Scientific Computing HW 1

Ryan Chen

September 4, 2024

1. (a) The expressions result in equal floating point numbers since

$$\begin{aligned} \mathrm{fl}((x * y) + (z - w)) &= \mathrm{fl}((z - w)) + (x * y) & \qquad \mathrm{fl}(a + b) = \mathrm{fl}(b + a) \\ &= \mathrm{fl}((z - w) + (y * x)) & \qquad \mathrm{fl}(a * b) = \mathrm{fl}(b * a) \end{aligned}$$

   (b) Let $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ be the relative error in computing $x + y, (x + y) + z, y + z, x + (y + z)$ respectively.

$$\begin{aligned} \mathrm{fl}((x + y) + z) &= ((x + y)(1 + \epsilon_1) + z)(1 + \epsilon_2) \\ &= (x + y)(1 + \epsilon_1) + z + ((x + y)(1 + \epsilon_1) + z)\epsilon_2 \\ &= x + y + (x + y)\epsilon_1 + z + (x + y + z)\epsilon_2 + (x + y)\epsilon_1\epsilon_2 \\ &= (x + y + z) + (x + y)\epsilon_1 + (x + y + z)\epsilon_2 + (x + y)\epsilon_1\epsilon_2 \end{aligned}$$

   Similarly,

$$\mathrm{fl}(x + (y + z)) = \mathrm{fl}((y + z) + x) = (y + z + x) + (y + z)\epsilon_3 + (y + z + x)\epsilon_4 + (y + z)\epsilon_3\epsilon_4$$

   In general the expressions do not result in equal floating point numbers.

   (c) Write H for oneHalf. Floating point multiplication by $1/2$ is always exact; to multiply a floating point number by $1/2$, simply decrease the exponent by 1. In light of this, the exact result of $(x * H) + (y * H)$ equals the exact result of $x + y$ but with its exponent decreased by 1. In other words, the exact result of $(x * H) + (y * H)$ equals the exact result of $(x + y) * H$. Thus these expressions give equal floating point values.

   (d) In general floating point multiplication by $1/3$ is not exact, i.e. it produces relative errors. Write T for oneThird. Let $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5$ be the relative errors of computing $x * T, y * T, (x * T) + (y * T), x + y, (x + y) * T$ respectively.

$$\begin{aligned} \mathrm{fl}((x * T) + (y * T)) &= ((xT)(1 + \epsilon_1) + (yT)(1 + \epsilon_2))(1 + \epsilon_3) \\ &= [(x + y) + x\epsilon_1 + y\epsilon_2 + (x + x\epsilon_1 + y + y\epsilon_2)\epsilon_3]\, T \\ &= [(x + y) + x(\epsilon_1 + \epsilon_3) = y(\epsilon_2 + \epsilon_3) + x\epsilon_1\epsilon_3 + y\epsilon_2\epsilon_3]\, T \end{aligned}$$

$$\begin{aligned} \mathrm{fl}((x + y) * T) &= (x + y)(1 + \epsilon_4)T(1 + \epsilon_5) \\ &= ((x + y) + (x + y)\epsilon_4)(1 + \epsilon_5)T \\ &= [(x + y) + x(\epsilon_4 + \epsilon_5) + y(\epsilon_4 + \epsilon_5) + (x + y)\epsilon_4\epsilon_5]\, T \end{aligned}$$

   In general the expressions do not result in equal floating point numbers.

2. We state some preliminaries for this problem. Let f^n denote the $n-$fold composition of f. Define the left bit shift map

$$g : [0, 1] \to [0, 1], \quad g(x) = \begin{cases} 2x, & x < \frac{1}{2} \\ 2x - 1, & x \geq \frac{1}{2} \end{cases}$$

The name of the map comes from the fact that, for binary expansions of numbers in $[0, 1]$,

$$g(0.b_1 b_2 \dots) = 0.b_2 b_3 \dots, \quad b_i \in \{0, 1\}$$

We also state some lemmas.

Lemma 1. $f \circ g = f^2$.

Proof. Split into cases.

- If $0 \le x < \frac{1}{4}$ then $0 \le 2x < \frac{1}{2}$, so
$$f(g(x)) = f(2x) = 4x$$
$$f(f(x)) = f(2x) = 4x$$
- If $\frac{1}{4} \le x < \frac{1}{2}$ then $\frac{1}{2} \le 2x < 1$, so
$$f(g(x)) = f(2x) = 2 - 4x$$
$$f(f(x)) = f(2x) = 2 - 4x$$
- If $\frac{1}{2} \le x < \frac{3}{4}$ then $0 \le 2x - 1 < \frac{1}{2}$ and $\frac{1}{2} < 2 - 2x \le 1$, so
$$f(g(x)) = f(2x - 1) = 4x - 2$$
$$f(f(x)) = f(2 - 2x) = 4x - 2$$
- If $\frac{3}{4} \le x \le 1$ then $\frac{1}{2} \le 2x - 1 \le 1$ and $0 \le 2 - 2x \le \frac{1}{2}$, so
$$f(g(x)) = f(2x - 1) = 4 - 4x$$
$$f(f(x)) = f(2 - 2x) = 4 - 4x$$

Lemma 2. $f \circ g^n = f^{n+1}$ for all $n \ge 1$.

Proof. Repeatedly apply lemma 1.

$$f \circ g^n = f \circ \underbrace{g \circ \cdots \circ g}_{n \text{ times}} = f^2 \circ \underbrace{g \circ \cdots \circ g}_{n-1 \text{ times}} = \cdots = f^{n+1}$$

Lemma 3. The sequence $f^n(x_0)$ (in n) is eventually $p-$periodic iff there exists N such that $f^{N+p}(x_0) = f^N(x_0)$.

Proof. ($\implies$) Eventual periodicity gives N such that $f^{n+p}(x_0) = f^n(x_0)$ for all $n \ge N$. In particular $f^{N+p}(x_0) = f^N(x_0)$.

($\impliedby$) Let $p = M - N$. Then for all $n \ge N$,

$$f^{n+p}(x_0) = f^{n-N+N+p}(x_0) = f^{n-N}(f^{N+p}(x_0)) = f^{n-N}(f^N(x_0)) = f^n(x_0)$$

Thus the sequence $f^n(x_0)$ is eventually $p-$periodic.

The last lemma is taken for granted since it is a relatively well–known fact.

Lemma 4. A number is rational iff it has an eventually periodic binary expansion.

(a) We find fixed points in the interval $[0, 1/2)$ by

$$x^* = f(x^*) \implies x^* = 2x^* \implies x^* = 0$$

We find fixed points in the interval $[1/2, 1]$ by

$$x^* = f(x^*) \implies x^* = 2 - 2x^* \implies x^* = \frac{2}{3}$$

Thus the fixed points are $x^* = 0, 2/3$.

(b) Fix $x_0 \in \mathbb{Q} \cap [0,1]$. By lemma 4, $x_0/2 \in \mathbb{Q}$ admits an eventually periodic binary expansion. Using the left bit shift map g, the sequence $g^n(x_0/2)$ is eventually periodic, say with period p, so lemma 3 gives N such that $g^{N+p}(x_0/2) = g^N(x_0/2)$. Note that $f(x_0/2) = x_0$ since $0 \le x_0/2 \le 1/2$. Then

$$\begin{aligned}
f^{N+p}(x_0) &= f^{N+p+1}\left(\frac{x_0}{2}\right) && f\left(\frac{x_0}{2}\right) = x_0 \\
&= f\left(g^{N+p}\left(\frac{x_0}{2}\right)\right) && \text{by lemma 2} \\
&= f\left(g^{N}\left(\frac{x_0}{2}\right)\right) && g^{N+p}(x_0/2) = g^N(x_0/2) \\
&= f^{N+1}\left(\frac{x_0}{2}\right) && \text{by lemma 2} \\
&= f^{N}(x_0) && f\left(\frac{x_0}{2}\right) = x_0
\end{aligned}$$

Thus by lemma 3, the sequence $f^n(x_0)$ is eventually $p-$periodic.

(c) Fix a period p. Referring to the argument in the last part, it suffices to find $x_0 \in \mathbb{Q} \cap [0,1]$ such that $g^{N+p}(x_0/2) = g^N(x_0/2)$, so that the sequence $f^n(x_0)$ is eventually $p-$periodic.
Write a binary expansion of $x_0 \in [0,1]$,

$$x_0 = 0.b_1 b_2 \dots, \quad b_i \in \{0,1\}$$

Imposing the condition $g^{N+p}(x_0/2) = g^N$ for some N gives

$$0.b_{N+p} b_{N+p+1} \dots = 0.b_N b_{N+1} \dots$$

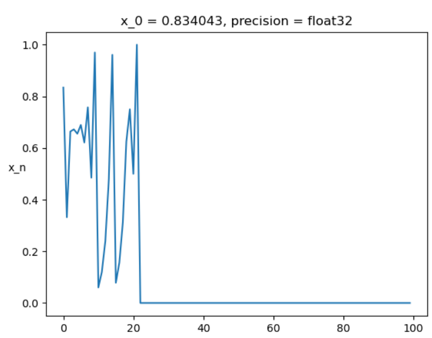
which is equivalent to $b_{n+p} = b_n$ for all $n \ge N$. Plugging into the expansion,

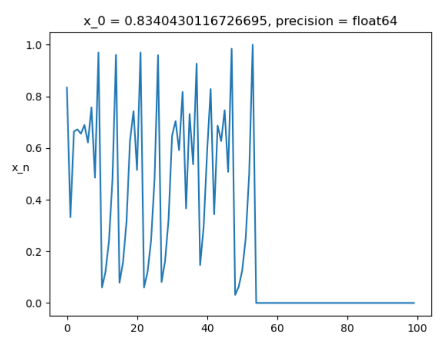
$$x_0 = 0.b_1 b_2 \dots b_{N-1} \overline{b_N b_{N+1} \dots b_{N+p-1}}$$

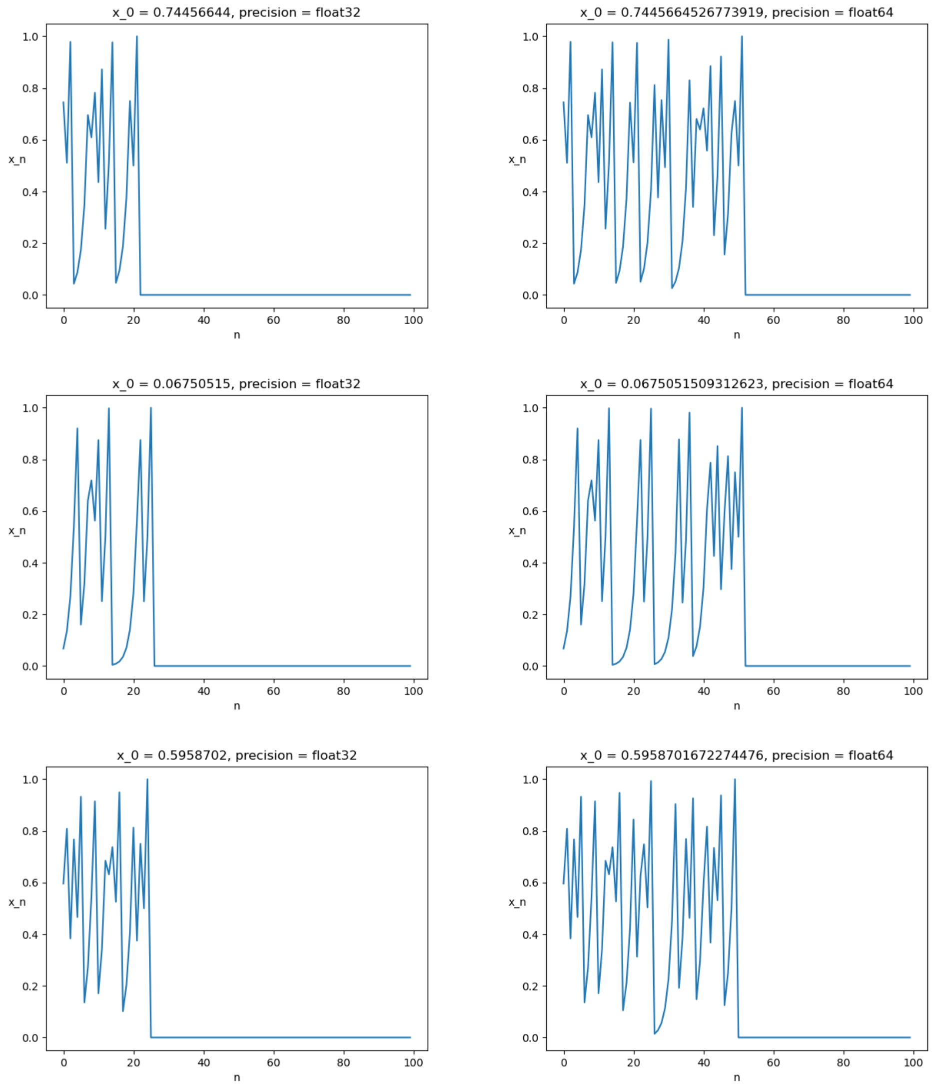
This is an eventually periodic expansion, so by lemma 4, $x_0 \in \mathbb{Q} \cap [0,1]$.
In conclusion, given p, the sequence $f^n(x_0)$ is eventually $p-$periodic for x_0 with binary expansion

$$x_0 = 0.b_1 b_2 \dots b_{N-1} \overline{b_N b_{N+1} \dots b_{N+p-1}} \in \mathbb{Q} \cap [0,1], \quad N \ge 1, \quad b_i \in \{0,1\}, \quad 1 \le i \le N+p-1$$

(d) Below are a few sequences of iterates.







The iterates eventually become zero, and this happens even faster for single precision.

(e) Floating point numbers have terminating binary expansions, so iteration under the left bit shift map g eventually vanishes. This, along with the facts that $f^{n+1} = f \circ g^n$ (lemma 2) and $f(0) = g(0) = 0$, implies that iteration of floating point numbers under f eventually vanishes. Moreover, single precision floating point numbers have fewer binary digits, which is why iteration under f vanishes faster than for double precision.