Floating-point numbers in Ginger

November 2, 2012

In this section I shall describe a new way to represent floating point numbers in Ginger. We consider $m \times m$ matrix multiplication and require the input entries in the set T=$\{a/b :\mid a \mid \leq 2^{N_a}, b \in \{1, 2, 2^2, 2^3, ..., 2^{N_b}\}\}$. Previously, it was shown that $p > (m+1)^2 \cdot 2^{4(N_a+N_b)}$ is necessary for making θ 1-1 which is required to make the mapping isomorphic from U to $\mathbb{Q}/p$. In this exercise I show that by modifying the definition of θ and the mapped field, one can obtain a better bound on p (i.e., $p > \max\{2m \cdot 2^{2N_a+2N_b}, 2N_a + 4N_b + \log_2 m\}$). I do not make any changes to step $C1$. The changes I propose, with the resulting proofs are described below:
Define θ as follows:

$$\theta : U \to \mathbb{F}$$
$$\frac{a}{b = 2^k} \mapsto (a \, mod \, p, k \, mod \, p)$$

The field $\mathbb{F}$ is the set of equivalence classes on the set $\mathbb{Z}/p \times \mathbb{Z}/p$ under the equivalence relation: (a,b) $\backsim (c, d)$ iff $\ell \equiv s \pmod{p}$ and $r_1 + d \equiv r_2 + b \pmod{p}$, where $a = \ell \cdot 2^{r_1}$ and $c = s \cdot 2^{r_2}$. We have written a and c in this form by factoring out all the powers of 2. Every integer can be written in this form (for integer greater than 1 this follows from fundamental theorem of arithmetic, and $1 = 1 \cdot 2^0$ and $0 = 0 \cdot 2^k$ where k is arbitrary).
The addition operation is defined by $(a, b) + (c, d) = (a \cdot 2^d + c \cdot 2^b, b + d)$
The multiplication operation is defined by $(a, b) \cdot (c, d) = (a \cdot c, b + d)$
$(1, 0)$ is the multiplicative identity. For any $(a, b) \in \mathbb{F}$ (except additive identity of course) $(a^{-1}, -b)$ is the multiplicative inverse. $(0, 0)$ is the additive identity (in fact $(0, 0) \backsim (0, k)$). For any $(a, b) \in \mathbb{F}$, $(-a, b)$ is the additive inverse. Now we show that θ preserves addition and multiplication rules for $q_1, q_2 \in U$.

$$\theta(q_1 + q_2) = \theta(\frac{a_1}{b_1} + \frac{a_2}{b_2}) = \theta(\frac{a_1 \cdot b_2 + a_2 \cdot b_1}{2^{k_1} \cdot 2^{k_2}}) = (a_1 \cdot b_2 + a_2 \cdot b_1, k_1 + k_2)$$
$$\theta(q_1) + \theta(q_2) = (a_1, k_1) + (a_2, k_2) = (a_1 \cdot 2^{k_2} + a_2 \cdot 2^{k_1}, k_1 + k_2)$$
$$so, \theta(q_1 + q_2) = \theta(q_1) + \theta(q_2)$$

Similarly, the multiplication rule holds:

$$\theta(q_1 \cdot q_2) = \theta(\frac{a_1 \cdot a_2}{b_1 \cdot b_2}) = \theta(\frac{a_1 \cdot a_2}{2^{k_1} \cdot 2^{k_2}}) = (a_1 \cdot a_2, k_1 + k_2)$$
$$\theta(q_1) \cdot \theta(q_2) = (a_1, k_1) \cdot (a_2, k_2) = (a_1 \cdot a_2, k_1 + k_2)$$
$$so, \theta(q_1 \cdot q_2) = \theta(q_1) \cdot \theta(q_2)$$

Claim: θ is a function from U to $\mathbb{F}$

Proof: We need to show that if $q_1 = q_2$ then $\theta(q_1) \equiv \theta(q_2)$

$$q_1 = q_2$$
$$\frac{a_1}{b_1} = \frac{a_2}{b_2}$$
$$a_1 \cdot b_2 = a_2 \cdot b_1$$
$$\ell \cdot 2^{r_1} \cdot 2^{k_2} = s \cdot 2^{r_2} \cdot 2^{k_1}$$

because we can write: $a_1 = \ell \cdot 2^{r_1}$ and $a_2 = s \cdot 2^{r_2}$.

$$\ell \cdot 2^{r_1+k_2} = s \cdot 2^{r_2+k_1}$$

so this implies $\ell = s$ and $r_1 + k_2 = r_2 + k_1$. To see why this is true first assume both sides are greater than one, so by fundamental theorem of arithmetic we can write them as a product of distinct primes. Now ℓ and s are numbers such that they contain all the other primes except 2.

$$\ell \cdot 2^{r_1+k_2} = s \cdot 2^{r_2+k_1}$$

means that on both sides the primes should be the same, and they should have the same powers. Since ℓ and s do not contain the prime 2, so therefore $r_1 + k_2 = r_2 + k_1$ as it's the power of 2. $\ell = s$ as it contains all the other primes except 2. Also knowing that the powers of 2 on both sides are equal trivially implies $\ell = s$. Now if both sides are equal to 1, then $\ell = s = 1$ and $r_1 + k_2 = r_2 + k_1 = 0$. If both sides are zero, this means a_1 and a_2 were both zero, so this implies $\ell = s = 0$ and $r_1 + k_2 = r_2 + k_1$ since for any given k_1 and k_2 we could choose arbitrary r_1 and r_2 (as we can write $0 = 0 \cdot 2^r$ where r is arbitrary). $\ell = s$ and $r_1 + k_2 = r_2 + k_1$ naturally means $\ell = s \pmod p$ and $r_1 + k_2 = r_2 + k_1 \pmod p$ which implies $\theta(q_1) \equiv \theta(q_2)$ (see definition of the new equivalence relation above) .

Claim: If $p > \max\{2m \cdot 2^{2N_a+2N_b}, 2N_a + 4N_b + \log_2 m\}$ then θ is 1-1 function.

Proof: We need to prove that if $\theta(q_1) \equiv \theta(q_2)$ then $q_1 = q_2$. Suppose for the sake of contradiction that:

$$q_1 \neq q_2$$
$$\frac{a_1}{b_1} \neq \frac{a_2}{b_2}$$
$$a_1 \cdot b_2 \neq a_2 \cdot b_1$$
$$\ell \cdot 2^{r_1+k_2} \neq s \cdot 2^{r_2+k_1}$$

now only two cases are possible:

case 1: $\ell \neq s$ and $r_1 + k_2 \neq r_2 + k_1$

$\theta(q_1) \equiv \theta(q_2)$ means that $\ell = s \pmod p$ and $r_1 + k_2 = k_1 + r_2 \pmod p$. Now $\ell \neq s$ implies that $\ell - s = hp$ where h is an integer other than zero (as $\ell = s \pmod p$). So, it follows:

$$\mid \ell - s \mid \geq p$$

$$| \ell | + | s | \geq p$$

$| \ell | \leq | \ell \cdot 2^{r_1} | \leq | a_1 | \leq m \cdot 2^{2N_a+2N_b}$ where the last inequality uses the bound on the numerator from Claim B.1. Similarly, $| s | \leq | s \cdot 2^{r_2} | \leq | a_2 | \leq m \cdot 2^{2N_a+2N_b}$. Therefore, it follows:

$$2m \cdot 2^{2N_a+2N_b} \geq p \tag{1}$$

now similarly, $r_1 + k_2 \neq r_2 + k_1$ and $r_1 + k_2 = r_2 + k_1 \pmod p$ implies that:

$$| (r_1 + k_2) - (r_2 + k_1) | \geq p$$

$$| (r_1 - r_2) - (k_1 - k_2) | \geq p$$

$$| r_1 - r_2 | + | k_1 - k_2 | \geq p$$

$| 2^{r_1} | \leq | \ell \cdot 2^{r_1} | \leq | a_1 | \leq 2^{2N_a+2N_b+\log_2 m}$. This implies $r_1 \leq 2N_a + 2N_b + \log_2 m$. Similarly, $r_2 \leq 2N_a + 2N_b + \log_2 m$. Hence, $| r_1 - r_2 | \leq 2N_a + 2N_b + \log_2 m$. Now, $b_1 = 2^{k_1} \leq 2^{2N_b}$(follows from the denominator bound in Claim B.1). This implies $k_1 \leq 2N_b$ and $k_2 \leq 2N_b$, $| k_1 - k_2 | \leq 2N_b$ then immediately follows. Using the above results:

$$2N_a + 2N_b + \log_2 m + 2N_b \geq p$$

$$2N_a + 4N_b + \log_2 m \geq p \tag{2}$$

Both (1) and (2) lead to contradiction as $p > \max\{2m \cdot 2^{2N_a+2N_b}, 2N_a + 4N_b + \log_2 m\}$.

case 2: $\ell \neq s$ or $r_1 + k_2 \neq r_2 + k_1$
so either $\ell \neq s$ or $r_1 + k_2 \neq r_2 + k_1$. Assuming $\ell \neq s$ and $\ell = s \pmod p$ (as $\theta(q_1) \equiv \theta(q_2)$) leads us to (1) as shown above which is a contradiction. On the other hand, assuming $r_1 + k_2 \neq r_2 + k_1$ with $r_1 + k_2 = r_2 + k_1 \pmod p$ leads to (2) which again results in a contradiction. Since, either $\ell \neq s$ or $r_1 + k_2 \neq r_2 + k_1$ must hold, so we get a contradiction.
Since we get a contradiction in all possible cases, so this means θ is a 1-1 function.