SOME MATH FOR NUMERICS

NICHOLAS F. MARSHALL

1. Introduction

In this note, we introduce four ideas underlying many numerical methods:

- asymptotic series,
- Richardson extrapolation,
- contraction mapping, and
- simple iteration.

To motivate these ideas, we study classical approximations of π and $\sqrt{2}$. We emphasize that our purpose for considering these classic approximations is only to provide familiar examples for introducing the four ideas listed above (these constants are readily available in all numerical programming environments).

2. Leibniz formula, asymptotic series, and Richardson extrapolation

2.1. Leibniz formula.

Recall that Leibniz formula for π states that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots,$$

or more formally

$$\pi = 4 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}.$$

Let $f(n)$ be the sum of the first $2n$ terms in the series

$$f(n) := 4 \sum_{k=0}^{2n-1} \frac{(-1)^k}{2k+1}. \tag{1}$$

By associating the even and odd terms in the sum defining $f(n)$

$$f(n) = 4\left(1 - \frac{1}{3}\right) + 4\left(\frac{1}{5} - \frac{1}{7}\right) + \cdots + 4\left(\frac{1}{4n-2} - \frac{1}{4n-1}\right),$$

it is clear that $f(n)$ is an increasing function of n. In the following, we will study the error function $\pi - f(n)$, which is positive, decreasing, and tends to 0 as $n \to \infty$. To develop some intuition, we compute the error $\pi - f(n)$ for a few values of n:

$$\begin{aligned} \pi - f(10^5) &= 5.000000031341045 \times 10^{-6} \\ \pi - f(10^6) &= 5.000000689037165 \times 10^{-7} \\ \pi - f(10^7) &= 4.999997615939833 \times 10^{-8}. \end{aligned} \tag{2}$$

2.2. **Asymptotic series.** Observe in (2) that when n increases by a factor of 10, the error $\pi - f(n)$ decreases by approximately a factor of 10. This reason for this consistent decrease is that $f(n)$ has an asymptotic series.

Definition 2.1 (Asymptotic series)**.** An asymptotic series (for the function $f(n)$ in terms of powers of $1/n$ as $n \to \infty$) is a formal series

$$f(n) = a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots,$$

that satisfies the follow condition: for each fixed integer $m \geq 0$ we have:

$$f(n) = \pi + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots + \frac{a_m}{n^m} + \mathcal{O}\left(\frac{1}{n^{m+1}}\right), \quad \text{as} \quad n \to \infty, \tag{3}$$

where the constant in the term $\mathcal{O}(1/n^{m+1})$ is allowed to depend on m.

Remark 2.1. The key aspect of this definition is that asymptotic series do not need to be convergent: they are formal series whose partial sums obey (3). There are many non-convergent asymptotic series that are useful for numerical methods.

It turns out, that the function $f(n)$ defined in (1) has an asymptotic series

$$f(n) = \pi + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots,$$

see Lemma 4.1; this explains why the error $\pi - f(n)$ decreases approximately linearly with respect to n (up to higher order effects). How can we use the existence of this asymptotic series to compute π more accurately?

2.3. **Richardson extrapolation.** There is a simple trick called Richardson extrapolation that can be used to improve approximations when asymptotic series exist. Assume that $f(n)$ has the asymptotic series

$$f(n) = \pi + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots.$$

Then,

$$f(2n) = \pi + \frac{a_1}{2n} + \frac{a_2}{4n^2} + \frac{a_3}{8n^3} + \cdots,$$

and it follows that

$$g(n) := 2f(2n) - f(n) = \pi + \frac{b_2}{n^2} + \frac{b_3}{n^3} + \cdots, \tag{4}$$

for some coefficients b_k. Observe that the a_1/n term was canceled out such that

$$g(n) = \pi + \mathcal{O}(1/n^2).$$

More concretely, if we set $n = 10^3$ and compute $g(n)$ we find

$$|\pi - g(10^3)| = 2.342570581959080 \times 10^{-11},$$

which is less than the error $|\pi - f(10^6)| \approx 5 \times 10^{-7}$ reported in (2). The error $|\pi - g(10^3)|$ is even less than we might have expected: if $f(n) = \pi + \mathcal{O}(1/n)$ and $g(n) = \pi + \mathcal{O}(1/n^2)$ then we might expect $|\pi - f(10^6)| \approx |\pi - g(10^3)|$. One possible explanation is that the constant in the error term associated with g is much smaller than that of f, but there turns out to be another reason: the asymptotic series for $f(n)$ in this example only has nonzero coefficients for odd powers of n (in particular $a_2 = 0$) so in fact $g(n) = \pi + \mathcal{O}(1/n^3)$.

Exercise. Implement and evaluate functions for (1) and (4) in your favorite programming language to develop some intuition about Richardson extrapolation.

3. Simple iteration, computing $\sqrt{2}$, and contraction mapping

3.1. **Simple iteration.** The verb *itero* means 'repeat, do again'. Iterative numerical methods involve repeating a procedure until the desired result is (hopefully) achieved. The simplest example of an iterative numerical method is simple iteration: given a function $f : \mathbb{R} \to \mathbb{R}$ and an initial point $x_0 \in \mathbb{R}$ we define

$$x_n = f(x_{n-1}), \tag{5}$$

for $n = 1, 2, \ldots$. When f is continuous and $x_n \to \alpha \in \mathbb{R}$, it follows that

$$f(\alpha) = \alpha,$$

that is, the simple iteration converges to a fixed point α of the function f. In the following section (§3.2), we present an example where a simple iteration converges to $\sqrt{2}$. Afterwards, in §3.3 we develop geometric intuition about when simple iterations converge, and then in §3.4 discuss contraction mapping, which is an important idea on which many iterative numerical methods are based.

3.2. **Computing $\sqrt{2}$.** Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right). \tag{6}$$

Observe that

$$f(\sqrt{2}) = \sqrt{2}.$$

Starting with $x_0 = 1$, we perform the simple iteration (5) and report the relative error $|x_k - \sqrt{2}|/\sqrt{2}$ for each iteration k.

k	$\lvert x_k - \sqrt{2}\rvert/\sqrt{2}$
1	2.928932×10^{-1}
2	6.066017×10^{-2}
3	1.734607×10^{-3}
4	1.501825×10^{-6}
5	1.127640×10^{-12}
6	1.570092×10^{-16}
7	1.570092×10^{-16}

FIGURE 1. The iteration (5) for $f(x) = \frac{1}{2}(x + 2/x)$ with $x_0 = 1$.

Observe that the relative error $|x_k - \sqrt{2}|/\sqrt{2}$ decreases rapidly until $k = 6$, where it stays around 10^{-16}. In theory, (see §3.4), the error should continue to decrease rapidly, but in practice, since these numerical calculations were performed on a computer using Double Precision numbers (the most common digital format for representing real numbers), the fact that the error stagnates around 10^{-16} is expected, and nothing to worry about. Roughly speaking, Double Precision numbers have sixteen digits of relative accuracy, that is, $x \in \mathbb{R}$ will be encoded as $\tilde{x} = x(1 + \gamma)$, for some $|\gamma| < 10^{-16}$. The topic of Double Precision numbers is very important and will be discussed separately; for now we ignore this issue and consider the theoretical question of when simple iterations converge.

Exercise. Implement (5) for (6) in your favorite programming language; compute the relative error at each iteration.

3.3. **Contraction mapping (intuition).** To develop some geometric intuition we visualize a few examples. In Figure 2, we plot two linear functions f_1 and f_2 with slops 0.7 and 1.3, respectively, as solid lines and plot the identity map $x \mapsto x$ as a dashed line. We represent the iteration as a dotted path that alternates between the function and the identity map.

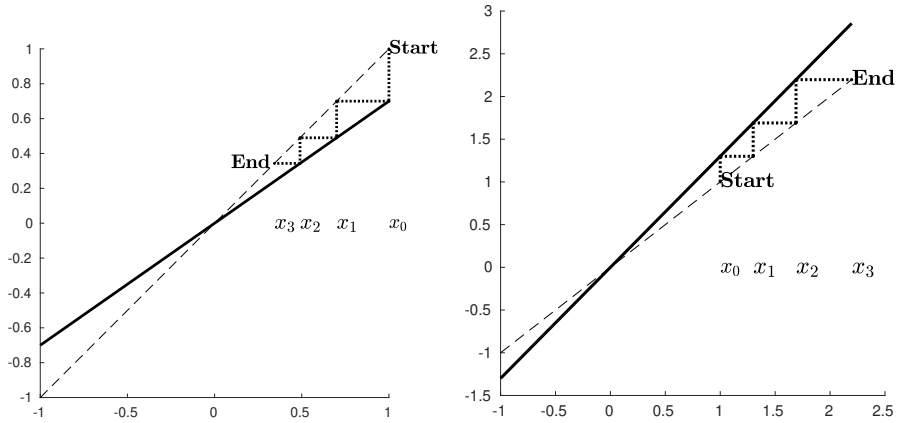

FIGURE 2. The iteration (5) for two linear functions with positive slope.

The sequence of points $x_0, x_1, x_2, x_3, \ldots$ approaches the fixed point for the function f_1, and diverges for the function f_2. Next we consider the iteration for two linear functions f_3 and f_4 with negative slopes -0.7 and -1.3, respectively, in Figure 3.

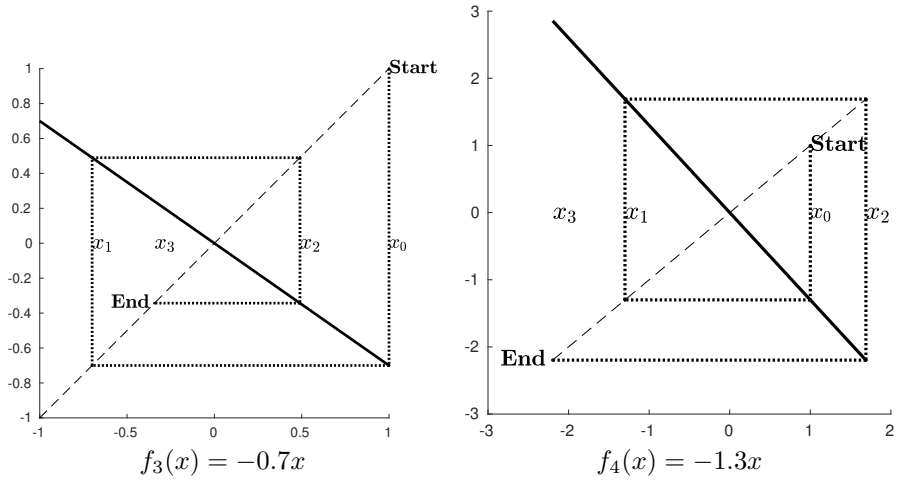

FIGURE 3. The iteration (5) for two linear functions with negative slope.

The path representing the iteration spirals inward for f_3 (the sequence $x_0, x_1, x_2, x_3, \ldots$ approaches the fixed point), and spirals outward for f_4 (the sequence diverges). In summary, these figures are a proof-by-picture that simple iteration on a linear functions f converges when $|f'| < 1$ and diverges when $|f'| > 1$ (unless $f(x_0) = x_0$).

3.4. **Contraction mapping (theory).** The following result turns the intuition developed in the previous section into a theorem; informally speaking, this theorem says that if $|f'| < 1$ then simple iteration on f converges to a fixed point, but to avoid assuming f is differentiable the result is stated using the more general condition $|f(x) - f(y)|/|x - y| < 1$.

Theorem 3.1 (Contraction mapping theorem)**.** *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a contraction, that is,*

$$|f(x) - f(y)| \leq c|x - y|, \quad \textit{for all} \quad x, y \in \mathbb{R},$$

for a fixed constant $0 < c < 1$. Then, f has a unique fixed point, and the iteration (5) *converges to this fixed point from any starting point.*

We give the proof which provides information about the convergence rate.

Proof of Theorem 3.1. Let $\varepsilon > 0$ be given. If $n_2 > n_1 \geq n$, then by writing x_{n_1} and x_{n_2} as telescopic series we have

$$|x_{n_1} - x_{n_2}| = \left| \sum_{k=n_1}^{n_2-1} (x_{k+1} - x_k) \right| \leq \sum_{k=n}^{\infty} |x_{k+1} - x_k|.$$

By the definition of the iteration (5) and the fact that f is a contraction we have

$$|x_{k+1} - x_k| = |f(x_k) - f(x_{k-1}| \leq c|x_k - x_{k-1}|.$$

Thus,

$$\sum_{k=n}^{\infty} |x_{k+1} - x_k| \leq \sum_{k=n}^{\infty} c^k |x_0 - x_1| \leq \frac{c^n}{1-c} |x_0 - x_1|.$$

It follows that $x_n \to \alpha$ as $n \to \infty$ for some $\alpha \in \mathbb{R}$, and moreover, that $|x_n - \alpha| \leq Cc^n$, where $C = |f(x_0) - x_0|/(1 - c)$. The fact that $C = 0$ if x_0 is a fixed point implies that the fixed point α is unique. □

Remark 3.1 (Iterations for a given accuracy). To ensure $|x_n - \alpha| \leq \varepsilon$ it suffices to choose n such that $\varepsilon \leq Cc^n$. Thus, any $n \geq (\ln(\varepsilon) - \ln(C))/\ln(c))$ is sufficient.

3.5. **Analyzing our example of computing $\sqrt{2}$.** Recall that when computing $\sqrt{2}$ we performed simple iteration on the function

$$f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right),$$

which has a fixed point $f(\sqrt{2}) = \sqrt{2}$, see Figure 4.

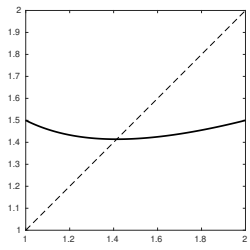

FIGURE 4. The function $f(x) = \frac{1}{2}(x + 2/x)$ (solid) and $x \mapsto x$ (dashed).

The rate of convergence of simple iteration depends (by the proof of the theorem) on the magnitude of the slope of the function near the fixed point. In Figure 4, the derivative f' appears to vanish at the fixed point. We can verify this by computing

$$f'(x) = \frac{1}{2}\left(1 - \frac{2}{x^2}\right),$$

and observing that $f'(\sqrt{2}) = 0$. Thus, as we get closer to the fixed point, the rate of convergence increases since the slope becomes closer the zero. By using the error analysis of the theorem iteratively it is possible to show that for this example if the relative error is currently δ, then after another iteration the error will be roughly δ^2 (in fact, this iteration is an example of Newton's Method, which is a iterative method designed to achieve this type of so called quadratic convergence). Observe that in Figure 1 the error is indeed roughly squared in each iteration (until we encounter numerical issues when the relative error reaches 10^{-16}).

4. Technical details

To complete this note, we give a short direct proof that $f(n)$ defined in (1) has an asymptotic series with at least two terms; the argument is straightforward to extend to show that $f(n)$ has a full asymptotic series.

Lemma 4.1. *If $f(n)$ is defined by (1), then,*

$$f(n) = \pi - \frac{1}{2n} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

Proof.

$$\frac{\pi}{4} = \arctan(1) = \int_0^1 \frac{1}{1+x^2}dx.$$

Recall that

$$\frac{1}{1-t} = \sum_{k=0}^{n} t^k + \frac{t^{n+1}}{1-t}.$$

Thus,

$$\frac{1}{1+x^2} = \sum_{k=0}^{n}(-1)^k x^{2k} + \frac{(-1)^{n+1}x^{2n+2}}{1+x^2}.$$

Integrating x from 0 to 1 gives

$$\frac{\pi}{4} = \sum_{k=0}^{n}\frac{(-1)^k}{2k+1} + (-1)^{n+1}\int_0^1 \frac{x^{2n+2}}{1+x^2}dx.$$

Integrating the integral on the right hand side by parts gives

$$\int_0^1 \frac{x^{2n+2}}{1+x^2}dx = \frac{(-1)^{n+1}}{4n+6} - \int_0^1 \frac{x^{2n+3}}{2n+3}\frac{-2x}{(1+x^2)^2}dx.$$

We have

$$\left|\int_0^1 \frac{x^{2n+3}}{2n+3}\frac{-2x}{(1+x^2)^2}dx\right| \leq \frac{2}{2n+3}\int_0^1 x^{2n+3}dx = \mathcal{O}\left(\frac{1}{n^2}\right).$$

We conclude that

$$\frac{\pi}{4} = \sum_{k=0}^{2n-1}\frac{(-1)^k}{2k+1} + \frac{1}{8n} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

□