This note is a followup from a discussion with Rick Regan in the comments section of a StackOverflow question. The aim is to give a short proof of the condition for roundtripping of conversions between two floating-point formats.

1 Preliminaries

We'll work with an idealised form of floating-point that ignores exponent bounds, NaNs, infinities and zeros. Roundtripping for negative floats doesn't behave substantially differently than for positive floats, so we also ignore negative floats and just work with positive floats. We'll only consider round-to-nearest rounding modes between the two formats, and we won't make any assumptions about which direction ties round in.

To be precise, here's what we'll mean by a floating-point number in what follows.

Definition 1. For integers $p \geq 1$ and $B \geq 2$, a **precision-p base-B floating-point number** is a positive rational number x that can be expressed in the form mB^e for some integers m and e, with $0 \leq m < B^p$.

We're interested in what happens when we convert a number in one floating-point format to another format and back again, with each of the two conversions rounding to nearest. Specifically, we want to give necessary and sufficient conditions on the two formats for this double conversion to recover the number we started with.

For the remainder of this note, we fix two floating-point formats. We choose integers $p \geq 1$, $q \geq 1$, $B \geq 2$ and $D \geq 2$. Write $\mathbf{B}$ for the set of all precision-p base-B floats, and $\mathbf{D}$ for the set of all precision-q base-D floats.

Definition 2. We will say that precision-p base-B float x in $\mathbf{B}$ **roundtrips through** $\mathbf{D}$ if the nearest precision-q base-D floating-point value to x rounds back to x. Or in other words, let y be the closest element of $\mathbf{D}$ to x. Then we require that x is the closest element of $\mathbf{B}$ to y. (In the event that x is exactly halfway between two elements of $\mathbf{D}$, we'll require that *both* those elements round back to x.)

We'll also say that the *format* $\mathbf{B}$ roundtrips through $\mathbf{D}$ if every element of $\mathbf{B}$ roundtrips through $\mathbf{D}$.

2 Roundtrip results

Now we can state the main propositions. There's a simple sufficient condition for roundtripping.

Proposition 3. *If $B^p \leq D^{q-1}$ then the format $\mathbf{B}$ roundtrips through $\mathbf{D}$.*

Under one additional hypothesis, this condition is also necessary.

Proposition 4. *If the format $\mathbf{B}$ roundtrips through $\mathbf{D}$,* and *B and D are not powers of a common base, then $B^p \leq D^{q-1}$.*

The *powers of a common base* condition excludes cases like $B = 4$ and $D = 8$, or $B = 5$ and $D = 25$, for example.

To prove the first proposition, we need a pair of lemmas relating to the spacing between successive floating-point numbers.

Lemma 5. *Suppose that x is an element of $\mathbf{B}$, and that $B^{e-1} < x \leq B^e$. Then any positive rational number y satisfying $|x - y| < \frac{1}{2}B^{e-p}$ rounds back to x.*

Proof. This follows from the observation that within the closed interval $[B^{e-1}, B^e]$, successive floats in $\mathbf{B}$ are spaced exactly B^{e-p} apart from one another. Some care must be taken with the corner case where $x = B^e$, where the next float up from x is $B^e + B^{e-p+1}$ rather than $B^e + B^{e-p}$. In particular, note that if we'd stated the condition on e in the form $B^{e-1} \leq x < B^e$, the conclusion of the lemma would be false. □

Lemma 6. *Suppose that x is any positive rational number, that f is the unique integer such that $D^{f-1} < x \leq D^f$, and that y is a closest element of $\mathbf{D}$ to x, choosing either possibility in the case of a tie. Then $|x - y| \leq \frac{1}{2}D^{f-q}$.*

Proof. This follows from the fact that $D^{f-1} \leq y \leq D^f$, and that successive floats in $\mathbf{D}$ are spaced exactly D^{f-q} apart in this interval. □

The proof of the sufficient condition is now straightforward.

Proof of Proposition 3. Let x be any precision-p base-B floating-point number, and let y be a closest precision-q base-D floating-point number to x (picking either one in the case of a tie). Choose integers e and f such that $B^{e-1} < x \leq B^e$ and $D^{f-1} < x \leq D^f$. Then we have:

$$
\begin{aligned}
|x - y| &\leq \frac{1}{2}D^{f-q} && \text{by Lemma 6} \\
&= \frac{1}{2}D^{f-1}D^{1-q} && \\
&< \frac{1}{2}xD^{1-q} && \text{by choice of } f \\
&\leq \frac{1}{2}xB^{-p} && \text{from the assumption that } B^p \leq D^{q-1} \\
&\leq \frac{1}{2}B^eB^{-p} && \text{by choice of } e \\
&= \frac{1}{2}B^{e-p} &&
\end{aligned}
$$

So $|x - y| < \frac{1}{2}B^{e-p}$, and it follows from Lemma 5 that y rounds back to x. □

For the necessary condition, we'll need the following result from elementary number theory.

Fact 7. *Suppose that B and D are integers larger than 1, and that B and D are not powers of a common base (or equivalently, $log(B)/log(D)$ is not rational). Then numbers of the form B^e/D^f are* dense *in the positive reals: any open subinterval (a, b) of the positive reals contains at least one such number.*

Proof. Take logs to base D and apply Kronecker's approximation theorem. □

Now we can prove the necessary condition.

Proof of Proposition 4. We prove the contrapositive: that if $B^p > D^{q-1}$ then there's at least one element x of **B** that fails to roundtrip through **D**.

To provide some intuition: we're looking for a region of the positive reals where the gap between successive elements of **B** is *smaller* than the gap between successive elements of **D**. In relative terms, the gap between successive elements of **B** is smallest just *before* a power of B, while the gap between successive elements of **D** is largest just *after* a power of D. So if there's an x that fails to roundtrip, a good place to look for it would be in an interval $[D^f, B^e]$ where D^f and B^e are very close to one another. In such an interval, the gap between successive elements of **D** is D^{f-q+1}, while the gap between successive elements of **B** is B^{e-p}. Our hypothesis that $B^p > D^{q-1}$ implies that $1 < B^p/D^{q-1}$, so by Fact 7 we should be able to find e and f so that $1 < B^e/D^f < B^p/D^{q-1}$. Then our gaps satisfy $B^{e-p} < D^{f-q+1}$, as required.

Making the above rigorous is a tiny bit fiddly: if we make D^f *too* close to B^e, then the interval (D^f, B^e) is too small to contain any floats, making it hard to find a suitable x. So we'll also need to give a lower bound on how close D^f and B^e get.

It turns out that we can always find a roundtrip failure of the form $x = B^e$: we'll arrange that x rounds down to an exact power of D, D^f, and then that *that* power of D in turn rounds *down* again to something smaller than B^e, breaking roundtripping.

For B^e to round down to D^f, we need:

$$D^f < B^e < D^f + \frac{1}{2}D^{f-q+1},$$

while for D^f to round down to some value strictly smaller than B^e, we need

$$D^f < B^e - \frac{1}{2}B^{e-p}.$$

Rearranging the inequalities above gives the condition

$$\frac{1}{1 - \frac{1}{2}B^{-p}} < B^e/D^f < 1 + \frac{1}{2}D^{1-q}.$$

So *if* we can find integer exponents e and f such that the above holds, then $x = B^e$ gives us a counterexample to roundtripping.

But from Fact 7, we can find a number of the form B^e/D^f in *any* open subinterval of the positive reals. The only thing we need to check is that the two inequalities above are compatible; that is, that

$$\frac{1}{1-\frac{1}{2}B^{-p}} < 1 + \frac{1}{2}D^{1-q},$$

so that the lower and upper bounds really do form a subinterval of the reals. But the above inequality can be rearranged to the equivalent inequality

$$\frac{1}{2} < B^p - D^{q-1},$$

which is true from our assumption that $B^p > D^{q-1}$.

So this completes the proof: find e and f satisfying the above inequality, and then $x = B^e$ will fail to roundtrip. □

Example 8. Consider precision-53 binary (as used in double-precision IEEE 754 floating-point, for example). We have

$$2^{53} \leq 10^{17-1}$$

and it follows that conversion from a finite IEEE 754 binary64 float to a decimal string with 17 significant digits produces a value that rounds back to the original float. 16 significant digits are insufficient.

Example 9. Consider conversions from 10-bit binary to 4-digit decimal and back again. The necessary and sufficient condition for roundtripping is that

$$2^{10} \leq 10^{4-1},$$

which is false (but only just). So roundtripping does not hold, and we should be able to find an explicit 10-bit binary float which fails to round-trip through 4-digit decimal.

Following the above proof, we look for exponents e and f such that

$$\frac{1}{1-\frac{1}{2}2^{-10}} < \frac{2^e}{10^f} < 1 + \frac{1}{2}10^{-3}$$

which simplies to the condition

$$\frac{2048}{2047} < \frac{2^e}{10^f} < \frac{2001}{2000}.$$

The smallest positive exponents satisfying this pair of inequalities are $f = 1929$ and $e = 6408$ (in fact, this f is the *only* positive value smaller than 10000 for which there's a solution). So $x = 2^{6408}$ fails to roundtrip: the nearest 4-digit decimal to x is 10^{1929}, and the nearest 10-bit binary value to 10^{1929} is $2^{6408} - 2^{6398}$.

To complete the picture, we're missing one special case. I'll state it here without proof for now.

Proposition 10. *Suppose that B and D are powers of a common base C, and choose C to be the maximal possible common base. The the floating-point format* **B** *roundtrips through* **D** *if and only if*

$$B^p \leq CD^{q-1}.$$

Example 11. In the special case where $B = D$, we have $C = B = D$, and the proposition above tells us that **B** roundtrips through **D** if and only if $p \leq q$, which is exactly what we'd expect.