THE CONVERSION OF HENSEL CODES TO RATIONAL NUMBERS

T. M. Rao
Department of Computer Science
University of Lagos
Lagos Nigeria

R. T. Gregory
Department of Computer Science
University of Tennessee
Knoxville, TN 37916

ABSTRACT

In a finite-segment p-adic number system one of the difficult problems is concerned with converting Hensel codes back into rational numbers. An algorithm for this conversion is proposed which is based on a sophisticated table look-up procedure.

1. Introduction

The use of finite-segment p-adic arithmetic as a practical method for performing error-free computation on a digital computer was first proposed by Krishnamurthy, Rao, and Subramanian [15], [16] and by Alparslan [1]. Since that time other researchers have become interested in the subject. See, for example, Beiser [3], Farinmade [5], Gregory [6], [7], Hehner and Horspool [9], [10], and Lewis [17].

Krishnamurthy et al. use the term Hensel code* to describe the first r digits of the infinite p-adic expansion of a rational number $\alpha = a/b$. They demonstrate that arithmetic operations on rational numbers can be replaced by corresponding arithmetic operations on their Hensel codes under certain conditions. One of the difficulties, however (once a computation has been completed), is in finding a simple method for determining the rational equivalent of the Hensel code which represents the solution. It is this difficulty which we address in this paper. The notation used is similar to that in Gregory [6], [7].

2. Finite-segment p-adic numbers

Modern introductions to Hensel's field of p-adic numbers Q_p (of which the rational numbers Q form a subfield) are contained in Bachman [2], Borevich and Shafarevich [4], Koblitz [12], and Mahler [18]. It is not necessary, however, to have a complete understanding of the theoretical aspects of p-adic numbers in order to work with finite-segment p-adic numbers (i.e., Hensel codes) because there is a simple algorithm for computing Hensel codes which does not require that the infinite p-adic expansions be known. In other words, it is possible to introduce Hensel codes without ever mentioning p-adic numbers. However, we do not choose to do this and a brief review of the connection between Hensel codes and p-adic numbers follows.

Let p be a prime and suppose we are given the rational number α with its unique, periodic, p-adic expansion

$$(2.1)\quad \alpha = a_n p^n + a_{n+1}p^{n+1} + a_{n+2}p^{n+2} + \dots$$
$$= p^n(c_0 + c_1 p + c_2 p^2 + \dots),$$

where $0 \le a_j < p$, for $j = n, n+1, \dots$, and $a_n \neq 0$.

Notice that

$$(2.2)\quad c_i = a_{n+i}, \quad i = 0, 1, 2, \dots$$

and n may be positive, negative, or zero.

Because of the one-to-one correspondence between (2.1) and the abbreviated representation

$$(2.3)\quad \alpha = a_n\, a_{n+1}\, a_{n+2} \cdots \quad (p),$$

where only the coefficients of the powers of p are exhibited, we can use the power series expansion and the abbreviated representation interchangeably. In fact, we shall refer to both (2.1) and (2.3) as the p-adic expansion for α.

The abbreviated notation is analogous to the representation of the decimal expansion of α. In fact, we complete the analogy by introducing a p-adic point as a device for displaying the sign of n. Thus, we write

$$(2.4)\quad \alpha = \begin{cases} a_n a_{n+1} \cdots a_{-1} \,.\, a_0 a_1 a_2 a_3 \cdots & \text{for } n < 0 \\ .\, a_0 a_1 a_2 a_3 \cdots & \text{for } n = 0 \\ .0 \dots 0\, a_n \cdots & \text{for } n > 0. \end{cases}$$

Since any rational number $\alpha = a/b$ has a unique representation

$$(2.5)\quad a/b = (c/d)p^n,$$

with $(c, d) = (c, p) = (d, p) = 1$, it follows, from (2.1), that the p-adic expansion for c/d is

$$(2.6)\quad c/d = .c_0 c_1 c_2 \cdots \quad (p).$$

Hensel codes

If we wish to work with r p-adic digits, the

*Named for K. Hensel who first proposed the system of p-adic numbers during the first decade of the twentieth century.

fixed-point Hensel code consists of the leftmost r digits in (2.4), including the leading zeros when $n > 0$, with the p-adic point in its proper place. However, we shall work primarily with normalized floating-point Hensel codes, in which case the p-adic point will always be as in (2.6), and the Hensel code will be written as an ordered pair consisting of a mantissa m_α and an exponent e_α.

2.7 Example From [7, p.111], we see that the p-adic expansion for $\alpha = 2/3$, with $p = 5$, is

$$2/3 = .4131313\ldots \quad (5).$$

It is easily verified (using radix-5 arithmetic from left to right) that $3(.41313\ldots) = .2000\ldots$ and this is the p-adic representation of 2. Similarly, the expansions for 2/15 and 10/3 are

$$2/15 = 4.131313\ldots \quad (5)$$

and

$$10/3 = .04131313\ldots \quad (5)$$

respectively. Hence, if $r = 4$, we have the fixed-point Hensel codes

$$\begin{aligned} H(5,4,2/3) &= .4131 \\ H(5,4,2/15) &= 4.131 \\ H(5,4,10/3) &= .0413, \end{aligned}$$

and the normalized floating-point Hensel codes

$$\begin{aligned} \hat{H}(5,4,2/3) &= (.4131,0) \\ \hat{H}(5,4,2/15) &= (.4131,-1) \\ \hat{H}(5,4,10/3) &= (.4131,1). \end{aligned}$$

Notice that $\hat{H}(5,4,10/3)$ contains one more "significant digit" than $H(5,4,10/3)$, and we can avoid many difficulties if we use normalized floating-point Hensel codes rather than fixed-point Hensel codes. See [6, pp. 292-296].

It is shown in [15] that negative rational numbers have a valid radix compliment representation. Thus, from Example 2.7, we can write

$$\begin{aligned} \hat{H}(5,4,-2/3) &= (.1313,0) \\ \hat{H}(5,4,-2/15) &= (.1313,-1) \\ \hat{H}(5,4,-10/3) &= (.1313,1). \end{aligned} \tag{2.8}$$

If r is even, positive and negative integers have easily recognizable Hensel codes in the sense that the last r/2 digits are zero if the integer is positive, and p-1 if the integer is negative. Thus, for example.

$$\begin{aligned} \hat{H}(5,4,13) &= (.3200,0) \\ \hat{H}(5,4,-13) &= (.2244,0) \end{aligned} \tag{2.9}$$

3. Residue equivalent of the Hensel code

We now describe an algorithm for mapping a rational number α onto its Hensel code $H(p,r,\alpha)$ which does not involve finding the (infinite) p-adic expansion and truncating it to r digits. It is based on the following result. See [7, p. 122].

3.1 Theorem Let $\alpha = a/b$, where $a/b = (c/d)p^n$, with $(c, d) = (c, p) = (d, p) = 1$. Let the Hensel code for c/d be

$$H(p,r,c/d) = .c_0c_1\cdots c_{r-1}.$$

Then $c_{r-1}\cdots c_1c_0$ is the radix-p representation for the integer $|cd^{-1}|_{p^r}$.

In other words,

$$|cd^{-1}|_{p^r} = c_0 + c_1p + \ldots + c_{r-1}p^{r-1}.$$

3.2 Example Let $\alpha = 2/3$, $p = 5$, and $r = 4$. Then $p^r = 625$ and

$$\begin{aligned} |2/3|_{625} &= |2\cdot 3^{-1}|_{625} \\ &= |2\cdot 417|_{625} \\ &= 209. \end{aligned}$$

Since $209_{ten} = 1314_{five}$, we reverse the order of the digits and obtain

$$H(5,4,2/3) = .4131$$

and this agrees with the result in Example 2.7.

If $\alpha = -2/3$ we have

$$\begin{aligned} |-2/3|_{625} &= |(-2)\cdot 417|_{625} \\ &= 416. \end{aligned}$$

Since $416_{ten} = 3131_{five}$, we reverse the order of the digits and obtain

$$H(5,4,-2/3) = .1313$$

which agrees with (2.8).

3.3 Remark Notice that when $\alpha = (c/d)p^n$, as in Theorem 3.1, with $\hat{H}(p,r,\alpha) = (m_\alpha, e_\alpha)$, we have

$$\begin{aligned} m_\alpha &= H(p,r,c/d) \\ &= .c_0c_1\cdots c_{r-1} \end{aligned}$$

and

$$e_\alpha = n.$$

Rational numbers with the same Hensel code.

It is easy to show that a/b and ka/kb (where $k \neq 0$ is any integer) have the same Hensel code. We also have the following result. See [7, p. 132].

3.4 Theorem Let $\alpha = a/b$ and $\beta = g/h$ with $(b, p) = (h, p) = 1$. Then $H(p,r,\alpha) = H(p,r,\beta)$ if and only if $ah \equiv bg \pmod{p^r}$.

Since we are interested primarily in normalized floating-point Hensel codes the following corollary (which follows directly from the theorem) is of interest.

3.5 Corollary Let $\alpha = (c/d)p^s$ and $\beta = (e/f)p^t$ as in (2.5). Then the mantissas of their normalized floating-point Hensel codes are equal (i.e. $m_\alpha = m_\beta$) if and only if $cf \equiv de \pmod{p^r}$.

To obtain a better understanding of what we are doing when we construct a normalized

floating-point Hensel code, consider the rational number $\alpha = (c/d)p^n$ as in (2.5). When we form the integer $|c\cdot d^{-1}|_{p^r}$ we are using the mapping $|\cdot|_{p^r} : Q \to I_{p^r}$ to map the rational number c/d onto the integer $|c\cdot d^{-1}|_{p^r}$ in the set

$$I_{p^r} = \{0,1,\ldots,p^r-1\}. \tag{3.6}$$

Following this step (assuming that the integers in (3.6) are represented in decimal notation) we change the representation of the integer $|c\cdot d^{-1}|_{p^r}$ from radix-10 to radix-p and, from Theorem 3.1, the radix-p digits are equal to the first r coefficients in the p-adic expansion (2.1) but with the order reversed.

The order of the digits in a Hensel code is not critical, however. For example, Knuth [11, p. 179] places the p-adic point on the right, rather than the left, so that (2.6) could be written

$$c/d = \ldots c_2c_1c_0. \quad (p), \tag{3.7}$$

and we could choose to write

$$H(p,r,c/d) = c_{p-1}\cdots c_1c_0. \tag{3.8}$$

However, we shall go along with the current literature and write the digits in reverse order.

Uniqueness

Theorem 3.4 and Corollary 3.5 clearly imply that we have infinitely many rational numbers mapped onto each Hensel code. Hence, unless we can establish a one-to-one mapping, there is no hope of mapping a Hensel code onto a unique rational equivalent. (This is the difficulty mentioned in the introduction which is addressed in this paper.)

We can establish a one-to-one mapping if we restrict the set of rational numbers to a finite subset of Q, called the order-N Farey fractions,

$$F_N = \{\alpha = a/b:\ 0 \le |a| \le N \text{ and } 0 < b \le N\}, \tag{3.9}$$

where N is a positive integer (defined, in our case, in Theorem 3.11). Let us define the set $\hat{H}_{p,r}$ to be the set of Hensel codes $\hat{H}(p,r,\alpha)$ for the order-N Farey fractions, that is,

$$\hat{H}_{p,r} = \{\hat{H}(p,r,\alpha) : \alpha \in F_N\}. \tag{3.10}$$

Then the following theorem is fundamental.

3.11 Theorem Let p be a prime and let r be a positive integer. Define N to be the largest positive integer which satisfies the inequality

$$N \le [(p^r-1)/2]^{\frac{1}{2}}.$$

Then the mapping $f : F_N \to \hat{H}_{p,r}$ is one-to-one and onto and thus has an inverse $f^{-1} : \hat{H}_{p,r} \to F_N$.

Proof See Rao [19]. ∎

Under the restriction of Theorem 3.11 to rational numbers which are order-N Farey fractions, we now examine the problem of finding the rational equivalent of a normalized floating-point Hensel code.

3.12 Remark From this point on the term "Hensel code" will always refer to a normalized floating-point Hensel code unless otherwise specified.

4. A simple table look-up procedure

Several methods for mapping Hensel codes onto their rational equivalents have been proposed in the literature; for example, the successive addition method, the multiplication method and the common denominator method [15], [19]. In this section we propose a method which makes use of a simple table look-up procedure. (A more sophisticated table look-up procedure is developed in subsequent sections.)

Assume that $\alpha \in F_N$, where

$$\begin{aligned} \alpha &= a/b \\ &= (c/d)p^n, \end{aligned} \tag{4.1}$$

with (c,d) = (c,p) = (d,p) = 1, and that

$$\hat{H}(p,r,\alpha) = (m_\alpha, e_\alpha) \tag{4.2}$$

is known. Recall that $e_\alpha = n$ and

$$m_\alpha = H(p,r,c/d). \tag{4.3}$$

Thus, obtaining a/b, given m_α and e_α, is simply a matter of obtaining c and d.

To obtain c and d, given m_α, we examine a table of fixed-point Hensel codes constructed as follows. Let

$$D = \{i_1, i_2, \ldots, i_k\} \tag{4.4}$$

be a set of integers satisfying the conditions

$$1 \le i_1 \le i_2 \le \ldots \le i_k \le N \tag{4.5}$$

and

$$(i_j, p) = 1, \quad j = 1,2,\ldots,k. \tag{4.6}$$

The k x k array M contains elements M_{st}, where

$$M_{st} = H(p,r,i_s/i_t). \tag{4.7}$$

Given $\hat{H}(p,r,\alpha) = (m_\alpha, e_\alpha)$, where α satisfies (4.1), we find c and d by comparing m_α successively with the elements of the array M in the order $M_{11}, M_{12}, M_{21}, M_{13}, M_{22}, M_{31}, \ldots, M_{kk}$ until we find a match. When a match is found, i_s and i_t are c and d.

It is easy to compute the size of the table M. The number of integers in the interval [1,N], which are not multiples of p, is

(4.8) $$k = N - \lfloor N/p \rfloor$$

and M contains k^2 elements.

Because of the size of this table (the number of elements is of order N^2) and the fact that the elements in the table are not ordered (making a sequential search necessary), this table look-up procedure is very inefficient and not recommended in practice. Consequently, in subsequent sections we develop a more sophisticated table look-up procedure.

5. The weight of α

In Theorem 3.11 we introduced the one-to-one and onto mapping $f : F_N \to \hat{H}_{p,r}$, where, for $\alpha \in F_N$,

(5.1) $$f(\alpha) = \hat{H}(p,r,\alpha).$$

We now introduce the mapping $g : \hat{H}_{p,r} \to I_{p^r}$, where, for $\hat{H}(p,r,\alpha) = (.c_0c_1\ldots c_{r-1}, e_\alpha)$,

(5.2) $$g(\hat{H}(p,r,\alpha)) = \begin{cases} c_0+c_1p+\ldots+c_{r-1}p^{r-1} & n < 0 \\ \left|p^n(c_0+c_1p+\ldots+c_{r-1}p^{r-1})\right|_{p^r} & n \geq 0, \end{cases}$$

with $e_\alpha = n$.

The composition of these two mappings, $w = gf$, is the mapping $w : F_N \to I_{p^r}$, where, for $\alpha \in F_N$,

(5.3) $$\begin{aligned} w(\alpha) &= g(f(\alpha)) \\ &= g(\hat{H}(p,r,\alpha)). \end{aligned}$$

Thus, we can map the order-N Farey fractions onto integers in the set I_{p^r}, where it is easy to prove that w has the following properties.

5.4 Theorem Let $\alpha, \beta \in F_N$. If $\alpha + \beta \in F_N$ and $\alpha\beta \in F_N$, then

(i) $w(\alpha+\beta) = \left|w(\alpha) + w(\beta)\right|_{p^r}$

(ii) $w(\alpha\beta) = \left|w(\alpha)\, w(\beta)\right|_{p^r}$

and

(iii) $w(-\alpha) = p^r - w(\alpha)$.

5.5 Definition The integer $w(\alpha)$ is called the weight of α.

From Theorem 3.1, if $c/d \in F_N$ with $(c,p) = (d,p) = 1$, then

(5.6) $$w(c/d) = \left|c\cdot d^{-1}\right|_{p^r}.$$

Also, from (5.2) and (5.3), it is clear that

(5.7) $$w(c/d) = w(c/b)$$

for any denominator of the form $b = d\cdot p^n$, with $0 < b \leq N$.

We introduce the set

(5.8) $$\tilde{F}_N = \{a/b \in F_N : (a,b) = 1 \text{ and } (b,p) = 1\}$$

which has the property that, if $\alpha,\beta \in \tilde{F}_N$ and $\alpha \neq \beta$, then $w(\alpha) \neq w(\beta)$.

The relation between $w(\alpha)$, for $\alpha \in F_N$, and its residue representation is shown in the following

5.9 Theorem If $\alpha = a/b = (c/d)p^n \in F_N$, then

$$w(\alpha) = \begin{cases} \left|a\cdot b^{-1}\right|_{p^r} & n \geq 0 \\ \left|a\cdot d^{-1}\right|_{p^r} & n < 0. \end{cases}$$

Proof (i) Let $n = 0$. In this case $a = c$ and $b = d$. Thus,

$$\begin{aligned} w(\alpha) &= w(c/d) \\ &= \left|c\cdot d^{-1}\right|_{p^r} \\ &= \left|a\cdot b^{-1}\right|_{p^r}. \end{aligned}$$

(ii) Let $n > 0$. In this case $a = c\cdot p^n$ and $b = d$. Thus,

$$\begin{aligned} w(\alpha) &= \left|p^n(c_0+c_1p+\ldots+c_{r-1}p^{r-1})\right|_{p^r} \\ &= \left|p^n \left|c_0+c_1p+\ldots+c_{r-1}p^{r-1}\right|_{p^r}\right|_{p^r} \\ &= \left|p^n\left|c\cdot d^{-1}\right|_{p^r}\right|_{p^r} \\ &= \left|a\cdot b^{-1}\right|_{p^r}. \end{aligned}$$

(iii) Let $n < 0$. In this case $a = c$ and $b = d\cdot p^{-n}$. Thus,

$$\begin{aligned} w(a/b) &= w(a/d) \\ &= \left|a\cdot d^{-1}\right|_{p^r}. \end{aligned}$$ ■

Observe that the mapping $w : \tilde{F}_N \to I_{p^r}$ is one-to-one but it is not onto. However, if we denote the set of images of $\tilde{F}_N$ by

(5.10) $$\hat{I}_{p^r} = \{w(\alpha) : \alpha \in \tilde{F}_N\}$$

then $w : \tilde{F}_N \to \hat{I}_{p^r}$ is both one-to-one and onto. (See Appendices A and B.)

Maximal sets of Farey fractions

We now show that $\tilde{F}_N$ has some interesting properties and, because of the bijection just stated, these properties apply to $\hat{I}_{p^r}$ as well.

5.11 Theorem If $\alpha = a/b \in \tilde{F}_N$, then at least one of the rational numbers $\alpha + 1$ or $\alpha - 1$ is also in $\tilde{F}_N$.

Proof

$$(a/b) \pm 1 = (a \pm b)/b$$

and $|a|, |b| \leq N$ imply

$$|\min(a+b, a-b)| \leq N. \quad \blacksquare$$

5.12 Corollary If $t \in \hat{I}_{p^r}$, then at least one of the integers $t + 1$ or $t - 1$ is also in $\hat{I}_{p^r}$.

Consequently, although the set of integers $\hat{I}_{p^r}$ is not a set of consecutive integers (i.e., it has some gaps in it) the integers in $\hat{I}_{p^r}$ appear in groups of at least two consecutive integers (and not as singles). We now examine how large these groups can be.

5.13 Definition Let $a/b \in \tilde{F}_N$. Then the maximal set of order-N Farey fractions generated by a/b is the set

$$F_{a,b} = \left\{\frac{a-k_2b}{b}, \ldots, \frac{a-b}{b}, \frac{a}{b}, \frac{a+b}{b}, \ldots, \frac{a+k_1b}{b}\right\}$$

where k_1 and k_2 are the largest integers satisfying

(i) $$a + k_1b \leq N$$

and

(ii) $$|a - k_2b| \leq N.$$

The fraction a/b is called a generator for $\tilde{F}_{a,b}$. However, there are other generators and it is easy to prove the following

5.14 Theorem Any element of $\tilde{F}_{a,b}$ can be used as a generator for $\tilde{F}_{a,b}$.

Hence, the following is needed.

5.15 Definition If $a + jb$ is the least positive numerator among the elements of $\tilde{F}_{a,b}$, then $(a+jb)/b$ is called the prime generator for $\tilde{F}_{a,b}$. Using this definition we have the obvious

5.16 Corollary $\tilde{F}_{a+jb,b} = \tilde{F}_{a,b}$.

Incidentally, we can have other maximal sets with the same denominator b. For example, the set with prime generator x/b would be written $\tilde{F}_{x,b}$. Unless otherwise stated, x/y is assumed to be the prime generator for $\tilde{F}_{x,y}$.

5.17 Example Let $p = 5$ and $r = 4$. Then $N = 17$ and for $\alpha = 1/3$ we have

$$\tilde{F}_{1,3} = \left\{-\frac{17}{3}, -\frac{14}{3}, -\frac{11}{3}, -\frac{8}{3}, -\frac{5}{3}, -\frac{2}{3}, \frac{1}{3}, \frac{4}{3}, \frac{7}{3}, \frac{10}{3}, \frac{13}{3}, \frac{16}{3}\right\}$$

We now show that $\tilde{F}_N$ is a disjoint union of these maximal sets. Let D be the set of integers in (4.4) satisfying (4.5) and (4.6). For each $i_j \in D$ define the set of integers

(5.18) $$C_j = \left\{c_1^j, c_2^j, \ldots, c_m^j\right\},$$

where

(5.19) $$1 \leq c_1^j \leq c_2^j \leq \ldots \leq c_m^j \leq i_j,$$

and

(5.20) $$(c_s^j, i_j) = 1, \qquad s = 1,2,\ldots,m.$$

Clearly, $m = \phi(j)$, where $\phi(j)$ is Euler's function representing the number of integers less than j which are relatively prime to j.

5.21 Theorem The set $\tilde{F}_N$ can be expressed as

$$\tilde{F}_N = \bigcup_{j=1}^{k} \bigcup_{m=1}^{\phi(j)} \tilde{F}_{x,y}$$

where $x = c_m^j$ and $y = i_j$, and where

$$\tilde{F}_{x_1,y_1} \cap \tilde{F}_{x_2,y_2} = \emptyset$$

for all $x_1/y_1 \neq x_2/y_2$.

Proof

If $a/b \in F_N$, then $b = i_j$ for some j. Obviously, integers q and s exist for which $a = bq + s$. Clearly $s < b \leq N$ and $(s,b) = 1$ because if v divides s and b, then v divides a which implies v divides both a and b (a contradiction). Thus, $s = c_n^j$ for some n which implies $a/b \in \tilde{F}_{s,b}$. On the other hand, if $a/b \in \tilde{F}_{z,b}$ then clearly $a/b \in F_N$. It remains to be shown that maximal sets are disjoint. Suppose

$$\tilde{F}_{x_1,y_1} \cap \tilde{F}_{x_2,y_2} \neq \emptyset .$$

Choose a/b in the intersection and generate the maximal set $\tilde{F}_{a,b}$. From (5.14) we deduce that both

$$\tilde{F}_{a,b} = \tilde{F}_{x_1,y_1}$$

and

$$\tilde{F}_{a,b} = \tilde{F}_{x_2,y_2},$$

and these imply that

$$\tilde{F}_{x_1,y_1} = \tilde{F}_{x_2,y_2}$$

(a contradiction). $\blacksquare$

5.22 Corollary Let $2 \leq b \leq N$ and $(b,p) = 1$. Then the number sets $\tilde{F}_{z,b}$ in $\tilde{F}_N$ is exactly $\phi(b)$.

Maximal sets of integers

Let $\tilde{F}_{a,b}$ be a maximal set of Farey fractions and let $G_{a,b}$ be the set

(5.23) $$G_{a,b} = \{t-k_2, \ldots, t-1, t, t+1, \ldots, t+k_1\},$$