Capítulo II

Identidad en Precisión Variable

1. Representación de los números racionales
2. Instrumentación de la función identidad

II - 1

Representación de los números racionales

En este apartado se introduce un sistema de representación numérica para un computador capaz de representar valores racionales sin error. Dejando a un lado la notación numérica simbólica, nos referiremos en lo sucesivo a la representación fraccionaria posicional de los números como una forma de expresión directa de su valor.

Conceptos sobre representación numérica

Como paso previo a la definición del modelo de representación que se propone se establecen los principios de la expresión fraccionaria de los números racionales. A continuación se formalizan las relaciones existentes entre los números que intervienen en el problema y los datos que se manejan para su tratamiento, entendiendo como tales, los valores numéricos representados y manipulados por la máquina. En este aspecto se analizan los formatos de representación posicional de números racionales en coma fija y en coma flotante como punto de partida en la formalización del esquema de representación numérica propuesto.

Representación fraccionaria de los números racionales

Los números reales forman parte integrante de muchos problemas ya sea por los tipos de datos que manejan o por la naturaleza de las operaciones que intervienen. La manipulación y el almacenamiento de estos valores son tareas imprescindibles para el procesamiento de dichos problemas mediante un computador.

La representación fraccionaria posicional permite conocer directamente el valor del número, realizar comparaciones y operar de forma directa. Estas características son deseables en un formato de codificación numérica, sin embargo, no todos los números reales se pueden expresar de esta forma. Los números irracionales disponen de una cantidad infinita de cifras fraccionarias significativas y, por tanto, no es posible su expresión en notación posicional. Además, sólo un subconjunto numerable de números irracionales es computable, es decir, existe para ellos un procedimiento que es capaz de generar cada cifra del valor real [Gianantonio, 1993], [Turing, 1937]. Para estos números, la forma de expresar su valor exacto es procedimiental, donde se describe el método de obtención del valor en lugar del valor en sí, por ejemplo: π, $\sqrt{3}$, sen 2, $\lg_3 4$, e^5, ... Cabe recalcar que, si bien es posible obtener cualquier cantidad de sus cifras significativas en un tiempo finito, siempre será una aproximación respecto a su valor exacto. Por otra parte, es necesario considerar también las imprecisiones derivadas de las limitaciones físicas de la máquina para contener dichas cifras.

Los números racionales sí permiten su representación posicional exacta. Cualquier número racional expresado en notación fraccionaria puede representarse mediante una *cantidad finita de cifras significativas*.

La representación posicional de un número racional es la siguiente:

$$x \in \mathbb{Q} \Leftrightarrow \exists\, L(m_f), L(m_p)>0 \in \mathbb{N} \;/$$

$$/\; x = \alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\ \alpha_i\ ,\ \alpha_{i-1}\ \ldots\ \alpha_1\ \alpha_0\gamma_{L(mp)-1}\ \ldots\ \gamma_1\gamma_0\ \gamma_{L(mp)-1}\ \ldots\ \gamma_1\gamma_0\ \ldots \qquad [2.1]$$

donde para una base de representación B: $\alpha_i \in \{0..B-1\} \wedge \gamma_i \in \{0..B-1\}$

Esta representación está formada por la concatenación de una secuencia de cifras alrededor de la coma fraccionaria, llamada parte fija, con una serie periódica de cifras que forman un ciclo. A la secuencia de dígitos de menor longitud que se repite se denomina el periodo del número. En la notación que se emplea estas cifras se representan tan sólo una vez y se señalan trazando sobre ellas un arco de circunferencia.

$$x = \alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\,,\ \alpha_{i-1}\ldots\ \alpha_1\alpha_0\ \overparen{\gamma_{L(mp)-1}\ \ldots\ \gamma_1\gamma_0} \qquad [2.2]$$

La expresión anterior muestra el valor exacto de un número racional con una cantidad de cifras finita. Existe una equivalencia entre la representación fraccionaria posicional y la notación simbólica de fracción para expresar los números racionales.

$$\forall\ x \in \mathbb{Q},\ \exists\ a,\ b\ \in \mathbb{Z}\ /\ x = \frac{a}{b} \wedge b \neq 0 \qquad [2.3]$$

Aunque las expresiones [2.2] y [2.3] representan al mismo número de forma exacta, la notación simbólica de fracción no es única ya que existen infinitas fracciones para un mismo número.

Como casos particulares de la configuración del periodo en la representación fraccionaria, cabe mencionar los siguientes:

- *Todas las cifras son cero*: con la información correspondiente a las cifras no periódicas es suficiente para conocer con exactitud el valor del número.

  $$\alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\,,\ \alpha_{i-1}\ldots\ \alpha_1\alpha_0\,\overparen{\gamma_{L(mp)-1}\ \ldots\ \gamma_1\gamma_0} =$$
  $$= \alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\,,\ \alpha_{i-1}\ldots\ \alpha_1\alpha_0 \Leftrightarrow \forall\ i\ \in \{0..L(m_p)-1\}\ \gamma_i = 0 \qquad [2.4]$$

- *Todas las cifras son iguales a la base de la numeración menos uno* (en base 10 iguales a 9, en base 2 iguales a 1): en este caso el número se expresa sumando uno a la cifra fija menos significativa y eliminando las cifras periódicas.

  $$\alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\,,\ \alpha_{i-1}\ldots\ \alpha_1\alpha_0\,\overparen{\gamma_{L(mp)-1}\ \ldots\ \gamma_1\gamma_0} =$$
  $$=\alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\,,\ \alpha_{i-1}\ldots\ \alpha_1(\alpha_0+1) \Leftrightarrow \forall\ i\ \in \{0..L(m_p)-1\}\ \gamma_i =(B-1) \qquad [2.5]$$

La cantidad de cifras, $L(m_p)$, que forman el periodo de un número racional está en estrecha relación con el denominador de la fracción irreducible que lo representa según la siguiente expresión.

$$\forall\ x = \alpha_{L(mf)}\alpha_{L(mf)-1}\ldots\alpha_i\ ,\ \alpha_{i-1}\ldots\ \alpha_1\alpha_0\,\overset{\frown}{\gamma_{L(mp)-1}\ldots\ \gamma_1\gamma_0} \in \mathbb{Q}\ ,\ \exists\ a,b\ \in \mathbb{Z}\ /$$

$$/\ x = \frac{a}{b} \wedge b \neq 0 \wedge \mathrm{mcd}(a, b) = 1 \Rightarrow L(m_p) = \Im\ (b) < b \qquad [2.6]$$

La cantidad de cifras fraccionarias periódicas es siempre menor que b. En concreto, la definición de la función $\Im$ que determina $L(m_p)$ consiste en las siguientes relaciones de congruencia [Belski y Kaluzhnin, 1980], [Guelfond, 1979]:

Si $mcd(B, b) = 1 \Rightarrow B^{L(mp)} \equiv 1 \bmod (b)$

Si no

$$\exists\, b', m \in \mathbb{Z} \,/\, b = m \cdot b' \wedge mcd(B, b') = 1 \wedge m = \prod_{i=1}^{p} f_i^{\,j}$$

siendo $f_1, \ldots, f_p$ la lista de factores primos de la base de representación B y j un natural, de tal forma que:

$$\frac{a}{b} = \frac{1}{m} \cdot \frac{a}{b'} \wedge mcd(B, b') = 1 \wedge B^{L(mp)} \equiv 1 \bmod (b')$$

Por las reglas de la aritmética modular, $\frac{1}{m}$, al estar compuesta por factores primos de la base, genera un valor fraccionario de periodo cero.

donde,

$\equiv$: Relación de congruencia.

mod : Operador módulo.

Existen infinitos valores de que cumplen la ecuación anterior, ya que todo múltiplo del valor $L(m_p)$ satisface la ecuación:

$$B^{L(mp)} \equiv 1 \bmod (b) \Rightarrow \forall\, j \in \mathbb{N},\, B^{j \cdot L(mp)} \equiv 1 \bmod (b) \qquad [2.7]$$

Este hecho justifica que cualquier grupo de cifras que incluya al periodo una cantidad entera de de veces y cuya longitud sea múltiplo de $L(m_p)$ constituye en sí mismo un periodo. Asimismo, todos los números racionales son periódicos, si bien éste no se representa cuando todas sus cifras son cero.

La cantidad de números periódicos con periodo distinto de cero está relacionada con la base de la numeración. Existen números racionales decimales de periodo cero que producen un periodo binario distinto de cero, mientras que al ser 2 un factor de 10, todo número binario de periodo cero produce un número decimal también de periodo cero. Esta observación es de gran relevancia para la representación numérica y la aritmética de computadores, ya que en un dominio binario existen números fraccionarios de periodo cero en decimal que producirán errores en la representación binaria ante la imposibilidad de codificar infinitas cifras fraccionarias. De ahí la importancia de definir un modelo que permita representar y operar con estos números de manera exacta.

Sin pérdida de generalidad se considera a lo largo de este trabajo números racionales no negativos menores que uno. Es decir,

$$x \in [0,\, 1) \subset \mathbb{Q} \Leftrightarrow 0{,}\alpha_{L(mf)-1}\alpha_{L(mf)-2}\ldots\alpha_i\alpha_{i-1}\ldots\,\alpha_1\alpha_0\,\overparen{\gamma_{L(mp)-1}\,\ldots\,\gamma_1\gamma_0} \qquad [2.8]$$

y donde $\alpha_{L(mf)} = 0$.

A la secuencia de cifras $\alpha_{L(mf)-1}\ldots\alpha_1\alpha_0$ de longitud de $L(m_f)$ es a lo que se denomina *mantisa fija*:

$$m_f = \alpha_{L(mf)-1}\ldots\alpha_1\alpha_0 \qquad [2.9]$$

A la secuencia de cifras $\gamma_{L(mp)-1}\,\ldots\,\gamma_1\gamma_0$ de longitud $L(m_p)$ se denomina *mantisa periódica*:

$$m_p = \gamma_{L(mp)-1}\,\ldots\,\gamma_1\gamma_0 \qquad [2.10]$$

La mantisa periódica sólo tiene sentido para posiciones a la derecha de la coma fraccionaria. La cadena de dígitos que la forman es suficiente para conocer toda la secuencia infinita de cifras fraccionarias que contiene la representación posicional de los números racionales a partir de la mantisa fija.

$$\widehat{m}_p = \gamma_{L(mp)-1}\,\ldots\,\gamma_1\gamma_0\,\gamma_{L(mp)-1}\,\ldots\,\gamma_1\gamma_0 \qquad [2.11]$$

El número racional queda formado por la concatenación de las cifras de la mantisa fija con las de la mantisa periódica a la derecha de la coma fraccionaria. Para una mayor claridad en la argumentación, se utilizará en lo sucesivo la siguiente expresión para referirse a los números racionales:

$$x \in [0,\ 1) \subset \mathbb{Q},\ x = 0{,}m_f\widehat{m}_p \qquad [2.12]$$

En función de la naturaleza de su mantisa se realiza una clasificación de éstos números racionales basándose en las siguientes definiciones.

Definición 1

Se denomina *número periódico* a todo número racional que contenga mantisa periódica distinta de cero:

$A \in [0,\ 1) \subset \mathbb{Q},\ A = 0{,}m_{fA}\widehat{m}_{pA} \wedge m_{fA} = \alpha_{L(mfA)-1}\ldots\alpha_0 \wedge m_{pA} = \gamma_{L(mpA)-1}\ldots\gamma_0\ /$

$/\ L(m_{pA}) > 0 \wedge \exists\ i \in \{0..L(m_{pA}) - 1\},\ \gamma_i \neq 0$ [2.13]

Definición 2

Se denomina *número no periódico* a todo número racional que contenga mantisa periódica igual a cero:

$A \in [0,\ 1) \subset \mathbb{Q},\ A = 0{,}m_{fA}\widehat{m}_{pA} \wedge m_{fA} = \alpha_{L(mfA)-1}\ldots\alpha_0 \wedge m_{pA} = \gamma_{L(mpA)-1}\ldots\gamma_0\ /$

$/\ L(m_{pA}) > 0 \wedge \forall\ i \in \{0..L(m_{pA}) -1\},\ \gamma_i = 0$ [2.14]

Definición 3

Se denomina *número periódico puro* a todo número periódico que contenga sólo mantisa periódica distinta de cero:

$A \in [0,1) \subset \mathbb{Q},\ A =0{,}m_{fA}\widehat{m}_{pA} \wedge m_{fA} = \alpha_{L(mfA)-1}\ldots\alpha_0 \wedge m_{pA} = \gamma_{L(mpA)-1}\ldots\gamma_0\ /$

$/\ L(m_{pA}) > 0 \wedge L(m_{fA}) = 0 \wedge \exists\ i \in \{0..L(m_{pA}) - 1\},\ \gamma_i \neq 0$ [2.15]

Definición 4

Se denomina *número periódico mixto* a todo número racional que contenga mantisa periódica no nula y mantisa fija:

$A \in [0,1) \subset \mathbb{Q},\ A =0{,}m_{fA}\widehat{m}_{pA} \wedge m_{fA} = \alpha_{L(mfA)-1}\ldots\alpha_0 \wedge m_{pA} = \gamma_{L(mpA)-1}\ldots\gamma_0\ /$

$/\ L(m_{pA}) > 0 \wedge L(m_{fA}) > 0 \wedge \exists\ i \in \{0..L(m_{pA}) - 1\},\ \gamma_i \neq 0$ [2.16]

Como aportación en este trabajo se presenta la siguiente definición que complementa a las anteriores. Los números de este tipo serán utilizados en el desarrollo de la algoritmia de los operadores matemáticos que se describen en capítulos posteriores.

Definición 5

Se denomina *número periódico unidad de grado s* a todo número periódico puro cuya mantisa periódica esté formada por una secuencia de $s-1$ ceros seguidos de un uno en la posición menos significativa.

$$A \in [0,1) \subset \mathbb{Q},\ A = 0,m_{fA}\widehat{m}_{pA} \wedge m_{fA} = \alpha_{L(mfA)-1}\ldots\alpha_0 \wedge m_{pA} = \gamma_{s-1}\ldots\gamma_0 \,/$$

$$/\ L(m_{pA}) > 0 \wedge L(m_{fA}) = 0 \wedge \forall\ i \in \{1..s-1\}\ \gamma_i = 0 \wedge \gamma_0 = 1 \qquad [2.17]$$

Los métodos para transformar un número racional desde una notación simbólica de fracción, $\frac{a}{b}$, a una notación fraccionaria posicional y viceversa se indican a continuación:

- La conversión de un número racional formulado en notación simbólica de fracción a expresión fraccionaria se realiza mediante la división indicada expresamente en la fracción. En el cociente se obtiene la expresión fraccionaria posicional y los restos de esta división entre b son valores β acotados: $0 \leq \beta < b$.

- La conversión de número racional expresado en notación posicional a notación simbólica de fracción se realiza mediante la siguiente expresión,

$$0,m_f\,\widehat{m_p} = \frac{m_f\,m_p - m_f}{\underbrace{(B-1)\cdots(B-1)}_{L(m_p)}\ \underbrace{0\ \cdots\ 0}_{L(m_f)}} \qquad [2.18]$$

  donde el numerador de la fracción se constituye por la resta entre el entero formado por la concatenación de las cifras de la mantisa fija y periódica menos las cifras de la mantisa fija y, el denominador consiste en el entero formado por la concatenación de $L(m_p)$ cifras iguales a la base de la representación menos uno con $L(m_f)$ ceros en la parte menos significativa.

A modo de ejemplo se presentan las siguientes correspondencias entre esquemas de representación de los números racionales:

Base 10:

$$0{,}1208\widehat{367} = \frac{1208367 - 1208}{9990000} = \frac{1207159}{9990000}$$

donde, $L(m_f) = 4$ y $L(m_p) = 3$

Base 2:

$$0{,}101\widehat{10} = \frac{10110 - 101}{11000} = \frac{10001}{11000}$$

donde, $L(m_f) = 3$ y $L(m_p) = 2$

Representación en coma fija

La representación de un número en coma fija consiste en una secuencia de cifras de una longitud de palabra establecida, $L(n)$, correspondientes a su representación aritmética posicional desde un orden de magnitud inicial (m_i) hasta un orden de magnitud final (m_f), tal que $m_i \geq m_f$. En general, se añade una posición más para expresar el signo del número mediante un convenio discreto. De este modo: $L(n) = m_i - m_f + 2$.

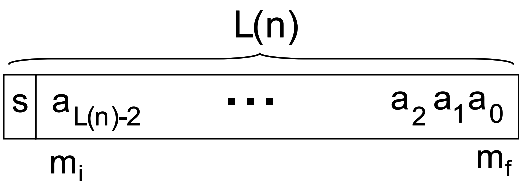

Figura 2-1: Esquema general de la representación en coma fija

El valor del número codificado en este formato se obtiene con el mismo procedimiento que en un sistema numérico posicional teniendo en cuenta únicamente los coeficientes asignados a tales potencias de la base.

$$A_{cfija} = (-1)^{s} \cdot \sum_{i=0}^{L(n)-2} a_i \cdot B^{mf+i} \qquad [2.19]$$

La denominación de coma fija proviene del hecho de que, del conjunto de cifras de la representación del número, se destina una de ellas a la codificación del signo, una parte fija de cifras, $L(m_e)$, a la representación de su parte entera y el resto, $L(m_f)$, a la representación de su parte fraccionaria. El orden de magnitud correspondiente a las unidades ocupa una posición intermedia, $m_i \geq 0 \geq m_f$, como describe la figura 2-2 [Cilio y Corporal, 1999], [Kum et al, 1997].

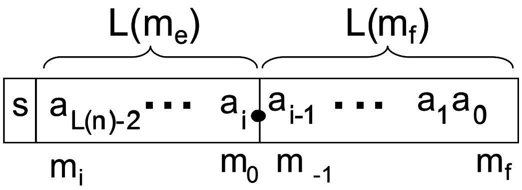

Figura 2-2: Estructura de la representación en coma fija

En lo que respecta a las características de representación numérica se destacan las siguientes:

1) El intervalo de los valores representables (R) viene determinado por la cantidad de cifras destinadas a la representación,

   $$R = [-(B^{L(me)} - B^{-L(mf)}),\ B^{L(me)} - B^{-L(mf)}] \qquad [2.20]$$

   donde B es la base de representación.

2) El intervalo de representación es discreto con una cuantización relacionada con la cantidad de cifras fraccionarias:

   $$q = B^{-L(mf)+1} \qquad [2.21]$$

Este esquema sólo es capaz de representar de manera exacta números racionales que contienen una cantidad de cifras enteras menor o igual a $L(m_e)$ y una cantidad de cifras fraccionarias menor o igual a $L(m_f)$. Esta limitación inherente a la coma fija supone un gran inconveniente para el cálculo matemático ya que existen números que no son representables sin error al no caber en alguna de las partes de que se compone el formato y, aunque se dedique una cantidad de cifras finita arbitraria a la representación numérica, tampoco es posible la codificación sin error de valores racionales con una parte fraccionaria periódica que se repita indefinidamente. Estas circunstancias se soslayan en la práctica realizando una aproximación a la representación exacta más cercana, lo que provoca imprecisiones y errores en los cálculos resultantes.

Por sus características, la representación en coma fija constituye una correspondencia entre el conjunto $\mathbb{Q}$ y los números racionales representados sin error por el formato. El redondeo al valor representable exacto provoca que una vez codificado un número resulte imposible conocer su valor original.

En la siguiente tabla se muestran algunos ejemplos de la representación en coma fija de valores racionales sin signo para una cantidad de cifras determinada: $L(n) = 10$; $L(m_e) = 5$; $L(m_f) = 5$. El conjunto de valores pretende ser lo más heterogéneo posible y debe interpretarse sólo como muestra de las características del formato.

$x \in \mathbb{Q}$	$A_{cfija}(x)$	|Error| $\lvert A_{cfija}(x) - x \rvert$
$123 \cdot 10^{1234567890}$	No representable	–
$123 \cdot 10^{123585}$	No representable	–
$123 \cdot 10^{3585}$	No representable	–
$123 \cdot 10^{7}$	No representable	–
123456	No representable	–
123,456	00123,45600	0
12345,67890123	12345,67890	0,00000123
0,123456	00000,12346	0,000004
0,00000000000012	00000,00000	0,000000000000123
$123 \cdot 10^{-7}$	00000,00001	$23 \cdot 10^{-7}$
$123 \cdot 10^{-3585}$	00000,00000	$123 \cdot 10^{-3585}$
123,45454545…	00123,45455	0,0000045454545…

0,017171717…	00000,017172	0,0000028282828…

Tabla 2-1: Ejemplos de representación en coma fija

Como se observa en la tabla anterior, los números con más de $L(m_e)$ cifras enteras se salen del rango de representación y no se pueden codificar por el formato. Para otros valores con más cifras que las disponibles en el esquema se realiza un redondeo que provoca un error de representación de una importancia proporcional al tamaño del número.

Representación en coma flotante

En este formato se expresan separadamente las cifras del número y su orden de magnitud. La longitud de palabra de la representación, L(n), se distribuye entre el signo, el exponente L(e) y la mantisa L(m).

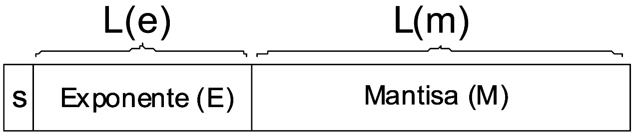

Figura 2-3: Esquema general de la representación en coma flotante

El valor del número se obtiene mediante la siguiente expresión,

$$A_{cflotante} = (-1)^{s} \cdot M \cdot B^{E} \qquad [2.22]$$

siendo B la base de la representación y, M y E el valor numérico de la mantisa y del exponente respectivamente.

La expresión separada del exponente permite la expresión de números de muy distinto valor. Esta característica proporciona una mayor flexibilidad y potencia expresiva a la vez que propicia su utilización en sistemas de propósito general [Patterson y Hennessy, 2002], [Sun Microsystems, 2000]. Su estandarización [IEEE, 1987], [IEEE, 1985] ha jugado un papel crucial en el desarrollo y extensión de este formato [Nielsen, 1997].

Se destacan las siguientes características:

1) La representación en coma flotante de un número no es única. Para evitar la múltiple codificación, la mantisa se expresa normalizada.

2) El intervalo de valores representables (R) depende de la cantidad de cifras que se destinan tanto a la mantisa como al exponente, alcanzando como se ha indicado, valores límite mucho mayores que en la anterior representación. Por ejemplo, el intervalo según la norma IEEE-754 tiene la siguiente expresión:

$$R = [-(2-2^{-L(m)})\cdot 2^{2^{L(e)-1}-1}, (2-2^{-L(m)})\cdot 2^{2^{L(e)-1}-1}] \qquad [2.23]$$

3) El intervalo de representación es discreto con una mayor separación entre los números representados conforme aumenta su orden de magnitud.

Al igual que en la representación en coma fija, tan solo se pueden expresar de manera exacta un conjunto finito de elementos. Todos los números expresados sin error son racionales con una cantidad finita de cifras y quedan fuera, además de los números irracionales, todo aquel número racional que no quepa en la longitud de palabra establecida. Para el resto de los valores que se encuentran en el intervalo de representación se realizan aproximaciones por algún número representable discreto. Por estos motivos, el formato establece una correspondencia entre el conjunto $\mathbb{Q}$ y los números racionales que representa de modo exacto y, debido al redondeo, no es posible recuperar el valor original desde la codificación realizada.

Se observa que en relación con el esquema anterior, se consigue cubrir un mayor espacio de la recta real aunque la cantidad de números representables de manera exacta es la misma que en coma fija [Sun Microsystems, 2000], [Kornerup y Matula, 1991], [Sterbenz, 1974], [Wilkinson, 1964].

En la tabla 2-2 se muestra un ejemplo de la representación con los mismos números y condiciones que en el formato de coma fija. En este caso se toma L(m) = 5, L(e) = 5. La mantisa se encuentra normalizada con el primer uno significativo a la izquierda de la coma fraccionaria.

$x \in \mathbb{Q}$	$A_{cflotante}(x)$		|Error|
	Exponente	Mantisa	$\|A_{cflotante}(x) - x\|$
$123 \cdot 10^{1234567890}$	No representable		–
$123 \cdot 10^{123585}$	No representable		–
$123 \cdot 10^{3585}$	03587	1,2300	0
$123 \cdot 10^{7}$	00009	1,2300	0
123456	00005	1,2346	4
123,456	00002	1,2346	0,004

12345,67890123	00004	1,2346	0,32109877
0,123456	-00001	1,2346	0,000004
0,00000000000012	-00013	1,2300	0
$123 \cdot 10^{-7}$	-00005	1,2300	0
$123 \cdot 10^{-3585}$	-03583	1,2300	0
123,45454545…	00002	1,2345	0,0045454545…
0,017171717…	-00002	1,7172	0,00000028282828

Tabla 2-2: Ejemplos de representación en coma flotante

Por su amplia utilización cabe realizar una mención especial a los errores que produce el procesamiento en coma flotante. Entre los innumerables ejemplos de falta de exactitud se describen los siguientes como representativos a pesar de lo sumamente sencillos.

Ejemplo1:

Sea la siguiente operación,

$$r = a - b - c$$

para los números: $a = 0,6$; $b = 0,35$; $c = 0,25$.

El cálculo con el formato en coma flotante de simple precisión de IEEE obtiene $r = 0,0000000298023223877$. El resultado exacto es $r = 0,0$.

Ejemplo 2:

Sean las siguientes matrices:

$$A = [-10^{18}, 2246, 10^{27}, 10^{25}, 22, 10^{5}]$$

$$B^{T} = [10^{38}, 33, 10^{29}, -10^{22}, 1044, 10^{42}]$$

El cálculo de su producto con el formato de representación de doble precisión de IEEE es $A \times B^{T} = 0$. El resultado exacto es $A \times B^{T} = 97,086$.

Se observa, por tanto, que mediante el sistema de representación de datos en coma flotante IEEE-754 se obtiene un resultado incorrecto en mayor o menor grado. El tamaño de este error puede no hacer viable su aplicación en determinados problemas.

Representación de números racionales en formato de doble mantisa

En los esquemas de representación de coma fija y coma flotante se ha comprobado que, a pesar de que su espacio de valores representable está contenido en $\mathbb{Q}$, existen infinitos números racionales no expresables aunque se permitan longitudes arbitrarias de $L(n)$. Es decir, por mucho que se aumente la cantidad de cifras representables y las longitudes de palabra es imposible la codificación exacta de tales valores.

Se plantea avanzar en la expresión numérica proponiendo un formato de representación que suministre una mayor potencia expresiva y sea capaz de cubrir un conjunto numérico más amplio. Esta nueva función de correspondencia abarca también la representación de los valores racionales que sistemáticamente quedan fuera de los anteriores esquemas, en concreto los valores racionales periódicos. El esquema de representación que se propone supone una extensión de los anteriores formatos al adoptar algunas de sus características. Se toma como base el método de Hehner y Horspool [Hehner y Horspool, 1979] y el esquema de coma flotante. De este último toma su flexibilidad y añade una segunda mantisa que representa explícitamente el periodo de los valores racionales.

El formato distribuye la cantidad de cifras significativas de la representación del número, $L(n)$, en cuatro partes: signo, exponente, $L(e)$, mantisa fija, $L(m_f)$, y mantisa periódica, $L(m_p)$. La mantisa fija constituye la parte fraccionaria del número racional no periódica, mientras que la mantisa periódica representa las cifras que forman la parte repetitiva. El exponente, al igual que en el formato de coma flotante, expresa el orden de magnitud del número.

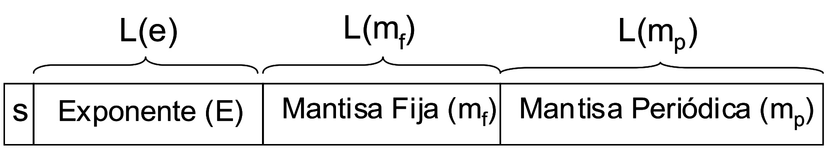

Figura 2-4: Esquema general del formato de representación de doble mantisa

La mantisa del número se confecciona por la concatenación de la mantisa fija y la mantisa periódica una cantidad indefinida de veces, tal como se muestra en la figura 2-5. De este modo se representan números con infinitas cifras fraccionarias a partir de una codificación con una cantidad de cifras finita.

Mantisa (M) =	m_f	m_p	m_p	m_p	• • •

Figura 2-5: Esquema de la formación de la mantisa

Para evitar la múltiple codificación de un número y conseguir una estructura compacta, el modelo contempla una serie de características adicionales que condicionan la representación:

1) La mantisa del número se normaliza situando su primer dígito con valor distinto de cero a la derecha de la coma fraccionaria. Si no existe mantisa fija se normaliza la mantisa periódica rotando sus cifras componentes hacia la izquierda.

   $$M = 0{,}m_f\widehat{m}_p \,/\, M \in [B^{-1}, 1) \qquad [2.24]$$

   donde B es la base de la representación.

2) La mantisa fija no debe contener secuencias de dígitos iguales a los de la mantisa periódica en su parte menos significativa.

   Sea $m_f = \alpha_{L(mf)-1}\alpha_{L(mf)-2}\ldots\alpha_1\alpha_0 \wedge m_p = \gamma_{L(mp)-1}\gamma_{L(mp)-2}\ldots\gamma_1\gamma_0$

   Entonces:

   $$\forall\, i \in [0..L(m_p)-1],\ \alpha_{i-1}\alpha_{i-2} \ldots \alpha_1\alpha_0 \neq \gamma_{i-1}\gamma_{i-2} \ldots \gamma_1\gamma_0 \qquad [2.25]$$

3) La mantisa periódica tiene que estar formada por la menor cantidad de dígitos periódicos.

   Sea $m_p = \gamma_{L(mp)-1}\gamma_{L(mp)-2}\ldots\gamma_1\gamma_0$

   Entonces: $\forall\, i \in [1..L(m_p)-1] \,/\, L(m_p) \bmod i = 0 \Rightarrow$

   $\Rightarrow \gamma_{i-1}\ldots\gamma_1\gamma_0 \neq \gamma_{2i-1}\ldots\gamma_1\gamma_i \ \wedge$

   $\wedge\ \gamma_{i-1}\ldots\gamma_1\gamma_0 \neq \gamma_{3i-1}\ldots\gamma_{2i+1}\gamma_{2i} \ \wedge$

   $\ldots$

11,001001000011111101101010100010001000010110100011000010001101001100010011000110011000101000101110000000011011100000011100110100010010100100000010010011100000100001
00010100110011111100110001110100

$$\wedge\ \gamma_{i-1}\ldots\gamma_1\gamma_0 \neq \gamma_{L(mp)-1}\ldots\gamma_{L(mp)-i+2}\gamma_{L(mp)-i+1} \qquad [2.26]$$

El valor del número se obtiene mediante una expresión similar a la del formato de coma flotante:

$$A_{\mathbb{Q}} = (-1)^{s} \cdot M \cdot B^{E} \qquad [2.27]$$

Siendo B la base de la representación, s el signo expresado en un convenio discreto y, M y E el valor numérico de la mantisa completa y del exponente respectivamente.

El valor de M se calcula con la expresión matemática descrita en [2.18]:

$$M = \frac{m_f\, m_p \;-\; m_f}{\underbrace{(B-1)\cdots(B-1)}_{L(m_p)}\ \underbrace{0\ \cdots\ 0}_{L(m_f)}} \qquad [2.28]$$

En caso de que no exista mantisa fija o la mantisa periódica sea igual a cero, su campo correspondiente en la codificación no contendrá ningún dígito, evitándose la representación de cifras innecesarias y no significativas. En estos casos el valor de la mantisa se calcula mediante las siguientes expresiones:

$$L(m_f) = 0 \Rightarrow M = \frac{m_p}{\underbrace{(B-1)\cdots(B-1)}_{L(m_p)}} \qquad [2.29]$$

$$m_p = 0 \Rightarrow M = \frac{m_f}{1\underbrace{00\cdots00}_{L(m_f)}} \qquad [2.30]$$

Mediante la utilización de una cantidad variable de cifras para cada campo en función de sus necesidades de codificación se aporta una mayor flexibilidad y potencia expresiva al modelo. Esta característica da la posibilidad de mantener campos vacíos si es necesario y de ajustar la mantisa periódica a su longitud.

Si no se consideran las limitaciones físicas del espacio material de representación de la máquina y se asumen longitudes arbitrarias de los campos del formato, el conjunto de números representables de manera exacta con este método abarca a todo el conjunto $\mathbb{Q}$. En esta situación, la función de representación propuesta corresponde con una aplicación biyectiva entre el conjunto de los números racionales.

Sea F la función de correspondencia que constituye el formato de representación:

$$F: \mathbb{Q} \rightarrow \mathbb{Q} \qquad [2.31]$$

Se resaltan sus características de interés siguientes:

- *Inyectiva*: cada valor racional tiene una codificación distinta. La cantidad de cifras fraccionarias de los números racionales es finita o bien es infinita periódica con periodo finito.

  $$\forall\ x_1, x_2 \in \mathbb{Q},\ F(x_1) \neq F(x_2) \qquad [2.32]$$

- *Sobreyectiva*: cualquier valor racional se puede representar con el formato propuesto.

  $$\forall\ x \in \mathbb{Q},\ \exists\ s, e, m_p, m_f \ /\ F(x) = x \qquad [2.33]$$

- *Existencia de inversa*: debido a la conjunción de las dos propiedades anteriores, la función de correspondencia dispone de inversa sobre el conjunto $\mathbb{Q}$, es decir, es posible construir una función que obtenga el valor racional inicial de cada codificación.

  $$\exists\ F^{-1}: \mathbb{Q} \rightarrow \mathbb{Q}\ /\ \forall\ x \in \mathbb{Q},\ x = F^{-1}(F(x)) \qquad [2.34]$$

Como consecuencia de estas propiedades todo número racional normalizado según el formato tendrá una expresión característica, compuesta por un signo, un exponente, una mantisa fija y una mantisa periódica.

Esta aplicación representa sin error los elementos de un conjunto numérico más amplio que el de las representaciones en coma fija o en coma flotante anteriores, las cuales son un caso particular del mismo. El nuevo formato constituye una función computable que mejora notablemente la capacidad de expresión de otros métodos que también muestran el valor numérico directo y evita los errores en la introducción de datos por el usuario a un computador que provoca el cambio de base de codificación desde decimal a binario. Por otra parte, admite procedimientos de cálculo que procesen números representados en este nuevo formato y que mantengan su capacidad de proveer un resultado exacto.

La tabla 2-3 muestra la representación con este formato de los ejemplos numéricos utilizados en los formatos anteriores. Se utilizan 10 dígitos repartidos entre el exponente, la mantisa fija y periódica.

$x \in \mathbb{Q}$	**Formato propuesto, F(x)**			**|Error|**
	Exponente	Mantisa fija	Mantisa periódica	**|F(x) − x|**
$123 \cdot 10^{1234567890}$	No representable			--
$123 \cdot 10^{123585}$	123588	0,123	--	0
$123 \cdot 10^{3585}$	03588	0,12300	--	0
$123 \cdot 10^{7}$	10	0,12300	--	0
123456	6	0,123456	--	0
123,456	3	0,123456	--	0
12345,67890123	5	0,12345678	--	0,0000012
0,123456	0	0,123456	--	0
0,0000000000001	-12	0,123	--	0
$123 \cdot 10^{-7}$	-4	0,123	--	0
$123 \cdot 10^{-3585}$	-03582	0,123	--	0
123,45454545…	3	0,123	45	0
0,017171717…	-1	--	17	0

Tabla 2-3: Ejemplos de representación en el formato propuesto

Como muestra la tabla anterior, siempre que la cantidad de cifras significativas del número sea menor o igual que los dígitos disponibles se permite su representación sin error. Sin embargo se observa que la instrumentación que se realice del formato va a estar sometida a las restricciones que imponga el espacio material de representación.

En el siguiente apartado se describe la instrumentación propuesta que relaciona los valores que intervienen en un problema con su codificación correspondiente en el marco de una arquitectura de procesador.

II - 2

Instrumentación de la función identidad

De acuerdo con la formulación del problema, la concepción de un procesador necesita de un método que proporcione una codificación de los valores numéricos con los que opera. Con este propósito se concibe el procedimiento de representación de los operandos en un formato adecuado para su procesamiento. La operatoria contempla operaciones de transferencia y almacenamiento flexible de los números.

Desde un enfoque matemático, se observa que un formato de representación corresponde con la función identidad,

$$f \equiv \text{identidad}: \mathbb{Q} \rightarrow \mathbb{Q} \quad [2.35]$$

tal que,

$$\forall\, x \in \mathbb{Q},\ \text{identidad}(x) = x \quad [2.36]$$

Formalmente, el procedimiento de representación numérica se considera como una instrumentación, $\Gamma_{\text{identidad}}$, de la función identidad cuyo objetivo es el de codificar en un computador elementos pertenecientes al conjunto de los números racionales.

La instrumentación según un diseño particular proporciona una codificación del número original, la cual, de acuerdo con la ecuación [1.2], cumple que:

$$\forall\, x \in \mathbb{Q},\ |\,\Gamma_{identidad}(x) - x| \leq \varepsilon \qquad [2.37]$$

Cualquier función de correspondencia entre valores racionales y su codificación en un computador (coma fija, coma flotante y el formato propuesto en el apartado anterior) encaja como instrumentación $\Gamma_{identidad}$, ya que realiza implementaciones concretas de métodos que representan valores racionales con un cierto grado de aproximación al valor exacto. Estos formatos constituyen instrumentaciones distintas de la función identidad.

La posibilidad de representación exacta de un formato está directamente relacionada con la existencia de inversa de la instrumentación en el conjunto numérico. Esta propiedad, reflejada en la expresión siguiente para el conjunto de los números racionales, establece un procedimiento de caracterización de formatos de representación sin error.

$$\forall\, x \in \mathbb{Q},\ \Gamma_{identidad}(x) = x \ \Leftrightarrow \forall\, x \in \mathbb{Q},\ \exists\ \Gamma^{-1}_{identidad} \qquad [2.38]$$

Las características del esquema de representación propuesto en el apartado anterior garantizan la representación exacta de cualquier valor racional con la única restricción de las limitaciones físicas de la máquina sobre la que se implemente y no por restricciones impuestas por el propio formato numérico. En consecuencia se escoge esta función de correspondencia para construir la *instrumentación* Γ *de la función identidad*.

El carácter de precisión variable de la función se adquiere mediante algún parámetro $\overline{d}$ adicional que regule el grado de aproximación de la función. Debido a las características del formato que se propone se asegura que para cualquier número racional existe una instancia de $\overline{d}$ con los que se alcanza su valor exacto.

11,00100100001111110110101010001000100001011010001100001000110100110001001100011001100010100010111000000001101110000011100110100010010100100000010010011100000100010001010011001111110011000111010

$$\forall\, x \in \mathbb{Q},\ \exists\ \bar{d}\ /\ \Gamma^{VP}_{identidad}(x,\ \bar{d}) = x \qquad [2.39]$$

El parámetro $\overline{d}$ admite distintas configuraciones. Una posible consiste en establecer $\overline{d}$ como un vector de números naturales que indican la cantidad de cifras que, como máximo, contienen los campos de la representación, es decir,

$$\overline{d} \equiv < d_1, d_2, d_3 >, \text{ con } d_i \in \mathbb{N} \qquad [2.40]$$

siendo, d_1, d_2 y d_3 la cantidad de cifras límite para la codificación del exponente, la mantisa fija y la mantisa periódica respectivamente.

Las condiciones sobre las longitudes de los campos vienen dadas por la siguiente expresión:

$$(L(e) \leq d_1) \wedge (L(m_f) \leq d_2) \wedge (L(m_p) \leq d_3) \qquad [2.41]$$

Otra configuración establece $\overline{d}$ como la cantidad máxima de cifras que se destinan a la representación del número en su conjunto sin especificar qué cantidad de dígitos se asigna a cada campo concreto. Esta opción aporta flexibilidad a la representación.

$$\overline{d} \equiv d \in \mathbb{N} \qquad [2.42]$$

La relación que ilustra esta condición se muestra en la ecuación siguiente:

$$1 + L(e) + L(m_f) + L(m_p) \leq d \qquad [2.43]$$

Configuraciones intermedias pueden establecer el tamaño de algunos campos del número y limitar el crecimiento conjunto de otros.

La interpretación de la restricción $\overline{d}$ en la instrumentación de la función corresponde con las limitaciones del espacio material de representación y constituye el tamaño máximo de la palabra o de la memoria disponible para la codificación del número o alguna de sus partes.

Como muestra la figura 1-1, el parámetro $\overline{d}$ es gestionado por la unidad de control de precisión para establecer las características de la arquitectura acordes con su valor. La instrumentación de esta gestión está estrechamente relacionada con la cantidad y capacidad de los registros de la unidad y presenta varias posibilidades de diseño. La primera alternativa consiste en implantar un juego de registros de tamaño fijo para mantener los datos. En este caso, las restricciones impuestas suponen un límite a la cantidad de espacio utilizado de los registros, sin poder sobrepasar su longitud. Otra posibilidad consiste en la creación de diversos juegos de registros de distintas longitudes y escoger el más favorable al valor de $\overline{d}$ para realizar los cálculos y contener los datos, ignorando el resto. También puede considerarse la construcción de registros mediante lógica reconfigurable con capacidad para adecuar dinámicamente su longitud y número en función de los requerimientos del problema. La técnica de gestión de la precisión que se utilice debe ser transparente al resto de módulos la unidad, de modo que en lo sucesivo, el desarrollo de los algoritmos y la instrumentación de las funciones considera en todo momento registros de una longitud fija w establecida.

A partir de los conceptos presentados hasta el momento se adopta la instrumentación $\Gamma^{VP}_{identidad}$ como formato de representación flexible de los números racionales. Esta relación corresponde con la *instrumentación Γ en precisión variable de la función identidad*. Sus características más importantes son las siguientes:

- Representación de los números mediante una notación fraccionaria en coma flotante en la que se codifica el periodo de los valores racionales con un conjunto de cifras específico que constituye la mantisa periódica.
- Adopción de una representación flexible de acuerdo con los requerimientos definidos por un parámetro $\overline{d}$ que interviene en la codificación.
- Concepción de operadores aritméticos que preserven las propiedades de exactitud y flexibilidad.

Una arquitectura que instrumente la función *identidad* para los números racionales mediante $\Gamma^{VP}_{identidad}$ representa una arquitectura en precisión variable.

$$A_{\Phi}^{VP} = \{\Gamma^{VP}_{identidad}\} \qquad [2.44]$$

Construcción de la función identidad

En este apartado se consideran los aspectos relativos a la arquitectura del procesador que implementan el formato de representación propuesto. Se destaca el interés por el diseño de $\Gamma^{\mathrm{VP}}_{\mathrm{identidad}}$ según las características de su instrumentación así como la inclusión del operador identidad como primitiva de la arquitectura junto con otros operadores aritméticos. Se adopta la base de representación binaria para la codificación de las cifras por razones obvias.

El planteamiento es esencialmente de modelización y habrá de ser la realización aplicada sobre la base de requerimientos reales la que incorpore y plantee los aspectos concretos relativos a la implementación: tecnología a utilizar, camino crítico, etc. En realidad, se espera que investigaciones posteriores profundicen en los aspectos de realismo de implementación hardware. No obstante, por cuestión de completitud, se ha decidido abordar en este apartado un supuesto de realización aportando líneas a seguir como punto de partida para su posterior desarrollo y perfeccionamiento.

Teniendo en cuenta lo anterior, el diseño propuesto para su implementación hardware se sustenta en la disposición de los campos que conforman el número en una palabra de longitud finita y en aportar flexibilidad en la distribución de las cifras fraccionarias entre la parte fija y periódica. La longitud w de los registros habrá sido establecida previamente por el módulo de gestión de la precisión según se ha mencionado. El formato consiste en cuatro campos diferenciados (figura 2-4): *signo, exponente, mantisa fija y mantisa periódica*. El significado y la codificación de cada campo es el siguiente:

- *Signo (s):* indica el signo del número mediante un solo bit con el convenio: 0 positivo, 1 negativo.
- *Exponente (e):* almacena el exponente del formato en coma flotante expresado en representación sesgada.
- *Mantisa fija (m_f):* secuencia de cifras que forma la parte fraccionaria del número racional no periódica. La mantisa fija se encuentra normalizada con el primer bit significativo a la derecha de la coma

fraccionaria. En caso de que no exista mantisa fija este campo no contendrá ninguna cifra.

- *Mantisa periódica (m_p):* secuencia de cifras que almacena la parte fraccionaria periódica del número racional.

El valor del número se obtiene según la expresión siguiente:

$$N = (-1)^{s} \cdot 0{,}m_f\widehat{m}_p \cdot 2^{E} \qquad [2.45]$$

donde,

E: Es el valor del exponente teniendo en cuenta su representación sesgada: $e = E_{S.}$

$0{,}m_f\widehat{m}_p$: Es el valor obtenido al establecer a la derecha de la coma las cifras fraccionarias de m_f concatenadas con la parte periódica m_p indefinidamente.

Este valor es equivalente al de la expresión [2.22] y tiene en cuenta la representación concreta de cada campo.

Las partes del número se colocan consecutivamente ocupando un registro de longitud fija. El signo, el exponente y la mantisa disponen de una cantidad de posiciones determinada para su representación, donde ésta última está formada por las cifras fijas y periódicas del número. Con esta instrumentación, el valor del parámetro $\overline{d}$ que limita la precisión viene dado por la cantidad de cifras que se destinan al exponente y la mantisa en su conjunto. En lo que sigue se tomará esta configuración como significado propio de $\overline{d}$ y se usará un par de números naturales para su expresión, como se muestra a continuación,

$$\overline{d} \equiv \langle d_{L(e)}, d_{L(m)} \rangle, \text{ con } d_{L(e)} \in \mathbb{N} \wedge d_{L(m)} \in \mathbb{N} \qquad [2.46]$$

tal que, las longitudes de los campos cumplen la siguiente relación:

$$(L(e) \leq d_{L(e)}) \wedge (L(m_f) + L(m_p) \leq d_{L(m)}) \qquad [2.47]$$

Gráficamente, el registro que contiene el dato mantiene la siguiente distribución de las longitudes de sus campos:

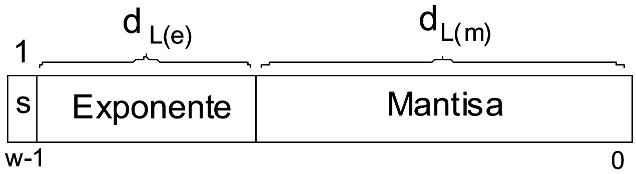

Figura 2-6: Distribución de las cifras de los campos del número

El reparto flexible de las cifras de la mantisa entre la parte fija y periódica de la misma requiere de un puntero que marque la separación entre ambas y permita su procesamiento separado. Para completar todas las cifras asignadas al campo de mantisa se concatenan las cifras del periodo formando un ciclo y se almacena su longitud y la de la mantisa fija junto con el puntero anterior. Estos tres datos se asocian al registro que contiene el número y se colocan adyacentes a él, como se observa en la estructura que ilustra la siguiente figura:

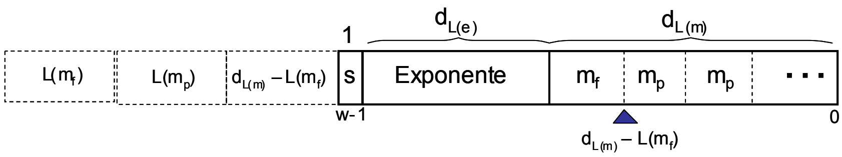

Figura 2-7: Estructura de la implementación del formato de doble mantisa

La mantisa de los números que carecen de parte fija se construye únicamente mediante la concatenación de las cifras de la mantisa periódica. En este caso, el puntero que marca la separación entre las mantisas toma el valor $d_{L(m)}$.

Cuando la longitud de alguna parte del número sobrepasa el límite impuesto por $\overline{d}$ no es posible su expresión exacta y será necesario ajustar la codificación al valor representable más cercano. El criterio que se propone es el de representar completamente el exponente y el signo y aplicar el recorte en la mantisa. Cuando ésta queda afectada por la limitación de cifras $d_{L(m)}$ se considera su expresión aproximada y se descartan las últimas cifras siguiendo el orden de magnitud de los dígitos.

Se presentan varias situaciones en relación con el valor de la condición $\overline{d}$ y la cantidad de cifras de cada campo de la representación:

- Si la cantidad de cifras $<d_{L(e)}, d_{L(m)}>$ es suficiente para la codificación completa del exponente y la mantisa en su conjunto se realiza su representación exacta.

$$(d_{L(e)} > L(e)) \wedge (d_{L(m)} \geq L(m_f) + L(m_p)) \Rightarrow \textit{Representación exacta} \qquad [2.48]$$

- Si la cantidad de cifras $d_{L(m)}$ no es suficiente para la representación completa de la mantisa del número se obtiene una representación aproximada.

  $$(d_{L(e)} > L(e)) \wedge (d_{L(m)} < L(m_f)+L(m_p)) \Rightarrow \textit{Representación aproximada} \quad [2.49]$$

  Se plantean dos escenarios en la codificación según el campo del número al que afecte el recorte de dígitos:

  - Si no es posible la codificación completa de la mantisa periódica carece de sentido su expresión incompleta y se elimina. Se asigna un valor cero al indicador de la longitud del periodo y utilizan sus dígitos para construir una mantisa fija extendida que aproveche la longitud total disponible.

    $$d_{L(m)} < L(m_f) + L(m_p) \quad [2.50]$$

    Tras esa reasignación de cifras se redondea el valor aproximado de la mantisa con el propósito de obtener un resultado con el menor error posible. Como ilustra la figura siguiente, el redondeo se realizará con las cifras del periodo que quedan fuera del límite. Se aplica cualquier sistema de redondeo referido en la bibliografía utilizando para ello la cantidad de dígitos necesaria [Even y Seidel, 2000], [Bruguera y Lang, 2000], [Park et al, 1999], [Quach et al, 1991] y [IEEE, 1985].

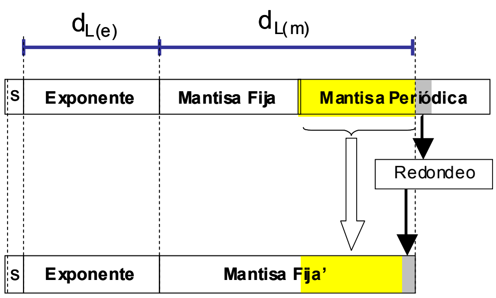

Figura 2-8: Transferencia de dígitos entre mantisas periódica y fija y redondeo

  - Si tampoco es posible la codificación completa de la mantisa fija se desecha directamente la mantisa periódica asignando al indicador de su longitud un valor cero y se procede a ajustar la mantisa fija mediante un redondeo sobre los dígitos sobrantes.

$$d_{L(m)} < L(m_f) \qquad [2.51]$$

En ambos casos, la posición de comienzo de la mantisa periódica que marca el puntero asociado es irrelevante.

- Finalmente, si la cantidad de dígitos disponibles $d_{L(e)}$ no abarca la representación del orden de magnitud, el número no es representable con las condiciones impuestas. En este caso el resultado mostrará una expresión de error.

$$d_{L(e)} \leq L(m_e) \Rightarrow \textit{Número no representable} \qquad [2.52]$$

El error que se produce en la codificación aproximada será cuantificado por el orden de magnitud del número. Por este motivo hay que seleccionar con especial cuidado el valor de la restricción $\overline{d}$ y considerar no sólo aspectos relativos al valor de los datos a codificar sino también al tipo y número de operaciones que intervienen. Las características del formato de representación mencionadas en el apartado anterior aseguran que para cualquier conjunto de valores racionales existe un tamaño de registro adecuado para su representación exacta. Si bien, será conveniente buscar una longitud de compromiso entre complejidad de la representación y capacidad de expresión.

La consideración de los campos de la mantisa en su conjunto, tanto en su representación exacta como aproximada, permite equiparar el formato de representación propuesto de doble mantisa con el clásico de coma flotante y presenta la capacidad adicional de codificación sin error de un conjunto de números racionales periódicos. En concreto, si se establece la estructura de los registros según el tamaño de los campos que indica la norma IEEE-754 se obtiene una representación compatible con la misma. Este hecho faculta el procesamiento de los números mediante la operatoria existente para coma flotante. Sin embargo, se obtienen ventajas de concebir métodos de operación adecuados para el nuevo formato que aprovechen su capacidad de expresión exacta y produzcan igualmente resultados sin error.

Las tablas siguientes muestran algunos ejemplos característicos de la codificación que realiza el formato de representación en comparación con el esquema estándar de representación en coma flotante. Se emplea una longitud de registro de 32 bits distribuidos entre 1 para el signo, 8 para el exponente y 23 para la mantisa ($d_{L(e)} = 8 \wedge d_{L(m)} = 23$) y se establece un sesgo para el exponente de $2^{L(e)-1}-1$. El conjunto de muestras pretende ser lo más heterogéneo posible y abarca números racionales periódicos y no periódicos tanto en decimal como en binario.

$N_i \in \mathbb{Q}$	Valor decimal	Valor binario
N_1	0,125	0,001
N_2	0,2	0,0011001100110011...
N_3	0,6	0,1001100110011001...
N_4	2,18	10,0010111000010100011111.
N_5	$7,\widehat{5}$	111,1000111000111...
N_6	25,04	11001,000010100001111010.
N_7	$37,\widehat{21}$	100101,0011011001...
N_8	$64,\widehat{01369863}$	1000000,000000111...

(a)

N_i	Representación según el formato IEEE de Simple Precisión		
	Exp.	Mantisa	Error
N_1	01111100	0000000000000000000000	0
N_2	01111100	1001100110011001100110	$1{,}19\cdot10^{-8}$
N_3	01111110	0011001100110011001101	$7{,}15\cdot10^{-8}$
N_4	10000000	0001011100001010001111	$1{,}72\cdot10^{-7}$
N_5	10000001	1110001110001110001110	$2{,}11\cdot10^{-7}$
N_6	10000011	1001000010100011110110	$2{,}75\cdot10^{-6}$
N_7	10000100	0010100110110010011011	$1{,}16\cdot10^{-6}$
N_8	10000101	0000000000001110000010	$3{,}86\cdot10^{-6}$

(b)

N_i	Representación según el formato propuesto			
	Exp.	m_f	m_p	Error
		Mantisa conjunta		
N_1	01111101	1	0	0
		10000000000000000000000		
N_2	01111101	--	1100	0
		11001100110011001100110		

N_3	01111111	-- ¦ 1001 10011001100110011001100	0
N_4	10000001	100 ¦ 01011100001010001111 1000101110000 101000111	$2,91 \cdot 10^{-7}$
N_5	10000010	1 ¦ 111000 11110001110001110001110	0
N_6	10000100	11001 ¦ 00001010001111010111 11001000010100011110110	$1,38 \cdot 10^{-7}$
N_7	10000101	1001 ¦ 0100110110 10010100110110010011011	0
N_8	10000110	1000000 ¦ 000000111 10000000000001110000001	0

(c)

Tabla 2-4: Ejemplos de representación de números racionales
(a) Valores racionales en decimal y binario
(b) Representación en el formato IEEE Simple Precisión
(c) Representación en el formato propuesto

Los ejemplos de las tablas anteriores ponen de manifiesto que el formato propuesto realiza la representación exacta siempre que la cantidad de cifras significativas de los números sea menor o igual que las posiciones disponibles y es capaz de codificar números con una cantidad de cifras infinita gracias a la existencia de una mantisa periódica.

El modelo de representación deja abierta la posibilidad de plantear codificaciones especiales para expresar aquellas situaciones anómalas o de excepción que se pueden producir: representación del cero, infinito, desbordamiento y error [Schulte, 2000], [IEEE, 1985].

Evaluación empírica

Una vez conocidas las características del método de representación propuesto así como su capacidad de expresión se realiza una serie de pruebas con ánimo de analizar su comportamiento. Los experimentos se centran en estudiar la codificación de la parte significativa del número considerando que tanto el exponente como el signo se representan correctamente en todos los casos. Se han realizado dos conjuntos de pruebas mediante una simulación del formato en un entorno de programación en C[1].

Experimentos I

En este primer bloque de experimentos se analiza la relación de números con mantisa fija y periódica así como la cantidad de cifras necesarias en la codificación de los valores racionales.

La siguiente tabla muestra los 80 primeros términos de la serie armónica en su representación binaria fraccionaria posicional antes de su codificación por el formato.

1/n	m_f	m_p
1/1	0	0
1/3	--	01
1/5	--	0011
1/7	--	001
1/9	--	000111
1/11	--	0001011101
1/13	--	000100111011
1/15	--	0001
1/17	--	00001111
1/19	--	000011010111100101
1/21	--	000011
1/23	--	00001011001
1/25	--	00001010001111010111
1/27	--	000010010111101101
1/29	--	0000100011010011110111001011
1/31	--	00001
1/33	--	0000011111
1/35	--	000001110101
1/37	--	000001101110101100111111001000

1/n	m_f	m_p
1/2	1	0
1/4	01	0
1/6	0	01
1/8	001	0
1/10	0	0011
1/12	00	01
1/14	0	001
1/16	0001	0
1/18	0	000111
1/20	00	0011
1/22	0	0001011101
1/24	000	01
1/26	0	000100111011
1/28	00	001
1/30	0	0001
1/32	00001	0
1/34	0	00001111
1/36	00	000111
1/38	0	000011010111100101

[1] *Entorno de desarrollo*: C++ Builder 5.0 Professional. Borland software corporation.
Entorno hardware: Procesador Intel Pentium III 500 MHz 512 MB RAM.

11,00100100001111110110101010001000100001011010001100001000110100110001001100011001100010100010111000000001101110000001110011010001001010010000001001001110000010001
0001010011001111100110001110100

		1010011
1/39	--	000001101001
1/41	--	000001100011111100111
1/43	--	00000101111101
1/45	--	000001011011
1/47	--	0000010101110010011000 1
1/49	--	000001010011100101111
1/51	--	00000101
1/53	--	000001001101010010000111001111101100101011011110001 1
1/55	--	000001001010011110 01
1/57	--	0000010001111101 11
1/59	--	000001000101011011000111100101111101110101001001110000110 1

1/40	000	0011
1/42	0	000011
1/44	00	0001011101
1/46	0	00001011001
1/48	0000	01
1/50	0	00001010001111010111
1/52	00	000100111011
1/54	0	000010010111101101
1/56	000	001
1/58	0	0000100011010011110111001011
1/60	00	0001

1/61	--	00000100001100100101110001010 01111101111001101101000111010 11	1/62	0	00001
1/63	--	000001	1/64	000001	0
1/65	--	000000111111	1/66	0	0000011111
1/67	--	00000011110100100010011000110 10101111110000101101110110011 10010101	1/68	00	00001111
1/69	--	00000011101101011110011	1/70	0	000001110101
1/71	--	00000011100110110000101011010 001001	1/72	000	000111
1/73	--	000000111	1/74	0	000001101110101100 111110010001010011
1/75	--	00000011011010011101	1/76	00	000011010111100101
1/77	--	00000011010100110001110111101 1	1/78	0	000001101001
1/79	--	00000011001111011001000111010 0101010001	1/80	0000	0011

Tabla 2-5: Representación fraccionaria posicional binaria de los 80 primeros términos de la serie armónica

En las representaciones de la tabla 2-5 se observan diferencias en la codificación de los números según el denominador de la fracción que los genera. La mantisa fija está formada por una cantidad de ceros igual a la mayor potencia de dos divisor del denominador de la fracción. Si éste no fuera divisible entre dos entonces carece de mantisa fija. En cuanto a la mantisa periódica, se comprueba que todos los números contienen un periodo y que éste sólo está formado por el dígito cero cuando el denominador es potencia de dos.

El formato de representación propuesto contiene algunos cambios en relación con la codificación posicional que muestra la tabla 2-5 que no alteran la expresión directa del valor del número. Se mantiene un campo de exponente para expresar el orden de magnitud y las mantisas se expresan normalizadas para contener su valor en el intervalo [0,5; 1). El efecto inmediato que se provoca es la reducción de la longitud de la representación binaria.

Para comprobar las características de la mantisa de los números codificados con el formato se realiza un conjunto de pruebas de representación de números racionales en un amplio intervalo.

El perfil de estos experimentos es el siguiente:

- Codificación de números racionales aleatorios[2] en el modelo de representación propuesto.
- Los valores pertenecen al intervalo (0..10000], donde cada número se genera mediante la construcción de una fracción $\frac{a}{b}$, siendo a y b valores enteros aleatorios en el rango [1..10000].
- Realización de 10^9 representaciones numéricas.

Los resultados obtenidos se muestran en la siguiente tabla:

Prueba a/b	%
Números con mantisa fija	61,10
Números con mantisa periódica distinta de 0	99,58
Números que carecen de algún tipo de mantisa	39,31

Tabla 2-6: Características de la representación racional de a/b en el formato propuesto

Los datos de la tabla anterior muestran que existe un conjunto de números, 38,90 %, para los que no es necesaria la mantisa fija. Este porcentaje es significativamente menor que el 50% obtenido para denominadores impares en la representación posicional directa que muestra la tabla 2-5. En este formato, la representación de la mantisa fija será necesaria cuando los números a codificar sean mayores o iguales a uno o cuando el denominador de la fracción que lo genera sea potencia de dos. También se observa que la cantidad de representaciones con mantisa periódica es muy elevada y prácticamente todos los números disponen de ella. Este hecho se debe a la naturaleza binaria de la base de representación.

[2] Para la generación de números enteros aleatorios se han utilizado las funciones *randomize* y *random* de la librería *stdlib.h* perteneciente al entorno de programación anteriormente mencionado.

Se realiza un segundo conjunto de pruebas, del mismo tipo que las anteriores, pero considerando ahora fracciones de la serie armónica con la forma $\frac{1}{b}$, con $b \in [1..10000]$ para evitar simplificaciones entre factores comunes de los numeradores y denominadores de las fracciones generadoras. Los datos que resultan de esta prueba se muestran en la tabla siguiente.

Prueba 1/b	%
Números con mantisa fija	0,14
Números con mantisa periódica distinta de 0	99,86
Números que carecen de algún tipo de mantisa	100

Tabla 2-7: Características de la representación racional de 1/b en el formato propuesto

En la tabla anterior se observa que debido a la normalización de la representación la mayoría de los números carecen de mantisa fija, y sólo la contienen aquellos cuyo denominador es potencia de dos.

Para determinar la complejidad espacial de las representaciones, las pruebas siguientes se orientan a estudiar la relación existente entre la cantidad de dígitos necesarios en la codificación de un número racional y el numerador y el denominador de la fracción que lo genera. Para una mayor claridad se presentan los resultados de forma separada para la mantisa fija y periódica en relación con el numerador y denominador de la fracción generadora respectivamente. En el caso general, la longitud de las mantisas estará en una posición intermedia consecuencia de las simplificaciones entre factores comunes de la fracción.

La figura 2-9 muestra cómo la cantidad de cifras necesaria para contener la mantisa fija aumenta de acuerdo con el logaritmo entero del numerador de la fracción, tal y como sucede en el resto de las codificaciones de coma flotante. La mantisa periódica permanece a cero.

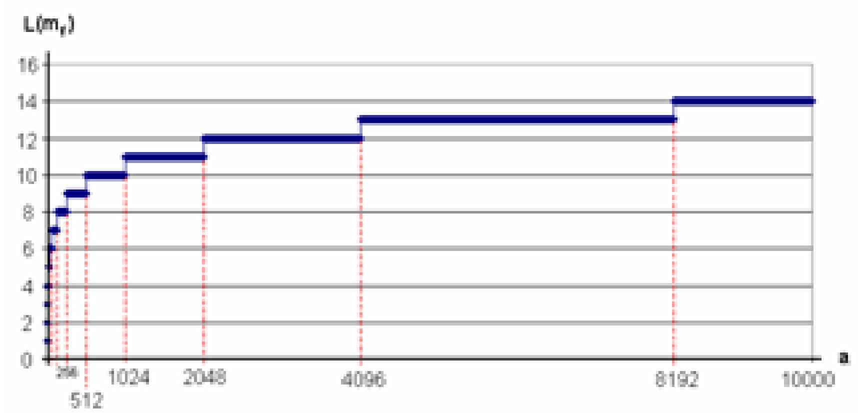

Figura 2-9: Cantidad de cifras necesaria en la representación de la mantisa fija de a/1

La figura 2-10 muestra una tendencia creciente de la cantidad de cifras de la mantisa periódica con respecto al denominador de la fracción. Esta circunstancia provoca que la cantidad de cifras del periodo dependa proporcionalmente de su valor, como era de esperar debido a su naturaleza.

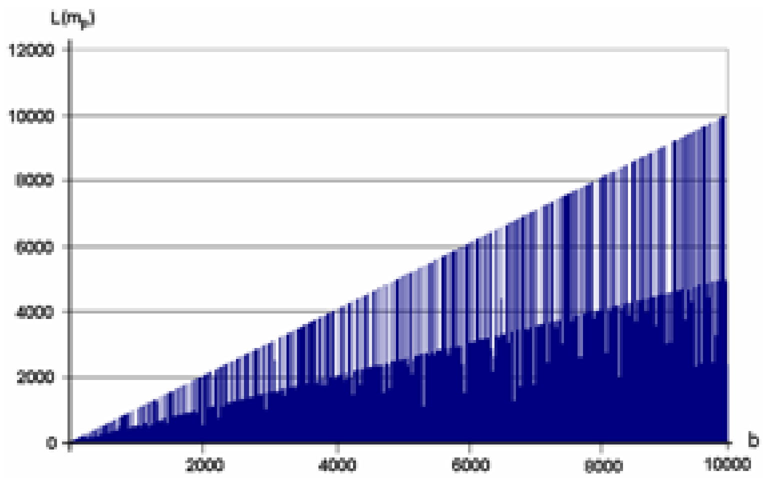

Figura 2-10: Cantidad de cifras necesaria en la representación de la mantisa periódica 1/b

Como se observa en las gráficas 2-9 y 2-10, la cantidad de cifras necesarias para la representación exacta de un número está en relación con el tamaño del numerador y denominador de la fracción según la siguiente expresión:

$$\forall\, x \in \mathbb{Q} \;/\; x = \frac{a}{b},\ \text{si } d_{L(m)} \geq b + lg_2\, a \Rightarrow \textit{Representación exacta} \qquad [2.53]$$

Esta cantidad de cifras puede ser menor debido a la presencia de factores comunes entre el numerador y el denominador de la fracción así como por denominadores con factores potencia de dos.

Experimentos II

El segundo grupo de pruebas tiene como objeto comparar el método propuesto y los métodos de representación convencionales presentes en la mayoría de los computadores de propósito general, en concreto el formato de representación IEEE-754. Para ello, se codifican números racionales periódicos y no periódicos tanto en el formato de representación propuesto como en el formato de representación IEEE-754 en simple y doble precisión. Los valores se toman en un intervalo normalizado entre 0 y 1 para centrar la atención en la precisión de las codificaciones realizadas de la mantisa y no en su orden de magnitud. De esta prueba se extraen conclusiones acerca de la expresividad de dicho formato y de las desviaciones que produce en la representación numérica racional. El perfil de los experimentos es el siguiente:

- Cada número no periódico pertenece al intervalo $[0, 1)$ y consta de una cantidad de 128 cifras fraccionarias aleatorias significativas.
- Cada número periódico pertenece al intervalo $(0, 1]$ y se construye mediante una fracción $1/b$, donde b es un valor entero aleatorio no potencia de 2 en el rango $[1..10000]$.
- Realización de 10^9 representaciones numéricas para cada caso.

Las pruebas se basan en la capacidad de expresión sin error de los números racionales con el método propuesto. El procedimiento consiste en comparar para cada número la codificación que proporcionan los formatos convencionales con las correspondientes cifras exactas de la representación en el formato propuesto, de forma que se establezca una medida absoluta del error cometido así como la desviación sobre la codificación correcta. De estas pruebas se obtiene una evaluación de la corrección de las codificaciones realizadas por los formatos de representación de IEEE-754 de simple y doble precisión en su capacidad de representación numérica.

Una primera orientación acerca de la calidad de la representación viene dada por la posición que ocupa el dígito de mayor orden de magnitud con valor distinto respecto al correspondiente de la representación exacta. Si bien no es indicativo de la dimensión del error cometido, sí que representan una medida de la similitud entre el valor exacto y el valor representado. La tabla 2-8 muestra el valor promedio de ésta primera posición con un dígito distinto de la mantisa. Se observa que en todos los casos su valor está muy cerca de la cantidad de cifras que se destinan a la representación.

Formato	Números no periódicos		Números periódicos	
	Posición	σ	Posición	σ
IEEE-754 Simple Precisión	23,36	8,97	23,08	8,34
IEEE-754 Doble Precisión	52,01	15,05	51,82	14,88

Tabla 2-8: Primera posición distinta en promedio

La tabla 2-9 muestra el promedio del error absoluto producido en valor de la mantisa sin considerar su orden de magnitud. Se debe señalar que debido a la naturaleza de formato en coma flotante estos errores se ven afectados por un exponente que puede amplificar su valor [Goldberg, 1991], [Bohlender, 1990].

Formato	Números no periódicos		Números periódicos	
	Error	σ	Error	σ

IEEE-754 Simple Precisión	$2{,}979 \cdot 10^{-8}$	$2{,}055 \cdot 10^{-13}$	$2{,}978 \cdot 10^{-8}$	$1{,}186 \cdot 10^{-13}$
IEEE-754 Doble Precisión	$5{,}551 \cdot 10^{-17}$	$1{,}856 \cdot 10^{-33}$	$5{,}549 \cdot 10^{-17}$	$1{,}027 \cdot 10^{-33}$

Tabla 2-9: Error promedio en la codificación IEEE-754 de simple y doble precisión

Conclusiones

En este capítulo se ha presentado la función identidad en precisión variable a partir de un formato de representación de números racionales y su implementación a bajo nivel. De la misma se destacan los siguientes aspectos:

- El formato de representación de doble mantisa permite la codificación de números racionales sin error y constituye una alternativa de expresión numérica racional exacta a la representación simbólica.

- La codificación de los números se fundamenta en su representación fraccionaria posicional, lo que permite realizar operaciones aritméticas directamente sobre los campos del formato.

- La instrumentación de la función identidad hace un uso flexible de registros de tamaño fijo y mantiene las características del formato. En la representación aproximada de los datos el esquema se comporta como un formato clásico de coma flotante.

- Las pruebas realizadas ponen de manifiesto la información que se pierde en la codificación fraccionaria de los números racionales mediante las técnicas tradicionales. Los análisis realizados deben tomarse para valorar la conveniencia de su utilización en determinados problemas.