

Introduction

All of us are of course familiar with the decimal number system. Those of us with a bit more grounding in mathematics might have experience with the binary, octal and hexadecimal number bases. The purpose of this paper is to explore the use of number systems with *nonintegral* bases. In particular, I approach the case where the base is a negative improper fraction.

For the most part, I will be arguing that there are practical applications for the number system in a negative improper function base. This paper will show that this number system retains many of the features of a conventional number system, for instance a practical addition algorithm. In fact, I will show that there are significant advantages of using an alternative base system.

In this first revision, I remove the fallacious Theorem 3, and added a few of my later insights, including some treatment of the case of positive rational bases. I've also corrected a few minor errors, and toned down a bit of the cockiness of this introduction.

The nature of this study is rather expository in nature, as I fear that an exhaustive analysis is as yet beyond my capabilities. Having said that, I am rather pleased with my first foray into mathematical research and hope that I have succeeded in making a contribution however small to our understanding of number theory.

DARREN ONG CHUNG LEE
25 August 2005

Definitions and Terminology

1. The *base*, b is a nonzero rational number
2. The *height*, h is an integer strictly larger than 1.
3. The *array* $[p_d, p_{d-1}, \dots, p_0]_b$ where all the *digits* p_i are nonnegative integers is defined as

$$\sum_0^d p_j b^j$$

For example:

$$[1, 9, 8, 7]_{10} = 1987$$

$$[1, 1, 1, 0]_2 = 14 \text{ or } 1110_2$$

$$[2, 13]_{10} = 33$$

$$[1, 2]_{\frac{5}{2}} = \frac{9}{2}$$

$$[1, 3, 1]_{-2} = -1$$

If in the context the base is clearly given, then $[p_d, p_{d-1}, \dots, p_0]_b$ may be simply written as $[p_d, p_{d-1}, \dots, p_0]$. If d is the largest integer where p_d is not zero then the array is known as a $d+1$ -digit array

4. The *field* ${}^h F_b$ of the height h and the base b is the set of possible values of $[p_d, p_{d-1}, \dots, p_0]_b$ subject to the restrictions:
 - all the digits p_i must strictly less than h
 - there are only finitely many digits in the array

In other words, ${}^h F_b$ is the set of arrays in base b using only the digits drawn from $(0, 1, 2, 3, \dots, h-1)$

5. An element in ${}^h F_b$ has *cardinality* c if it can be represented in exactly c distinct ways by $[p_n, p_{n-1}, \dots, p_0]_b$ with respect to the restrictions in definition 4. For example, the number 4 has cardinality 3 in ${}^5 F_2$ because it can be represented as $[1, 0, 0]_2$, $[2, 0]_2$ and $[4]_2$.
6. A rational number is *n-friendly* if its denominator in lowest terms only contains prime factors that are also prime factors of the integer n . Integers are n -friendly for all n . Rationals that are not n -friendly are *n-unfriendly*.

THEOREM 1: If $\frac{k}{g}$ is a rational number in lowest terms and $|k|$ is larger than $|g|$ then all the elements in ${}^{|k|}F_{\frac{k}{g}}$ have cardinality 1.

PROOF: If, conversely there exists a number expressible in two or more ways in ${}^{|k|}F_{\frac{k}{g}}$ then we must have for $0 \leq p_i, q_i \leq |k| - 1$

$$[p_d, p_{d-1}, \dots, p_0]_{\frac{k}{g}} = [q_d, q_{d-1}, \dots, q_0]_{\frac{k}{g}} \quad (1)$$

with at least one of the pairs p_i, q_i being unequal.

By Definition 3 we may express this equation as a polynomial, that is

$$(p_d - q_d)\left(\frac{k}{g}\right)^d + (p_{d-1} - q_{d-1})\left(\frac{k}{g}\right)^{d-1} + \dots + (p_0 - q_0) = 0 \quad (2)$$

By the rational root theorem, we know k divides $(p_0 - q_0)$. But the absolute value of $(p_0 - q_0)$ is strictly less than $|k|$ since both p_0 and q_0 are nonnegative integers strictly less than $|k|$. Thus k divides a number with an absolute value less than $|k|$. This is a contradiction. Hence Theorem 1 holds.

QED

THEOREM 2: If m and n are relatively prime positive integers with m strictly larger than n , then ${}^mF_{(-\frac{m}{n})}$ is the set of n -friendly rationals.

PROOF: It is obvious that ${}^mF_{(-\frac{m}{n})}$ contains only rational numbers. We still need to show that it doesn't contain rationals that are n -unfriendly.

Claim 2.1: ${}^mF_{(-\frac{m}{n})}$ cannot contain any n -unfriendly rationals

PROOF: $[p_d, p_{d-1}, \dots, p_0]_{(-\frac{m}{n})}$ when expressed as a rational in lowest terms cannot have a prime factor in its denominator not also in n . This is because by Definition 3

$$[p_d, p_{d-1}, \dots, p_0]_{(-\frac{m}{n})} = \sum_0^d p_j \left(-\frac{m}{n}\right)^j = \frac{\sum_0^d p_j (-m)^j n^{d-j}}{n^d} \quad (3)$$

In equation (3), both the numerator and denominator are integers. Reducing this to lowest terms will not add any prime factors to the denominator. Hence the claim holds.

★

Now we need to show that all n -friendly rationals are elements of ${}^mF_{\frac{m}{n}}$

Claim 2.2: $0, 1, -1, m$ and $\frac{1}{n}$ are all elements of ${}^mF_{(-\frac{m}{n})}$.

PROOF: We can verify immediately using Definition 3 that for base $(-\frac{m}{n})$

1. $[0] = 0$
2. $[1] = 1$
3. $[n, (m-1)] = -1$
4. $[n, (m-n), 0] = m$

Thus $0, 1, -1$ and m are elements of ${}^mF_{(-\frac{m}{n})}$

We then demonstrate that $\frac{1}{n}$ can be written as a two digit array. We must show $p_1(-\frac{m}{n}) + p_0 = \frac{1}{n}$ has solutions for $0 \leq p_1, p_0 \leq (m-1)$

We know that $nx - 1 \equiv 0 \pmod{m}$ has a solution for x between 1 and $(m-1)$ inclusive since m and n are relatively prime. So we can set p_0 to equal that value of x . So now $np_0 - 1 = my$ for some positive integer y . But this integer y cannot exceed $m-1$ otherwise $np_0 - 1 < np_0 < ny < my$. Thus y lies between 0 and $(m-1)$ inclusive and we may set $p_1 = y$. This gives us

$$np_0 = 1 + mp_1 \tag{4}$$

$$\left(-\frac{m}{n}\right)p_1 + p_0 = \frac{1}{n} \tag{5}$$

So we have found p_0, p_1 that satisfy $[p_1, p_0] = \frac{1}{n}$

Thus $\frac{1}{n}$ is an element of ${}^mF_{(-\frac{m}{n})}$ and the claim is proven.

★

Claim 2.3: For any base b , the $d+1$ digit array $[p_d, p_{d-1}, \dots, p_0]_b$ multiplied by b (or $[1, 0]_b$) yields the $d+2$ digit array $[p_d, p_{d-1}, \dots, p_0, 0]_b$

PROOF: This follows immediately from Definition 3.

★

Claim 2.4: The sum of any two elements of ${}^mF_{(-\frac{m}{n})}$ is an element of ${}^mF_{(-\frac{m}{n})}$

PROOF: We use an algorithmic approach. That is, we will create an algorithm to determine the correct array of a sum of two elements of ${}^mF_{(-\frac{m}{n})}$. First note that

$$[p_d, p_{d-1}, \dots, p_0]_b + [q_d, q_{d-1}, \dots, q_0]_b = [(p_d + q_d), (p_{d-1} + q_{d-1}), \dots, (p_0 + q_0)]_b \tag{6}$$

holds if none of the $(p_i + q_i)$ terms equal or exceed the height.

If however the sum of a pair of corresponding digits in the two arrays equals or exceeds the height, then we have to 'carry over' as we do in conventional, decimal addition.

The algorithm for the sum of two elements in ${}^m F_{(-\frac{m}{n})}$ proceeds as follows:

1. We arrange the digits as we do in conventional decimal addition: in three rows with the two summands occupying the top two rows and the answer row in the bottom. Corresponding digits are in the same columns. We leave an extra row on top to accommodate the digits being 'carried over'. We call the top row the *carry-over row* the bottom row the *answer row* and the two rows containing the arrays to be summed the *summand rows*.

	p_3	p_2	p_1	p_0
+	q_3	q_2	q_1	q_0

2. We add p_0 and q_0 . If the sum exceeds $(m-1)$ then we subtract m repeatedly until we get a number between 0 and $(m-1)$ inclusive. This number will be a_0 , the units digit of the answer. For every time we subtracted m we 'carry over' $(m-n)$ to the second column and n to the third column. We do this because $m = [n, (m-n), 0]$ as determined in Claim 2.2. We then erase p_0 and q_0 since they have been accounted for.

		n	$(m-n)$	
	p_3	p_2	p_1	
+	q_3	q_2	q_1	
				a_0

m was subtracted once in this step

3. We then repeat the process in the second column. This time we have to add the digits carried over to the second column (if any) to the sum of p_1 and q_1 . After subtracting m as many times as necessary from this sum, we determine a_1 analogously. For every time m is subtracted we carry $(m-n)$ to the next column and we carry n to the column after that one. We then erase all the summed terms in the second column, including the digits in the carry-over row.

	$3n$	$n+3(m-n)$		
	p_3	p_2		
+	q_3	q_2		
			a_1	a_0

m was subtracted thrice in this step

4. We apply the algorithm to the following columns until the summand and carry-over rows are all empty or we reach a point where we find the algorithm repeats while only adding zeroes to the answer row. Then, to get the answer simply read the digits in the answer row as an array. In other words the answer is $[\dots, a_3, a_2, a_1, a_0]$

There are actually two operations involved in this algorithm:

- The addition operation, where we erase two digits p_i, q_i in the summand rows and replace them with a new digit in the answer row. This is when the sum of the two digits does not exceed $m-1$.
- The carry over operation, where we erase two digits p_i, q_i in the summand rows, subtract m from the sum as necessary and place the remainder in the answer row, while carrying over $(m-n)$ to the next column and n to the column after that as many times as m was subtracted from the first sum.

Firstly note that the digital sum of the entire arrangement, that is the sum of all the unerased digits in the answer, summand and carrying rows is invariant. The first operation merely erases two digits in the summand rows and puts their sum in the answer row. The second does the same thing, except it first removes m a few times from the initial sum. Each m subtracted is accounted for, since the sum of the n term and the $(m-n)$ term carried over is still m . Thus neither operation affects the digital sum and so the digital sum of the entire arrangement is constant.

Secondly, if we modify the arrangement by multiplying all the digits in the units columns by $(-\frac{m}{n})^0$, all the digits in the second column by $(-\frac{m}{n})^1$, all the digits in the third column by $(-\frac{m}{n})^2$ etc counting all erased digits and empty spaces as 0, we find that the sum of all the modified digits is invariant as well. The first operation removes two numbers and adds their sum; the three numbers concerned are in the same column and thus when modified are multiplied by the same term. Hence the first operation does not affect the modified digital sum. The second does the same thing, except it first removes m a few times from the initial sum. Each m is accounted for, since by carrying over we are in essence replacing each m subtracted with $[n, (m-n), 0]$ and these two terms are equivalent by Claim2.2. Thus the two operations do not affect the modified digital sum either.

We are now ready to prove our algorithm works. We claim that at the end of the algorithm the modified digital sum of the carry-over and summand rows are all zero. Thus the array we get in the answer row is equivalent to the modified digital sum at the start of the algorithm. Note that the modified digital sum at the start of the algorithm is in fact the sum of the two summand arrays (refer to Definition 3 to see why) .

It may be the case that at some point all the digits in the top three columns become erased- in which case we are done. However sometimes the carry-over row is never empty no matter how many times we repeat the algorithm. For example, try adding $[2, 2]$ and $[1]$ in base $-\frac{3}{2}$, keeping in mind that the height is 3 and that $3 = [2, 1, 0]$ is the carry-over value. The algorithm goes on forever without terminating. Note that after a certain point of a non-terminating algorithm the answer row can only generate zeroes. If this isn't true, then the digital sum of the answer row must indefinitely increase. This is impossible

since the digital sum of the answer row cannot exceed the digital sum of the four rows, which is a constant.

Since both the summed arrays have finitely many digits, after a certain point of the algorithm both summand rows must always be empty as well. Thus, after a certain point in this nonterminating algorithm only the carry-over row is nonempty. Now consider a point in the algorithm where only the carry-over row remains nonempty, **before** we apply the algorithm to a column. Let c_0 be the value of the carry-over row in the *active* column (the column for which we are about to operate the algorithm) c_1 the value of the carry-over row in the column after the active one. Note that only these two values in the carry-over row are nonzero, since the carry over operation only carries over to the two next columns ($m-n$ to the next column and n to the column after that) These two values are therefore the only nonempty values left in the arrangement so $c_0 + c_1$ equals the constant digital sum. There are only finitely many possible values for c_1, c_0 so if we repeat the algorithm sufficiently many times, we will have two steps in the algorithm where for both steps the value c_0 and hence the value of c_1 is the same. Say this happens before the ϕ th and $(\phi+k)$ th steps. The modified digital sum for the summand and answer rows are the same for both points in the algorithm, since we have reached the point where both summand rows are empty and the answer row is generating zeroes. Thus the modified digital sum of the carry-over row for the ϕ th and $(\phi+k)$ th steps is the same as well. We get

$$c_1\left(-\frac{m}{n}\right)^\phi + c_0\left(-\frac{m}{n}\right)^{\phi-1} = c_1\left(-\frac{m}{n}\right)^{\phi+k} + c_0\left(-\frac{m}{n}\right)^{\phi+k-1} \quad (7)$$

$$\left(\left(-\frac{m}{n}\right)^k - 1\right)\left(c_1\left(-\frac{m}{n}\right) + c_0\right) = 0 \quad (8)$$

this implies $(c_1\left(-\frac{m}{n}\right) + c_0) = 0$ which means $(c_1\left(-\frac{m}{n}\right)^{\phi+k} + c_0\left(-\frac{m}{n}\right)^{\phi+k-1}) = 0$

Hence the modified digital sum in the carry-over row for the $(\phi+k)$ th step is zero. So when the 'pattern' of the carry-over row repeats after the summand rows are empty we are done: the modified digital sum for the top three rows is zero and so the array in the answer row will give us the value we want.

Thus for any two finite arrays in ${}^mF_{\left(-\frac{m}{n}\right)}$ we can find another array in ${}^mF_{\left(-\frac{m}{n}\right)}$ equal to their sum, as claimed.

★

Claim 2.5: *All integers are elements of ${}^m F_{(-\frac{m}{n})}$*

PROOF: By Claim 2.2 both 1 and -1 are elements of ${}^m F_{(-\frac{m}{n})}$. By applying Claim 2.4 repeatedly we find that all other integers are elements of ${}^m F_{(-\frac{m}{n})}$ as well.

★

Claim 2.6: *If X is an element of ${}^m F_{(-\frac{m}{n})}$ then so is $\frac{X}{n}$*

PROOF: say $X = [p_d, p_{d-1}, \dots, p_0]$. Then by Definition 3 and Claim 2.3

$$X = \sum_0^d p_j \left(-\frac{m}{n}\right)^j = p_0 \cdot [1] + p_1 \cdot [1, 0] + p_2 \cdot [1, 0, 0] + \dots + p_d \cdot [1, 0, 0, \dots, 0] \quad (9)$$

We note that $\frac{1}{n}$ is a two digit array in ${}^m F_{(-\frac{m}{n})}$ by Claim 2.2. Let $\frac{1}{n} = [v_1, v_0]$. If in equation (9) we replace $[1]$ by $[v_1, v_0]$, $[1, 0]$ by $[v_1, v_0, 0]$, $[1, 0, 0]$ by $[v_1, v_0, 0, 0]$ etc the value of all the terms are divided by n and so the value of the entire expression becomes $\frac{X}{n}$. Using the addition algorithm in Claim 2.4 we can thus obtain a single array fulfilling the conditions of ${}^m F_{(-\frac{m}{n})}$ (note that multiplication of a term by λ is equivalent to adding that term to itself λ times). The value of this array is $\frac{X}{n}$ thus proving that $\frac{X}{n}$ is indeed an element of ${}^m F_{(-\frac{m}{n})}$.

★

All n -friendly rationals can be expressed in the form $\frac{Z}{n^k}$ where Z is an integer and k a natural number. By Claim 2.5 we know that Z is an element of ${}^m F_{(-\frac{m}{n})}$. Applying Claim 2.6 to Z k times we find that $\frac{Z}{n^k}$ is also an element of ${}^m F_{(-\frac{m}{n})}$

QED

THEOREM 3: *If m and n are relatively prime positive integers with m larger than n , then ${}^mF_{(\frac{m}{n})}$ is the set of all rational numbers of the form $\sum_{i=0}^k a_i \frac{m^i n^{k-i}}{n^k}$ where k and the a_j are positive integers.*

PROOF: We derive a property similar to one used to prove Theorem 2.

Claim 3.1: *The sum of any two elements of ${}^mF_{(\frac{m}{n})}$ is itself an element of ${}^mF_{(\frac{m}{n})}$.*

PROOF: Note that for base $\frac{m}{n}$, $[n, 0] = m$. Using this equation in lieu of Claim 2.2, we can prove this assertion the same way Claim 2.4 was proven. This is left as an exercise for the reader.

★

Since $\sum_{i=0}^k a_i \frac{m^i n^{k-i}}{n^k}$ is simply a sum of numbers $a_0, a_1 \frac{m}{n}, a_2 \frac{m^2}{n^2} \dots a_k \frac{m^k}{n^k}$ we are done.

QED

The next line of investigation is, obviously to determine what rational numbers can be expressed in the form $\sum_{i=0}^k a_i \frac{m^i n^{k-i}}{n^k}$. This leads us immediately to this general formulation: given constant integers $\eta_1, \eta_2 \dots \eta_k$, which have a gcd 1, to determine the what integers can be expressed as a sum of the η_j . (Each η_j may be added more than once). To facilitate our investigation, we define $H(\eta_1, \eta_2 \dots \eta_k)$ as the set of integers that may be expressed as a sum of η_j , where each η_j may be in the sum more than once.

Firstly, we wish to investigate the largest number not in the set $H(\eta_1, \eta_2 \dots \eta_k)$. A proof of the general problem is unfortunately beyond me. I have however made progress in a particular case that is sufficiently general to illuminate our understanding of ${}^mF_{(\frac{m}{n})}$.

THEOREM 4: *If m , n and k are positive integers with m and n relatively prime then the largest integer not in $H(m^k, m^{k-1}n, mk - 2n^2 \dots n^k)$ is $(n - 1)\left(\frac{n^{k+1} - m^{k+1}}{n - m} - \frac{n^{k+1} - 1}{n - 1}\right) - 1$*

PROOF: We prove this assertion through induction on k . Claim 4.1 will serve as our basis case.

Claim 4.1: *Where m and n are relatively prime integers, the largest integer not in $H(m, n)$ is $(m-1)(n-1)-1$*

PROOF: Note first of all that

$$mx + ny = (m - 1)(n - 1) - 1$$

is unsolvable in non-negative integers x, y . This is equivalent to

$$m(x + 1) + n(y + 1) = mn$$

since m and n are relatively prime, we must have $y+1=m$ and $x+1=n$. But then we get $2mn=mn$, a contradiction.

WLOG $m > n$. We then need to show that all the integers from $(m-1)(n-1)$ to $m(n-1)$ are elements of $H(m, n)$. Consider the modular equation

$$mx \equiv c \pmod{n}$$

where c is positive. We know that this equation has solutions for all values c between 0 and $n-1$ inclusive. Thus

$$ny - mx = -c$$

has solutions for all those values of c as well. Note that $x < n$. So we know the equations

$$ny - mx = -1$$

$$ny - mx = -2$$

...

$$ny - mx = -(n - 1)$$

are all solvable. Adding these equations to $m(n-1)$ will generate a sum of the form $mx' + ny'$ for $m(n-1)-1, m(n-2)-2, \dots, (m-1)(n-1)$ respectively. Thus all these numbers, along with $m(n-1)$ are elements of $H(m, n)$. Since these are n consecutive integers, we know that all larger integers are elements of $H(m, n)$ since we can generate those by adding n to one of the integers from $(m-1)(n-1)$ to $m(n-1)$.

★

Assume that the statement of Theorem 3 is true for a value $k=v-1$. Consider $H(m^v, m^{v-1}n \dots n^v)$. This is equivalent to $H(m^v, n \times H(m^{v-1}, m^{v-2}n \dots n^{v-1}))$. In other words, the numbers that can be represented for case $k=v$ are those representable in the form

$$m^v x + n(m^{v-1}y_0 + m^{v-2}ny_1 \dots + n^v y_v) \quad (10)$$

In equation [10], we may set $x < n$ since if $x \geq n$, we can subtract nm^v as many times as we want from $m^v x$ by adding in a corresponding number of times to y_0 . By the induction hypotheses, we know that the largest multiple of n that cannot be represented by [10] is $n[(n-1)(\frac{n^k-m^k}{n-m} - \frac{n^k-1}{n-1}) - 1]$. Generally, where $\alpha < n$ the largest number $\alpha m^k \pmod n$ that cannot be similarly represented is $\alpha m^k + n[(n-1)(\frac{n^k-m^k}{n-m} - \frac{n^k-1}{n-1}) - 1]$. Thus the largest number that cannot be represented by [10] is the case we get when $\alpha = n-1$, that is

$$(n-1)m^k + n[(n-1)(\frac{n^k-m^k}{n-m} - \frac{n^k-1}{n-1}) - 1]$$

We complete our induction proof by simplifying this expression, obtaining

$$(n-1)(\frac{n^{k+1}-m^{kt1}}{n-m} - \frac{n^{kt1}-1}{n-1}) - 1$$

QED