# Roundtrip conversions between floating-point formats

Mark Dickinson

May 2, 2016

This note is a followup from a discussion with Rick Regan in the comments section of a Stack Overflow question[1]. The aim is to give a short proof of the condition for roundtripping of conversions between two floating-point formats.

Note: all results presented here are well-known. The proofs of the main propositions are, as far as I know, original.

## 1 Preliminaries

We work with an idealised form of floating-point that ignores exponent bounds, NaNs, infinities and zeros. Roundtripping for negative floats behaves entirely analogously to that for positive floats, so for simplicity we only consider positive floats in what follows. We consider only round-to-nearest rounding modes between floating-point formats, and we make no assumptions about which direction ties round in, so our results are equally applicable to the round-ties-to-even and the round-ties-away-from-zero rounding modes, for example.

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

**Definition 1.** Given 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 $x = mB^e$ 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. 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 say that a 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 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 Necessary and sufficient conditions

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 $B = 25$ and $D = 5$.

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

**Lemma 5.** *Suppose that $x$ is an element of $\mathbf{B}$, and that $B^{e-1} < x \leq B^e$ for some integer $e$. 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 $[B^{e-1}, B^e]$, adjacent floats in $\mathbf{B}$ are spaced exactly $B^{e-p}$ apart from one another, so if $y$ is within the interval $[B^{e-1}, B^e]$ then $x$ must be the closest element of $\mathbf{B}$ to $y$, and there's nothing more to show. The only way for $y$ to *not* be in the interval is if $x = B^e$ and $B^e < y < B^e + \frac{1}{2}B^{e-p}$. But then the next element of $\mathbf{B}$ up from $x$ is $B^e + B^{e-p+1}$, and again $x$ is the closest element of $\mathbf{B}$ to $y$. □

Note that in the above, the strictness of the inequalities in our condition on $e$ is important: if we replace the condition $B^{e-1} < x \leq B^e$ with $B^{e-1} \leq x < B^e$, the conclusion is no longer valid, since there will be values $y$ just below $B^{e-1}$ which round to the next float down from $B^{e-1}$.

**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.* Given our target interval $(a, b)$, it's enough to show that the interval $(1, b/a)$ contains at least one such number, $\gamma$ say. Given such a $\gamma$, there must be a power of $\gamma$ within $(a, b)$. (Take $k = \lfloor \log_\gamma a \rfloor$, then $\gamma^k \leq a < \gamma^{k+1}$, and $\gamma^{k+1} < (b/a)\gamma^k \leq (b/a)a = b$. So $\gamma^{k+1}$ lies in the interval $(a, b)$.)

So it's enough to prove the fact in the case $a = 1$. Suppose, for a contradiction, that the interval $(1, b)$ contains no elements of the form $B^e/D^f$. Then the set of all elements of the form $B^e/D^f$ is discrete (if $\gamma$ is any such element, then there can be no other elements within the interval $(\gamma/b, \gamma b)$), and so amongst all elements of the form $B^e/D^f$ larger than 1 there must be a smallest such element; call it $C$. Now *all* elements of the form $B^e/D^f$ must be integral powers of $C$ (let $\gamma$ be any such element, then $1 \leq \gamma/C^{\lfloor \log_C(\gamma) \rfloor} < C$; this contradicts the choice of $C$ as the smallest element *unless* we have $1 = \gamma/C^{\lfloor \log_C(\gamma) \rfloor}$), so we have $B = C^m$ and $D = C^n$ for some positive integers $m$ and $n$. Remembering that $C$ is rational, it follows that $C$ must be an integer, hence that $B$ and $D$ are powers of a common base, contradicting our assumptions. □

Now we can prove the necessary condition. We'll show the contrapositive, namely that if $B^p > D^{q-1}$ then there's some element $x$ of $\mathbf{B}$ that fails to roundtrip through $\mathbf{D}$. Before we embark on the proof proper, let's sketch the main idea. To find our roundtrip counterexample $x$, we look for a region of the positive reals where the gap between successive elements of $\mathbf{B}$ is *smaller* than the gap between successive elements of $\mathbf{D}$. In relative terms, the gap between successive elements of $\mathbf{B}$ is smallest just *before* a power of $B$, while the gap between successive elements of $\mathbf{D}$ 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 $\mathbf{D}$ is $D^{f-q+1}$, while the gap between successive elements of $\mathbf{B}$ 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 this idea rigorous turns out to be a bit fiddly: just because the gaps work out nicely in a particular interval, that doesn't guarantee that we can find a suitable $x$ in that interval: the interval might be too small to contain *any* suitable values $x$.

It turns out that we can always find a roundtrip failure $x$ of the form $x = B^e$: by careful choice of $e$ and $f$, we'll be able to find an interval $[D^f, B^e]$ so that

1. $B^e$ is sufficiently close to $D^f$ that $B^e$ rounds down to $D^f$ when converting from $\mathbf{B}$ to $\mathbf{D}$, but
2. $D^f$ is sufficiently far away from $B^e$ that $D^f$ rounds down to some value *smaller* than $B^e$ when converting back from $\mathbf{D}$ to $\mathbf{B}$.

Then $x = B^e$ gives us our roundtrip failure. Here's the formal proof.

*Proof of Proposition 4.* We prove the contrapositive: that if $B^p > D^{q-1}$ then there's at least one element $x$ of $\mathbf{B}$ that fails to roundtrip through $\mathbf{D}$. We'll find such an $x$ of the form $x = B^e$, and show that there exist integers $e$ and $f$ such that $B^e$ rounds down to $D^f$, but $D^f$ rounds down again to some value smaller than $B^e$.

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 $2^{6408} - 2^{6398}$.

There are other, smaller values in **B** that fail to roundtrip: the first such value larger than 1 is $1017 \cdot 2^{774}$, which lies in the interval $(10^{236}, 2^{784})$; values from **D** have a spacing of $10^{233}$ in this interval, while values from **B** are spaced $2^{774}$ apart; we have $10^{233}/2^{774} = 1.006429...$.

For floats larger than 1, the *first* interval where floats from **B** are spaced more closely than floats from **D** is $(10^{31}, 2^{103})$, but it turns out that there are no roundtrip counterexamples in this interval.

To complete the picture, we're missing one special case.

**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*

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

*Proof.* For sufficiency, we prove something more, namely that if

$$B^p \le CD^{q-1}$$

then every element $x$ of **B** is already an element of **D**. Then roundtripping follows immediately.

Given an element $x$ of **B**, we can write

$$x = mB^e$$

for some positive integer $m < B^p$ and integer $e$. Now choose $f$ such that

$$D^f \le B^e < D^{f+1},$$

then we can write

$$x = (mB^e D^{-f})D^f.$$

To show that $x$ is representable in **D**, we need to show that the coefficient $mB^e D^{-f}$ of $D^f$ is an integer, and that it's strictly less than $D^q$. It's an integer because $B^e D^{-f}$ is a power of $C$ that's larger than 1. To see the bound, first note that our choice of $f$ ensures that $CB^e \le D^{f+1}$. Now

$$\begin{aligned} mB^e D^{-f} &< B^{p+e} D^{-f} \\ &\le CD^{q-1} B^e D^{-f} \\ &\le D^{f+1} D^{q-1} D^{-f} \\ &= D^q, \end{aligned}$$

as required.

For the converse, suppose that $B^p > CD^{q-1}$. Choose integer exponents $e$ and $f$ such that $B^e/D^f = C$, and consider the interval $[D^f, B^e]$. The endpoints of the interval are representable in both formats simultaneously, so the interval maps to itself under rounding in either direction.

So to show that roundtripping fails, it's enough to show that there are more **B**-floats in this interval than **D**-floats, or equivalently, that the **B**-floats are denser than the **D**-floats.

But the spacing between any two adjacent **B**-floats in this interval is $B^{e-p}$, while the spacing between any two adjacent **D**-floats is $D^{f-q+1}$, and we have:

$$\begin{aligned} B^{e-p} &= CD^f B^{-p} \\ &< D^{1-q} D^f \\ &= D^{f-q+1}. \end{aligned}$$

This completes the proof. □

**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.

# References

[1] "How to find an original string representation for lower precision float values in Python?", `http://stackoverflow.com/a/35708911/270986`