Density results on floating-point invertible numbers

G. Hanrot\textsuperscript{a,}\textsuperscript{*}, J. Rivat\textsuperscript{b}, G. Tenenbaum\textsuperscript{b}, P. Zimmermann\textsuperscript{a}

\textsuperscript{a}INRIA Lorraine, Technopôle de Nancy-Brabois, 615 rue du Jardin Botanique, F-54602 Villers-lès-Nancy Cedex, France

\textsuperscript{b}Institut Élie Cartan, Université Henri Poincaré, B.P. 239, F-54506 Vandœuvre-lès-Nancy, France

Received 5 January 2001; received in revised form 18 April 2001; accepted 5 December 2001

Abstract

Let $F_k$ denote the $k$-bit mantissa floating-point (FP) numbers. We prove a conjecture of Muller according to which the proportion of numbers in $F_k$ with no FP-reciprocal (for rounding to the nearest element) approaches $\frac{1}{2} - \frac{3}{2}\log_2\frac{3}{2} \approx 0.06847689$ as $k \to \infty$. We investigate a similar question for the inverse square root.

Keywords: Floating-point number; Reciprocal; Inverse square root

1. Introduction

For integer $k \geq 3$, we consider the set $F_k$ of exponent-unbounded, $k$-bit mantissa, binary floating-point (FP) numbers, viz.

$$F_k := \{ m2^e : m, e \in \mathbb{Z}, \ 2^{k-1} \leq m < 2^k \} \cup \{0\}.$$ 

The result of an arithmetic operation with input values in $F_k$ does not necessarily belong to $F_k$. Therefore, it needs to be rounded. The IEEE-754 standard defines four different rounding modes. In this article, we only consider rounding real numbers $x$ to their nearest element in $F_k$, noted $x_{(k)}$. In case $x$ is the exact mean of two consecutive elements of $F_k$, $x_{(k)}$ is defined as the neighbour with even $m$. In particular, we have $x_{(k)} = 1$ if, and only if, $1 - 2^{-k-1} \leq x < 1 + 2^{-k}$. 

* Corresponding author.

E-mail addresses: guillaume.hanrot@loria.fr (G. Hanrot), rivat@iecn.u-nancy.fr (J. Rivat),
tenenb@iecn.u-nancy.fr (G. Tenenbaum), paul.zimmermann@loria.fr (P. Zimmermann).

0304-3975/02/$-$see front matter © 2002 Elsevier Science B.V. All rights reserved.

PII: S0304-3975(02)00222-0
We say that \( x \in F_k \) admits an **FP-reciprocal** if there exists \( y \in F_k \) such that \((xy)_{(k)} = 1\). One might expect that any \( x \in F_k \) admits exactly one FP-reciprocal. However, Muller showed in [4] that some elements of \( F_k \) have no FP-reciprocal (as \( \frac{3}{2} \) in \( F_6 \)), while others have two—as \( \frac{3}{2} \) in \( F_5 \), which admits \( \frac{1}{2} \) and \( \frac{1}{3} \) as FP-reciprocals. There cannot be more than two FP-reciprocals. Muller conjectures further that the proportion of numbers in \( F_k \) without an FP-reciprocal converges to \( \frac{1}{2} - \frac{3}{2} \log \frac{4}{3} \) as \( k \) tends to infinity. We prove this conjecture in a quantitative way.

**Theorem 1.** For \( r = 0, 1, 2 \), let \( \gamma_r(k) \) denote the number of \( x \in F_k \cap [1, 2[ \) having exactly \( r \) FP-reciprocals. Then

\[
\gamma_0(k)/2^{k-1} = \frac{1}{2} - \frac{3}{2} \log \frac{4}{3} + O(2^{-k/3}) = 0.0684768917 \ldots + O(2^{-k/3}),
\]

\[
\gamma_1(k)/2^{k-1} = 1 - \frac{3}{2} \log \frac{9}{8} + O(2^{-k/3}) = 0.8233254464 \ldots + O(2^{-k/3}),
\]

\[
\gamma_2(k)/2^{k-1} = -\frac{1}{2} + \frac{3}{2} \log \frac{3}{2} + O(2^{-k/3}) = 0.1081976622 \ldots + O(2^{-k/3}).
\]

Muller also considered the problem of finding, for given \( x \) and \( z \) in \( F_k \), an element \( y \in F_k \) such that \((xy)_{(k)} = z\). This is solved by Theorem 1 when \( z \) is a power of 2, and our argument can easily be adapted to handle the general case.

Our proof of Theorem 1 relies on Lemma 1 below, which is also the key argument for the modern proof of Voronoi’s formula [12] on the divisor problem, viz.

\[
\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O(x^{1/3} \log x),
\]

where \( \tau(n) \) denotes the number of divisors of \( n \) and \( \gamma \) is Euler’s constant.

While Theorem 1 is hence, in some sense, a consequence of Voronoi’s formula, our method extends to other problems of similar type. Lemmas 2 and 3 below are useful in many situations, and certainly deserve to be known outside number theory.

We say that a number \( x \in F_k \) admits an **FP-inverse square root** if there exists \( y \in F_k \) such that \((xy^2)_{(k)} = 1\). Such a \( y \) does not always exist: \( x = \frac{3}{2} \in F_3 \) has no FP-inverse square root in \( F_3 \)—the two numbers of \( F_3 \) around \( 1/\sqrt{x} \) are \( y_1 = \frac{3}{4} \) and \( y_2 = \frac{7}{8} \), and \((xy_1^2)_{(3)} = (\frac{27}{12})_{(3)} = \frac{7}{4} \) while \((xy_2^2)_{(3)} = \frac{7}{4} \). If \( x \) has an FP-inverse square root, then so do \( 4x \) and \( x/4 \), thus we may restrict the study to \( \frac{1}{2} \leq x < 2 \). We obtain the following result.

**Theorem 2.** Every number of \( F_k \) admits at most one FP-inverse square root.

**Theorem 3.** The number \( \delta^+(k) \) of elements \( x \in F_k \cap [1, 2[ \) admitting an FP-inverse square root satisfies

\[
\delta^+(k)/2^{k-1} = \frac{3\sqrt{2} - 3}{2} + O(2^{-k/3}) = 0.621320343 \ldots + O(2^{-k/3}).
\]
Theorem 4. The number $\delta^-(k)$ of elements $x \in F_k \cap [\frac{1}{2}, 1]$ admitting an FP-inverse square root satisfies

$$\delta^-(k)/2^{k-1} = \frac{3\sqrt{2} - 3}{2\sqrt{2}} + O(2^{-k/3}) = 0.4393398278 \ldots + O(2^{-k/3}).$$

2. FP-reciprocals—Proof of Theorem 1

The number $x = 1$ admits $y = 1$ as unique FP-reciprocal; $x \in F_k \cap ]1, 2]$ admits an FP-reciprocal if, and only if, $1 - 2^{-k-1} \leq xy \leq 1 + 2^{-k}$ for some $y \in F_k$. Writing $x = m/2^k$, $y = n/2^k$ with $2^{k-1} < m < 2^k$, $2^{k-1} < n < 2^k$ we see that the condition on $xy$ is equivalent to

$$2^{2k-1} - 2^{k-2} \leq mn \leq 2^{2k-1} + 2^{k-1}.$$  

With $y_1 := 2^{2k-1} - 2^{k-2} - 1$, $y_2 := 2^{2k-1} + 2^{k-1}$, since $]y_1/m, y_2/m\subset [2^{k-1}, 2^k]$, we see that the number of FP-reciprocals of $x$ equals the number of integers $n$ in the range $]y_1/m, y_2/m\$. This quantity is exactly $\lfloor y_2/m \rfloor - \lfloor y_1/m \rfloor$.

Put $M := y_2 - y_1 = 3 \cdot 2^{k-2} + 1$, $N := 2^{k-1}$. Then $y_2/m - y_1/m \geq 1$ if, and only if, $m \leq M$, so $x$ admits at least one FP-reciprocal when $m \leq M$ and has at most one when $m > M$. Therefore,

$$\gamma_0(k) = \sum_{M < m < 2N} \left(1 - \lfloor y_2/m \rfloor + \lfloor y_1/m \rfloor\right),$$

$$\gamma_2(k) = \sum_{N < m < M} \left(\lfloor y_2/m \rfloor - \lfloor y_1/m \rfloor - 1\right).$$

Introducing the first Bernoulli function $B_1(u) := u - \lfloor u \rfloor - \frac{1}{2}$, we obtain

$$\gamma_0(k) = 2N - M - 1 - \sum_{M < m < 2N} \frac{M}{m} + \sum_{M < m < 2N} B_1\left(\frac{y_2}{m}\right) - \sum_{M < m < 2N} B_1\left(\frac{y_1}{m}\right),$$

$$\gamma_2(k) = -M + N - 1 + \sum_{N < m < M} \frac{M}{m} - \sum_{N < m < M} B_1\left(\frac{y_2}{m}\right) + \sum_{N < m < M} B_1\left(\frac{y_1}{m}\right).$$

The sums involving $B_1$ will be handled by the following classical result.

Lemma 1. Let $f$ be a real valued, twice continuously differentiable function on an interval $I$ of length $|I| \geq 1$. Suppose that there exist $\lambda > 0$, $\alpha \geq 1$, such that

$$\lambda \leq |f''(x)| \leq \alpha \lambda \quad (x \in I).$$

Then we have

$$\sum_{n \in I} B_1(f(n)) = O(\alpha |I|^{1/3} + \lambda^{-1/2}).$$
In particular, we have, uniformly for all integers $N \geq 1$, intervals $I \subset [N, 2N]$ and real numbers $y \in [N^2, 3N^2]$,  
\[ \sum_{n \in I} B_1(y/n) = O(N^{2/3}). \] (3)

Applying (3) with $y = y_1$ and with $y = y_2$, we obtain  
\[ \gamma_0(k) = 2N - M - M \{ \log(2N/M) + O(1/M) \} + O(N^{2/3}), \]
\[ \gamma_2(k) = -M + N + M \{ \log(M/N) + O(1/N) \} + O(N^{2/3}). \]

The estimates of Theorem 1 for $\gamma_0(k)$ and $\gamma_2(k)$ follow on replacing $M$ and $N$ by their explicit values. The result on $\gamma_1(k)$ is then a consequence of the identity  
\[ \gamma_0(k) + \gamma_1(k) + \gamma_2(k) = 2^{k-1}. \]

Lemma 1 is closely connected with Voronoï’s asymptotic formula [12] on the divisor problem. Indeed, an elementary computation yields that  
\[ \sum_{n \leq x} \tau(n) - x \log x - (2^{1/2} - 1)x = 2 \sum_{n \leq \sqrt{x}} B_1(x/n) + O(1) \quad (x \geq 1). \] (4)

Thus, Lemma 1 implies Voronoï’s formula and, conversely, all known proofs of Voronoï’s theorem, including the original proof and Vinogradov’s elementary generalization [11] provide (3). Note that estimate (3) with the slightly weaker error term $O(N^{2/3} \log N)$ formally follows from (1) and (4) when $I = [1, N], \ y = N^2$. Thus, Theorem 1 may be seen as a consequence of Voronoï’s theorem.

Van der Corput’s method [1], now a classical tool in analytic number theory, yields a simple and short proof of Lemma 1. It will enable us to provide the reader with a self-contained proof of Lemma 1.

A comprehensive study on the statistical behaviour of fractional parts of $x/n$ and related sequences, also depending on van der Corput’s method, has been undertaken by Saffari and Vaughan [5,6,7].

2.1. Reduction to exponential sums

We write $e(u) = \exp(2i\pi u)$. The first Bernoulli function $B_1$ can be sharply approximated by trigonometric polynomials using the following handy result.

**Lemma 2** (Vaaler [10]). For $H \in \mathbb{N}$, $h \in \mathbb{Z}$, $1 \leq |h| \leq H$, let  
\[ 0 < b_H(h) := \pi \frac{|h|}{H + 1} \left( 1 - \frac{|h|}{H + 1} \right) \cot \left( \pi \frac{|h|}{H + 1} \right) + \frac{|h|}{H + 1} < 1. \]

Then, the trigonometric polynomial  
\[ B_H^*(x) = -\frac{1}{2i\pi} \sum_{1 \leq |h| \leq H} b_H(h) e(hx) \]
satisfies for \( x \in \mathbb{R} \)

\[
|B_1(x) - B^*_H(x)| \leq \frac{1}{2H + 2} \sum_{|h| \leq H} \left( 1 - \frac{|h|}{H + 1} \right) e(hx) = \frac{\sin^2 \pi(H + 1)x}{2(H + 1)^2 \sin^2 \pi x}.
\]

**Proof.** For \( x \notin \mathbb{Z} \) this is inequality (7.14) of Vaaler [10]—see also [2], Theorem A.6. For \( x \in \mathbb{Z} \), both sides are equal to \( \frac{1}{2} \), so the result remains true. \( \square \)

We now prove Lemma 1. We note at the outset that we may assume \( \lambda \leq 1 \) since the result is otherwise trivial. Applying Lemma 2 to the left-hand side of (2), we get

\[
\left| \sum_{n \in I} B_1(f(n)) \right| \leq \frac{|I|}{2H + 2} + \sum_{1 \leq |h| \leq H} \left( \frac{1}{2\pi|h|} + \frac{1}{2H + 2} \right) \left| \sum_{n \in I} e(hf(n)) \right|
\]

\[
\leq \frac{|I|}{2H} + \sum_{1 \leq |h| \leq H} \frac{2}{h} \left| \sum_{n \in I} e(hf(n)) \right|.
\]

We note that this upper bound, with possibly other numerical constants, could also be formally deduced from the classical Erdős–Turán inequality (see, e.g. [3], p. 112 and 114).

The exponential sum on the right-hand side above may be handled by the following basic result in van der Corput’s theory.

**Lemma 3** (van der Corput [1]). *Under the assumptions of Lemma 1, we have

\[
\sum_{n \in I} e(f(n)) = O(\alpha |I| \lambda^{1/2} + \lambda^{-1/2}).
\]

**Proof.** See, e.g., Theorem 5.9 of [9], Theorem 2.2 of [2], or Théorème I.4.5, p. 96 of [8]. \( \square \)

We derive from Lemma 3 that

\[
\sum_{1 \leq h \leq H} \frac{1}{h} \left| \sum_{n \in I} e(hf(n)) \right| = \sum_{1 \leq h \leq H} \frac{1}{h} O(\alpha |I| \sqrt{\lambda h} + \sqrt{1/\lambda h})
\]

\[
= O(\alpha |I| \sqrt{\lambda H} + \sqrt{1/\lambda})
\]

and so

\[
\left| \sum_{n \in I} B_1(f(n)) \right| = O \left( \frac{|I|}{H} + \alpha |I| \sqrt{\lambda H} + \sqrt{1/\lambda} \right).
\]

Selecting \( H = \lceil \lambda^{-1/3} \rceil \), we obtain the required estimate.
3. FP-inverse square roots

As noticed above, we may assume without loss of generality that \( \frac{1}{2} \leq x < 2 \). We have \( (xy^2)_k = 1 \) if, and only if, \( 1 - 2^{-k-1} \leq xy^2 \leq 1 + 2^{-k} \).

The number \( x = 1 \) admits \( y = 1 \) as unique FP-inverse square root. Uniqueness follows from the inequalities

\[
1 - 2^{-k-1} < \sqrt{1 - 2^{-k-1}} \leq y \leq \sqrt{1 + 2^{-k}} < 1 + 2^{-k}.
\]

Conversely, \( x = 1 \) is the only number in \( F_k \) having \( y = 1 \) as FP-inverse square root.

Let \( x \in F_k \cap [1, 2[ \) and assume that \( x \) admits \( y \) as an FP-inverse square root. We write \( x = \frac{m}{2^{k-1}} \) and \( y = \frac{n}{2^k} \) with \( 2^{k-1} \leq m, \ n < 2^k \). We have

\[
y_0 \leq mn^2 \leq y_2
\]

with \( N := 2^{k-1} \), \( y_0 := 4N^3 - N^2 \), \( y_2 := 4N^3 + 2N^2 \). Arguing as before, we obtain that the number of FP-inverse square roots of \( x \) is the number of integers in the interval

\[
\left[ \sqrt{y_0/m}, \sqrt{y_2/m} \right]. \quad (5)
\]

Similarly, for \( x \in F_k \cap [\frac{1}{2}, 1[ \) we write \( x = \frac{m}{2^k} \) with \( 2^{k-1} \leq m < 2^k \) and conclude that the number of FP-inverse square roots of \( x \) is the number of integers in the interval

\[
\left[ \sqrt{y_0/(2m)}, \sqrt{y_2/(2m)} \right]. \quad (6)
\]

**Proof of Theorem 2.** It suffices to show that both intervals above have length < 1 and hence contain each at most one integer. The length of interval (6) is smaller than that of interval (5) by a factor \( 1/\sqrt{2} \). The length of interval (5) does not exceed

\[
\frac{y_2 - y_0}{2\sqrt{my_0}} \leq \frac{3N^2}{2\sqrt{3mN^3}} = \frac{\sqrt{3}}{2} \frac{\sqrt{N}}{m} < 1
\]

since \( m \geq N \). \( \Box \)

**Proof of Theorem 3.** Retaining the above notation, and putting \( y_1 := 4N^3 - N^2 - 1 \), we may assert that \( \lfloor \sqrt{y_2/m} \rfloor - \lfloor \sqrt{y_1/m} \rfloor \) is 1 or 0 according to whether \( m/2^{k-1} \) has or not an FP-inverse square root. Thus

\[
\delta^+(k) = \sum_{N \leq m < 2N} \left( \lfloor \sqrt{y_2/m} \rfloor - \lfloor \sqrt{y_1/m} \rfloor \right)
\]

\[
= \sum_{N \leq m < 2N} \left( \sqrt{y_2/m} - \sqrt{y_1/m} \right) - \sum_{N \leq m < 2N} \{ B_1(\sqrt{y_2/m}) - B_1(\sqrt{y_1/m}) \}.
\]

Let us call the first of the above two sums the ***main term***, and the second one the ***remainder term***. We now turn to estimates for those two sums.
By the mean value theorem, we see that the main term is
\[
\sum_{N \leq m < 2N} \frac{1}{\sqrt{m}} \left\{ \frac{3N^{1/2}}{4} + O \left( \frac{1}{\sqrt{N}} \right) \right\} = \frac{3\sqrt{2}}{2} - \frac{3}{2} N + O(1).
\] (7)

To estimate the remainder term, we apply Lemma 1 with \( f(x) := \sqrt{y/x} \). We obtain that, uniformly for all integers \( N \geq 1 \), all intervals \( I \subset [N, 2N] \) and all real numbers \( y \in [N^3, 6N^3] \), we have
\[
\sum_{n \in I} B_1(\sqrt{y/n}) = O(N^{2/3}).
\] (8)

This is plainly sufficient. \( \square \)

**Proof of Theorem 4.** We argue exactly as in the proof of Theorem 3, except that we deal with interval (6) instead of interval (5), and so the main term is divided by \( \sqrt{2} \). \( \square \)

**Acknowledgements**

The authors take pleasure in thanking J.-M. Muller for his interest in, and some suggestions on, the present work.

**References**