 % chance that it will be rounded down to $-n - 1$.

When $0.5 \leq f < 1$, the sum $n + f + d$ will be greater than or equal to $n + 0.5$ and smaller than $n + 1.5$ and can therefore be rounded down to n or to $n + 1$. When f is 0.5, it will always be rounded down to n and when f is close to 1, it will almost always be rounded down to $n + 1$. When f is 0.73, there is a 46 % chance that the sum will be rounded down to $n + 1$. Meanwhile, $-n - f + d$ will be greater than $-n - 1$ and smaller than $-n$ and will therefore always be rounded down to $-n - 1$.

Therefore, even though the dither was uniform from 0 to 0.5 rather than 0 to 1, the mathematical expectation of the difference between the outputs of the two quantizers, and hence of the DACs if they are ideal, will be proportional to $n + f$:

When $0 \leq f < 0.5$:

mathematical expectation of the differential output is $n - (-n - 2f) = 2n + 2f$

When $0.5 \leq f < 1$:

mathematical expectation of the differential output is $n + 2(f - 0.5) - (-n - 1) = 2n + 2f$

Actually subtracting the outputs of two N -bit DACs gives you a total of 2^{N+1} possible levels. If you would apply a half-LSB offset to the requantizer of one of them and not to the other, they would together form an $(N + 1)$ -bit DAC. It is therefore understandable that half of the dither that a single DAC would need can suffice.

Regarding noise modulation, when $f = 0$ or $f = 0.5$ and $0 \leq d < 0.5$, the code for the positive side $n + f + d$ will always be rounded down to n . When $f = 0$ and $0 \leq d < 0.5$, the code for the negative side $-n - f + d$ will always be rounded down to $-n$. When $f = 0.5$ and $0 \leq d < 0.5$, the code for the negative side $-n - f + d$ will always be rounded down to $-n - 1$. This means that the dithered quantizer produces no noise whenever $n + f$ is an integer multiple of $1/2$. That is, noise modulation appears to be periodic with a period of $1/2$ LSB rather than 1 LSB.

Unfortunately it is impossible to get rid of noise modulation by adding the same dither to $n + f$ and to $-n - f$. That's easy to see for integer values of $n + f$, particularly $n + f = 0$. When $n + f = 0$, also $-n - f = 0$, so no matter what the probability density of the dither may be, $n + f + d$ and $-n - f + d$ will always be rounded to the same number, giving a zero and hence noiseless difference.

Noise modulation can be suppressed with some differential dither, though. As the noise modulation appears to be periodic with a period of $1/2$ LSB as a function of $n + f$, we

can make a second dither signal d_2 uniformly distributed on $-0.25 \leq d_2 < 0.25$, calculate $n + f + d_2 + d$ and $-n - f - d_2 + d$ and round those two numbers downwards.

A.2. Large dither to randomize dynamic non-linearity errors of non-ideal DACs

Ideally, adding some random integer to d won't change the differential output voltage, as both the positive and the negative DAC output signals change by the same amount, assuming the DACs aren't driven into clipping. (I have the strong suspicion that adding a random integer multiple of 1/2 to d also won't change anything, but I have not checked that.) Although d will need to have a fractional part to properly dither the requantization as discussed in section A.1, to keep the calculations simple, we will pretend in this section that d is an integer and ignore d_2 .

Real-life DACs will have so-called integral and differential non-linearity and gain errors. To keep things simple, I will only consider differential non-linearity and gain errors.

When the input to a DAC is k , its output signal should be kG , where G is a gain factor. When the positive and negative DACs have gain errors and differential non-linearity, their output signals will be $kG_p + \varepsilon_p(k)$ for the positive DAC and $kG_n + \varepsilon_n(k)$ for the negative DAC. Each DAC has a different G to show that their gains are not exactly the same. The ε 's show that there is some extra error that depends on the DAC and DAC code. I will assume that these extra errors are independent of each other.

Suppose the positive DAC gets an input code $n + d$ and the negative DAC gets a code $-n + d$. The difference between their outputs will then be:

$$(n + d)G_p + \varepsilon_p(n + d) - ((-n + d)G_n + \varepsilon_n(-n + d)) = n(G_p + G_n) + d(G_p - G_n) + \varepsilon_p(n + d) - \varepsilon_n(-n + d)$$

The following subsections will deal with a couple of special cases.

A.2.1. Gain error not taken into account, $d = 0$

If there were no gain error, $G_p = G_n = G$, and d would be 0, the differential output would be $2nG + \varepsilon_p(n) - \varepsilon_n(-n)$, so the sum of the desired signal and two error terms due to differential non-linearity that depend on the sample value n . Suppose you would play a small periodic signal with a period time that's an integer number of sample periods. Every time you play a certain sample value of the periodic waveform, you get the associated error $\varepsilon_p - \varepsilon_n$. The error will therefore also be periodic and can be expressed as a Fourier series. That is, it causes harmonic distortion.

A.2.2. Gain error not taken into account, d random integer

If there were no gain error, $G_p = G_n = G$, and d would be some random integer, the differential output would be $2nG + \varepsilon_p(n + d) - \varepsilon_n(-n + d)$, so the sum of the desired signal and two error terms due to differential non-linearity. Suppose you would play a small periodic signal with a period time that's an integer number of sample periods. Every time you play a certain sample value of the periodic waveform, you get a different error $\varepsilon_p - \varepsilon_n$ because d will be different. The error will therefore be randomized and sound more like noise and less like distortion.

When d can take on a huge number of different integer values and the ε 's are independent with zero mean, the error for each value of n will average out and get quite close to 0. When d can only have a few values, the averaging out won't work so well. For example, when there is a sample $n = 23$ in the periodic waveform and d can only be -2, -1, 0 and 1, $\varepsilon_p(n + d)$ will randomly switch between $\varepsilon_p(21)$, $\varepsilon_p(22)$, $\varepsilon_p(23)$ and $\varepsilon_p(24)$. As there are only four ε_p 's to average out, they will not average out as well as when there had been 100 of them.

All in all, distortion due to differential nonlinearity is partly converted into noise and this works better as the range of dither values is increased.

A.2.3. Gain error taken into account, d random integer

The full expression for the differential output signal is $n(G_p + G_n) + d(G_p - G_n) + \varepsilon_p(n + d) - \varepsilon_n(-n + d)$. Clearly, the term $d(G_p - G_n)$ that represents imperfect noise cancellation due to gain errors results in a noise floor that increases as the range of values for d increases. One therefore needs to make a compromise between the conversion of differential non-linearity errors from distortion to noise (A.2.2) and a noise floor increase.

A.3. Using dynamic element matching / mismatch shaping / data-weighted averaging

In section A.2, distortion due to differential non-linearity was partly converted into noise by adding a random integer part to the dither. It would be nicer if this noise could predominantly be put outside the audio band. As the Anagram Technologies DAC uses oversampling, the obvious place to put it is above the audio band. This is vaguely similar to the dynamic element matching and mismatch shaping techniques used for multibit sigma-delta modulators, though I don't know if any of them are directly applicable, nor if Anagram Technologies used any of those.

A.4. Conclusion

It looks like two DACs used to generate a differential output signal can be effectively dithered with a bit of common-mode and a bit of differential dither. When the only purpose of the dither is to make sure requantization errors sound like noise, the recipe is:

- Generate a uniformly distributed random signal d with $0 \leq d < 0.5$.
- Generate a second uniformly distributed random signal d_2 with $-0.25 \leq d_2 < 0.25$.
- Calculate $n + f + d_2 + d$ and $-n - f - d_2 + d$ and round those two numbers downwards. $n + f$ stands for the signal (integer and fractional part) that is to be processed.
- Drive one DAC with one rounded number and the other with the other rounded number and take the difference between their output signals.

When the purpose is also to randomize the effect of dynamic non-linearity errors of the DACs, the recipe is:

- Generate a uniformly distributed random signal d with $0 \leq d < 0.5$.
- Generate a second uniformly distributed random signal d_2 with $-0.25 \leq d_2 < 0.25$.
- Generate a random integer signal d_3 , the range of integers to be used for d_3 is a compromise between how well dynamic non-linearity errors are randomized and how much noise you get when there are gain errors or the common-mode suppression is imperfect.
- Calculate $n + f + d_2 + d + d_3$ and $-n - f - d_2 + d + d_3$ and round those two numbers downwards. $n + f$ stands for the signal (integer and fractional part) that is to be processed.
- Drive one DAC with one rounded number and the other with the other rounded number and take the difference between their output signals.

Chances are that using integer multiples of $1/2$ for d_3 will also work. In that case, $d + d_3$ can be seen as one uniformly distributed random signal, d_3 representing the bits before and the first bit after the binary point and d the other bits after the binary point.