Relative error due to a single bit-flip in floating-point arithmetic

Bradley R. Lowery

May 11, 2014

We consider the error due to a single bit-flip in a floating point number. We assume IEEE 754 double precision arithmetic, which encodes binary floating point numbers in a 64-bit word [1]. We assume that the bit-flip happens randomly so it has equi-probability (1/64) to hit any of the 64 bits. Since we want to mitigate the assumption on our initial floating-point number, we assume that it is uniformly picked among all normalized number. With this framework, we can summarize our findings as follows.

The probability for a single bit flip to cause a relative error less than $10^{-11}$ in a normalized floating-point number is above 25%; The probability for a single bit flip to cause a relative error less than $10^{-6}$ in a normalized floating-point number is above 50%; Etc.

Similar studies could be done for binary floating point numbers in a 32-bit word, or complex numbers; (so, s,c,d,z in LAPACK or BLAS notation); multiple bit-flips; or burst of bit-flips (that is multiple bit-flips which happen contiguously in the binary word).

A 64-bit word consists of 3 fields: a sign bit ($s$), a biased 11 bit exponent ($e$), and a 52 bit fraction ($f$). Excluding exceptions (to be explained) the encoded number is

$$(-1)^s(1.f)2^{e-1023}.$$ 

There are three exceptions: not a number (NAN), infinity (INF), and denormalized numbers. NANs are represented by the biased exponent of all ones and any nonzero fraction (sign does not matter). INFs are represented by the biased exponent of all ones and the zero fraction (positive or negative infinity is possible). Denormalized numbers have a biased exponent of all zeros. In this case the number represented is $(-1)^s(0.f)2^{1022}$. Zero is a special case of a denormalized number where the fraction is zero. We refer the reader to Overton [2] for more details on IEEE arithmetic.

We only make one mild assumption on the initial floating-point number. For example, we assume it is a normalized number and then derive probabilities from that assumption. In that case, we assume that our initial floating-point number is randomly picked among all normalized number. In some cases, one could derive more information if one assumes more on the initial floating-point number. For example, if the number is $2.0e+00$, then the probability that a bit-flip changes it to INF is zero. That would be a different study.

Let $x$ be the original 64-bit word and $x'$ be the 64-bit word after a single bit-flip between $x$ and $x'$ occurred. We will also denote the fields of $x'$ as $s'$, $e'$, and $f'$ when needed. Also, $e(k)$ is the $k$-th entry of the exponent field, and similar notation for other fields.

We begin by calculating the probability that $x'$ is normalized, NAN, INF, or denormalized, given that $x$ is normalized, NAN, INF, or denormalized. Table 4 summarizes the results. The narrative below explains where the results come from.
1. Let \( x \) be a normalized number. Therefore, the exponent is anything other than all ones or all zeros. For \( x' = \text{NAN} \) or \( x' = \text{INF} \) the exponent must be all ones. There are 11 exponents corresponding to normalized numbers that are susceptible to becoming all ones due to a bit-flip (bit strings of all ones except 1 zero). The probability that the zero bit in the exponent is flipped is \( \frac{1}{64} \). The result will be an INF only if \( f = 0 \). Therefore we have

\[
\Pr(\text{normalized to NAN}) = \left( \frac{11}{2^{11} - 2} \right) \left( \frac{1}{64} \right) \left( 1 - \frac{1}{2^{52}} \right) \approx 8.4005 \times 10^{-5}
\]

and

\[
\Pr(\text{normalized to INF}) = \left( \frac{11}{2^{11} - 2} \right) \left( \frac{1}{64} \right) \left( \frac{1}{2^{52}} \right) \approx 1.8653 \times 10^{-20}.
\]

For \( x' \) to be a denormalized number the exponent must be all zeros. There are 11 exponents corresponding to normalized numbers that are susceptible to becoming all zeros due to a bit-flip (bit strings of all zeros except 1 one). The probability that the one bit in the exponent is flipped is \( \frac{1}{64} \). For all factions the result will be a denormalized number. Therefore we have

\[
\Pr(\text{normalized to denormalized}) = \left( \frac{11}{2^{11} - 2} \right) \left( \frac{1}{64} \right) \approx 8.4005 \times 10^{-5}.
\]

Finally, we consider the probability that \( x' \) is a normalized number. If the exponent of \( x \) is not one of the 22 susceptible values then \( x' \) will be a normalized number. \( x' \) will also be a normalized number if the exponent is one of the susceptible exponents but the bit-flip is not in the corresponding entry.

\[
\Pr(\text{normalized to normalized}) = 1 - \left( \frac{22}{2^{11} - 2} \right) \left( \frac{1}{64} \right) \approx 9.9983 \times 10^{-1}.
\]

2. Let \( x = \text{NAN} \). Therefore, \( e = (1\ldots1)_2 \) and \( f \neq 0 \). \( x' \) will only be a normalized number if a bit in the exponent is flipped. Therefore,

\[
\Pr(\text{NAN to normalized}) = \frac{11}{64} \approx 1.7188 \times 10^{-1}.
\]

There are 52 fractions that are susceptible to becoming all zeros (bit strings of all zeros except 1 one). The probability that the 1 bit is flipped is \( \frac{1}{64} \). If this occurs the result will be an INF, therefore

\[
\Pr(\text{NAN to INF}) = \left( \frac{52}{2^{52}} \right) \left( \frac{1}{64} \right) \approx 1.8041 \times 10^{-16}.
\]

A single bit-flip can not produce a denormalized number, therefore

\[
\Pr(\text{NAN to denormalized}) = 0.
\]

The remaining cases will result in a NAN, therefore

\[
\Pr(\text{NAN to NAN}) = 1 - \frac{11}{64} - \left( \frac{52}{2^{52}} \right) \left( \frac{1}{64} \right) \approx 8.2812 \times 10^{-1}.
\]
Table 1: Probability of a 64 bit word changing type due to a single bit-flip.

|            | Normalized | Denormalized | NAN          | INF          |
|------------|------------|--------------|--------------|--------------|
| Normalized | $9.9983 \times 10^{-1}$ | $8.4005 \times 10^{-5}$ | $8.4005 \times 10^{-5}$ | $1.8653 \times 10^{-20}$ |
| Denormalized | $1.7188 \times 10^{-1}$ | $8.2812 \times 10^{-1}$ | 0          | 0           |
| NAN        | $1.7188 \times 10^{-1}$ | 0          | $8.2812 \times 10^{-1}$ | $1.8041 \times 10^{-16}$ |
| INF        | $1.7188 \times 10^{-1}$ | 0          | $8.1250 \times 10^{-1}$ | $1.5625 \times 10^{-2}$ |

Relative error

Normalized Numbers

When $x$ and $x'$ are normalized or denormalized numbers we would like to calculate the relative error in $x'$ as well as the probability that the relative error is below a given threshold.

Let $x$ and $x'$ be normalized numbers. Then we have the following cases.

1. If the flipped bit is in $s$, then $x' = -x$ and
   \[
   \frac{|x - x'|}{|x|} = 2.
   \]

2. If the flipped bit is the $k$-th entry of $f$, then
   \[
   |x - x'| = (0.0\ldots01\ldots0)_22^{e-1023} = (1.0)_2(2^{-k})2^{e-1023}.
   \]
Hence,
\[
\frac{|x - x'|}{|x|} = \frac{(1.0)_{2}}{(1.0)_{2}} (2^{-k}).
\]
Since \(1/2 < 1/(1.f)_{2} \leq 1\), we have
\[
2^{-k-1} < \frac{|x - x'|}{|x|} \leq 2^{-k}.
\]

3. If the flipped bit is in \(k\)-th entry of \(x\), then
\[
|x - x'| = |(1.f)_{2}(2^{-1023})(2^{e} - 2^{e'})|
\]
\[
= |(1.f)_{2}(2^{-1023})(2^{e}) ||1 - 2^{e'-e}| = |x||1 - 2^{e'-e}|.
\]

(a) If \(e(k) = 0\), hence \(e' > e\), then \(e' - e = 2^{11-k}\).
(b) If \(e(k) = 1\), hence \(e' < e\), then \(e' - e = -2^{11-k}\).

Therefore,
\[
\frac{|x - x'|}{|x|} = \begin{cases} 
2^{211-k} - 1 & \text{if } x(k) = 0, \\
1 - 2^{211-k} & \text{if } x(k) = 1.
\end{cases}
\]

Note that \(2^{211-k} - 1 \geq 1\) and \(1/2 \leq 1 - 2^{-211-k} < 1\) for \(k = 1, \ldots, 11\).

Let \(x\) be a normalized number and \(x'\) be a denormalized number. This can only happen if the bit-flip occurs in the exponent. The error is
\[
|(-1)^{s}(1.f)_{2}2^{e-1023} - (-1)^{s}(0.f)_{2}2^{-1022}| = |x||1 - (0.f)_{2}2^{1-e}|
\]
Since, \(0 \leq (0.f)_{2}/(1.f)_{2} < 1/2\) and \(e = 2^{11-k}\)
\[
\frac{1}{2} < 1 - 2^{-e} < \frac{|x - x'|}{|x|} \leq 1.
\]

The following theorem summarizes the results.

Theorem 1. Let \(x\) is a normalized number. The relative error due to a single bit-flip is
\[
\frac{|x - x'|}{|x|} = \begin{cases} 
2 & \text{if bit-flip in } s, \\
2^{-k-1} < \frac{|x - x'|}{|x|} \leq 2^{-k} & \text{if bit-flip in } f, \\
2^{211-k} - 1 \geq 1 & \text{if bit-flip in } e \text{ and } e(k) = 0, \\
1/2 \leq \frac{|x - x'|}{|x|} = 1 - 2^{-211-k} < 1 & \text{if bit-flip in } e, e(k) = 1, \text{ and } e' \neq 0 \\
1/2 \leq 1 - 2^{-211-k} < \frac{|x - x'|}{|x|} \leq 1 & \text{if bit-flip in } e \text{ and } e' = 0,
\end{cases}
\]
where \(k\) is the location of the bit-flip in the given field.

Using these results, assuming \(x\) is a random normalized number, we can calculate the probability for the relative error to be in the following intervals
1. \( \Pr(|x - x'|/|x| \geq 1) = 13/128 + (11/(2046))2^{-52}(1/64) \approx 0.10156. \) This is mainly due to the probability of a bit-flip in \( s \) and in \( e \) (flipping 0 to 1). See (1) and (3). The probability of each of these cases is 1/64 and 11/128, respectively. There is also a possibility for the relative error to be 1 if \( e' = (0 \ldots 0) \) and \( f = 0 \), hence \( x' = 0 \). This is a special case of (3). This will occur with a probability of \((11/2046)(1/64)2^{-52}\).

2. \( \Pr(1/2 < |x - x'|/|x| < 1) = 11/128 - ((2^{10} - 2)/2046)(1/64) - (11/(2046))2^{-52}(1/64) \approx 0.078133. \) This is mainly due to the probability of a bit-flip in \( e \) (flipping 1 to 0), which occurs with probability 11/128. See (1) and (3). However, these include the special case when the error is 1/2 and 1, which we do not include. The error will be 1/2 from this case if the first bit of the exponent is flipped, but the resulting number must still be normalized. This occurs with probability \((2^{10} - 2)/2046)(1/64)\). The error will be 1 for this case if \( x' = 0 \), which occurs with probability \((11/2046)(1/64)2^{-52}\).

3. \( \Pr(|x - x'|/|x| \leq 1/2) = 52/64 + (2^{10} - 2)(2046)(1/64) \approx 0.82030. \) The relative error will always be less than or equal to 1/2 if the bit-flip occurs in \( f \) (see (2)), which occurs with probability 52/64. Including the special case from (4) such that the relative error is 1/2 gives the stated probability.

4. \( \Pr(|x - x'|/|x| \leq 2^{-i}) = (53 - i)/64 \) for \( i = 2, \ldots, 52 \). The relative error will always be less than or equal to \( 2^{-i} \) if the bit-flip occurs in the \( j \)-th entry of \( f \) \((j \geq i)\). See (2). There are 53 – \( i \) entries of \( f \) greater than or equal to \( i \).

We can bound the probability that the error is less than some tolerance using the fourth case above. Below are the bounds on the probability for some error tolerances.

\[
\begin{align*}
\Pr(|x - x'|/|x| \leq 10^{-1}) &< 0.78125 \\
\Pr(|x - x'|/|x| \leq 10^{-2}) &< 0.73438 \\
\Pr(|x - x'|/|x| \leq 10^{-3}) &< 0.68750 \\
\Pr(|x - x'|/|x| \leq 10^{-4}) &< 0.62500 \\
\Pr(|x - x'|/|x| \leq 10^{-5}) &< 0.57812 \\
\Pr(|x - x'|/|x| \leq 10^{-6}) &< 0.53125 \\
\Pr(|x - x'|/|x| \leq 10^{-7}) &< 0.46875 \\
\Pr(|x - x'|/|x| \leq 10^{-8}) &< 0.42188
\end{align*}
\]

**Denormalized Numbers**

Let \( x \) be a denormalized number. If the bit-flip occurs in \( s \), then \( x' \) is also a denormalized number and the error is the same as in the normalized case:

\[
\frac{|x - x'|}{|x|} = 2.
\]

If the bit-flip occurs in \( k \)-th entry of \( f \), then \( x' \) will again be a denormalized number and

\[
|x - x'| = (0.00101\ldots0)2^{-1022} = (1.0)2(2^{-k})2^{-1022}.
\]
Hence,
\[
\frac{|x - x'|}{|x|} = \frac{(1.0)_2}{(0.f)_2} (2^{-k}).
\]
Since \(1 < 1/(0.f)_2 \leq 2^{52}\), we have
\[
2^{-k} < \frac{|x - x'|}{|x|} \leq 2^{52-k}.
\]
To improve these bounds we must be more specific on \(f\). Let the first nonzero entry in \(f\) be the \(t\)-th entry. Then \(2^{t-1} < 1/(0.f)_2 \leq 2^t\) and
\[
2^{t-k-1} < \frac{|x - x'|}{|x|} \leq 2^{t-k}.
\]
Finally, if the bit-flip occurs in then \(k\)-th entry of \(e\), then \(x'\) is a normalized number and
\[
|x - x'| = |(0.f)_2 2^{-1022} - (1.f)_2 2^{e' - 1023}|
\]
\[
= |2^{-1022}((0.f)_2 - 2^{-1}(1.f)_2 2^{e'})|
\]
\[
= |2^{-1022}((0.f)_2 - (0.f)_2 2^{e'})|
\]
\[
= |2^{-1022}(0.f)_2(1 - 2^{e'})|
\]
\[
= |x||1 - 2^{e'}|
\]
\[
= |x||1 - 2^{2^{11-k}}|.\]

For the last line we use \(e' = 2^{11-k}\). This is the same error as in the normalized case when the bit-flip in the exponent is from 0 to 1. The following theorem summarizes the results.

**Theorem 2.** Let \(x\) is a denormalized number. The relative error due to a single bit-flip is
\[
|x - x'|/|x| = 2 \quad \text{if bit-flip in } s, \quad (6)
\]
\[
2^{t-k-1} < |x - x'|/|x| \leq 2^{t-k} \quad \text{if bit-flip in } f, \quad (7)
\]
\[
|x - x'|/|x| = 2^{2^{11-k}} - 1 \geq 1 \quad \text{if bit-flip in } e, \quad (8)
\]
where \(k\) is the location of the bit-flip in the given field and \(t\) is the location of the first nonzero entry in \(f\).

**References**

[1] IEEE standard for binary floating-point arithmetic. ANSI/IEEE Std 754-1985, 1985.

[2] M.L. Overton, Society for Industrial, and Applied Mathematics. *Numerical Computing with IEEE Floating Point Arithmetic: Including One Theorem, One Rule of Thumb, and One Hundred and One Exercises*. Society for Industrial and Applied Mathematics, 2001.