Composite Polynomial Approximations to the Sign and Square Root Functions

Part C Dissertation submitted in partial fulfilment of the requirements for the degree of Master of Mathematics

Candidate Number: 1013048

Trinity Term 2020

To Ray, Toni, and Sam.

Abstract

Approximation theory is ubiquitous in scientific computing, in particular in the computation of matrix functions. For instance, the matrix sign function $\operatorname{sgn}(A)$ and matrix roots $A^{1/p}$ can be approximated by matrix iterations such as Newton's Method. Efforts have been made to improve the efficiency of Newton's Method in modern mathematical research by considering scaled iterative processes.

Matrix iterations can effectively be viewed as approximating a matrix function $f(A)$ by a composite rational function of A. Polynomials are often overlooked for their comparatively weaker convergence properties, but at a matrix argument they are much cheaper to compute than rational functions, since matrix polynomials are inversion-free.

Composing low-degree polynomials or rational functions is an effective way of rapidly producing such functions of much higher degree, and a remarkable result in rational approximation theory states that appropriately composing Zolotarev functions generates higher-order Zolotarev functions. This dissertation investigates the polynomial analogue by constructing a composite polynomial approximation to the sign function based on a greedy algorithm. We show that our construction is equivalent to a scaled Newton-Schulz iteration, and analyse its convergence with respect to the degrees of freedom of the approximation. Moreover, we show that the iteration can be used to obtain a novel composite polynomial approximation to the square root function.

Contents

1 Introduction **1**

2 Preliminary discussion **4**
- 2.1 Composite polynomials and degrees of freedom . . . 4
- 2.2 Best approximation and equioscillation . . . 5
  - 2.2.1 Rational best approximation and equioscillation . . . 7
- 2.3 Matrix iterations and Newton's Method . . . 8
  - 2.3.1 Newton-Schulz iterations and the matrix sign function . . . 10
  - 2.3.2 Improved Newton's Method . . . 12

3 Composite polynomial approximation to sgn(x) **16**
- 3.1 Zolotarev functions . . . 16
- 3.2 Greedy approximation to sgn(x) in $\mathcal{P}^{\text{comp}}_{(k,3)}$ . . . 18
- 3.3 Equioscillation behaviour . . . 20
- 3.4 Convergence analysis of the scaled Newton-Schulz iterates . . . 22
  - 3.4.1 Deriving a lower bound for the number of iterations . . . 24
  - 3.4.2 Convergence with respect to degrees of freedom . . . 27

4 Composite polynomial approximation to $\sqrt{x}$ **29**
- 4.1 Zolotarev functions and matrix roots revisited . . . 29
- 4.2 Deriving an approximation from scaled Newton-Schulz iterates . . . 31
- 4.3 Observations of the scaled Alternative Newton iterates . . . 32

5 Conclusion **36**
- 5.1 Where does this project lead? . . . 36

Appendices **38**

A Results for composite rational functions **38**
- A.1 Composite rational functions and degrees of freedom . . . 38
- A.2 Composite rational approximation to $\sqrt{x}$ . . . 39
  - A.2.1 Bounding the error on $[0, \alpha^2]$ . . . 40

A.2.2 Proof of Theorem A.4 42

References

Chapter 1

Introduction

Composite approximations of functions appear in a range of modern mathematical literature, with a prime example being in the computation of matrix functions. The study of matrix functions has become a growing area of research in recent years for its far-reaching applications in scientific computing, engineering and finance. For instance, there are many frameworks in which it is desirable to compute the pth root X of a stochastic matrix A, whose entries might encode the probabilities that a company moves from one credit rating to another over a given time period, or the progression of chronic diseases in patients at different stages in time [12]. In this case, the matrix root will not only satisfy $X^p = A$, but have entries corresponding to a shorter time period. For example, if A represents an annual transition matrix and $p = 2$, then X is a six-month transition matrix—a *matrix square root* of A[1].

Numerical methods for computing functions of matrices are well-established and plentiful; such methods are discussed in depth in [9]. In particular, using matrix iterations $X_{k+1} = g(X_k)$, typically where g is a polynomial or rational function, results in composite polynomial or rational approximations to a matrix function $f(A)$ in terms of an initial guess X_0. Matrix iterations are commonly used for the computation of matrix roots, in addition to the *matrix sign function* [17, 26, 27], and the unitary matrix arising from *polar decomposition* [19]. A material example is the Newton iteration for the matrix square root [10]

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A), \qquad X_0 = A, \tag{1.1}$$

which computes the root of a nonsingular, diagonalisable matrix $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on the closed negative real axis[2]. The limit of (1.1) is known as the *principal square root*, denoted by $A^{1/2}$. For this example, it is easy to show

[1] Further applications of matrix functions are discussed in [13, 15].

[2] Such a square root exists and is unique: see [9, Theorem 1.29].

that the iterates X_k are also diagonalisable, and if $D_k = \mathrm{diag}(\lambda_k^{(1)}, \dots, \lambda_k^{(n)})$ is the diagonal matrix arising from the eigendecomposition for X_k, then

$$\lambda_{k+1}^{(j)} = \frac{1}{2}\left(\lambda_k^{(j)} + \frac{\lambda_0^{(j)}}{\lambda_k^{(j)}}\right), \qquad j = 1, \dots, n.$$

That is, the Newton iteration (1.1) can be broken down into n independent scalar Newton iterations for computing the square roots of $\lambda_0^{(j)}$, namely the eigenvalues of A. Accordingly, we might consider using successive approximations of functions $f_{k+1} = g(f_k)$ on some interval containing $\Lambda(A)$, the spectrum of A, as a basis for determining matrix iterations $X_{k+1} = g(X_k)$. Indeed, the scalar Newton iterations above converge if the functions f_k defined by $f_0(x) = x$ and

$$f_{k+1} = g(x, f_k(x)), \qquad g(x, y) = \frac{1}{2}\left(y + y^{-1}x\right) \tag{1.2}$$

converge to $\sqrt{x}$ on $\Lambda(A)$, and we note that the matrix iteration (1.1) is given by $X_{k+1} = g(A, X_k)$.

Using composite rational functions such as (1.2) to compute matrix functions is the approach of much of the modern literature [6, 7, 20]; often polynomials are dismissed for their poor comparative performance at a scalar level. However, to evaluate rational functions at a matrix argument we must compute matrix inverses, either directly or by a matrix decomposition (such as the Cholesky factorisation $A = LL^*$ for positive-definite Hermitian matrices [18]), which is expensive in comparison to polynomial evaluations which only require matrix multiplications.

The goal of this dissertation is to construct and analyse composite *polynomial* approximations to scalar functions, with the discussion above serving as motivation for doing so. Whilst it is not always the case that results for scalar iterations hold for the matrix counterpart (see [9, Section 4.9]), there are still many interesting questions concerning composite polynomials that can be asked. For example, we know that every continuous function f on a closed, bounded interval has a unique best approximation in every class of polynomials or rational functions, and that the best approximation is characterised by an equioscillation property in the error function [29]. Whilst it is ambitious to believe that composite polynomials (or rational functions) are best approximations for their degree[3], one question we could ask is how quickly they converge with respect to their *degrees of freedom*—the number of parameters required to express them. Composing polynomials or rational functions is an effective method for generating such functions of much

[3] Surprisingly, there is a class of rational functions for which this property holds, called the *Zolotarev functions*. We discuss these in Chapter 3.

higher degree with respect to their degrees of freedom[4], so much so that we can construct a composite rational approximation to $\sqrt{x}$ for which the convergence is *doubly exponential*[5] with respect to the degrees of freedom [8]. In this dissertation, we consider such questions in the polynomial setting.

Outline

This dissertation is structured as follows. Chapter 2 is a preliminary discussion in which we review fundamental results from approximation theory and discuss variants of Newton's Method used for approximating the sign and square root functions. In Chapter 3, we introduce the Zolotarev functions, and construct a composite polynomial approximation to the sign function inspired by the recursive optimality property of the Zolotarev functions. We analyse the convergence of our construction, and make comparisons with the minimax approximation with respect to the degrees of freedom. Finally, we use the sign function approximation in Chapter 4 to obtain a composite polynomial approximation to $\sqrt{x}$. The appendix contains results concerning rational functions, included to provide a comparison to the polynomial results.

Notation

In this dissertation, we shall use the following notation without further discussion:

- $\mathbb{N} = \{0, 1, 2, \dots\}$;
- $C(I)$ denotes the set of continuous functions on an interval $I \subset \mathbb{R}$;
- $\|\cdot\|_{\infty,I}$ denotes the ∞-norm over I, defined by $\|f\|_{\infty,I} = \max_{x\in I} |f(x)|$;
- $\mathcal{P}_n(I)$ denotes the set of polynomials of degree at most n on I;
- $\mathcal{R}_{m,n}(I)$ denotes the set of rational functions of type (m, n) on I, namely

$$\mathcal{R}_{m,n}(I) = \{p/q : p \in \mathcal{P}_m(I),\, q \in \mathcal{P}_n(I)\}.$$

We omit I in notation when the interval is clear, and write e.g. $\mathcal{P}_m$, $\mathcal{R}_{m,n}$, $\|\cdot\|_\infty$.

[4] For example, an inductive argument shows that the f_k in (1.2) are rational functions of type $(2^{k-1}, 2^{k-1} - 1)$, yet have at most $4k$ degrees of freedom.

[5] By doubly exponential, we mean that an approximation with d degrees of freedom has maximum uniform error $O(\exp(-C_1 \exp(C_2 d)))$ for some constants $C_i > 0$.

Chapter 2

Preliminary discussion

In this chapter, we review the equioscillation property of best approximations and discuss composite approximations $f_{k+1} = g(f_k)$ used for computing the sign and square root functions. In particular, we introduce the *Newton-Schulz* iteration for approximating $\operatorname{sgn}(x)$, for which the iteration function g is a polynomial. These ideas will be important in later chapters, as we will construct scaled composite approximations $f_{k+1} = g_{k+1}(f_k)$, assessing their quality by comparing them to the corresponding unscaled approximations, and checking whether or not they are best approximations by counting the number of equioscillation points in the error. Before we proceed, let us make the definitions of composite polynomials and degrees of freedom precise.

2.1 Composite polynomials and degrees of freedom

Definition 2.1. A bivariate polynomial $p(x, y)$ is said to be of degree m if $p(x, x)$ is of degree m. A polynomial $p(x)$ is said to be (k, m)-*composite* if

$$p(x) = p_k(x, p_{k-1}(x, p_{k-2}(\cdots(x, p_1(x, 1))))) \tag{2.1}$$

for bivariate polynomials $p_i(x, y)$, $i = 1, \ldots, k$, each of degree m. We say that a (k, m)-composite polynomial is *pure* if the $p_i(x, y)$ in (2.1) are univariate, so that

$$p(x) = p_k(p_{k-1}(\cdots(p_1(x)))).$$

We will denote the space of pure (k, m)-composite polynomials by

$$\mathcal{P}^{\text{comp}}_{(k,m)} = \{p_k \circ \cdots \circ p_1 : p_i \in \mathcal{P}_m\}.$$

Clearly $\mathcal{P}^{\text{comp}}_{(k,m)} \subset \mathcal{P}_{m^k}$, but equality doesn't hold: see [22].

The introduction of bivariate polynomials seems unnecessarily complicated, but is essential to allow the iteration function g to depend on x as well as the previous iterate[1]. Many iterations we consider will take the form $f_{k+1}(x) = g(x, f_k(x))$.

Definition 2.2. The *degrees of freedom* of a polynomial $p_n \in \mathcal{P}_n$ is the number of parameters required to completely determine p_n. We say that p_n is a *plain polynomial* if p_n has $n+1$ degrees of freedom, that is, $p_n(x) = \sum_{k=0}^{n} \alpha_k x^k$ where $\{\alpha_k\}_{k=0}^{n}$ is a set of independent, non-zero scalars.

We observe that composing low-degree polynomials is an efficient way to generate polynomials of much higher degree, with respect to the degrees of freedom.

Example 2.3. Consider two plain polynomials $p_m \in \mathcal{P}_m$, $p_n \in \mathcal{P}_n$. Then $p_m \circ p_n$ has degree mn, whilst having only $m+n+2$ degrees of freedom. Moreover, a polynomial $p \in \mathcal{P}_{(k,m)}^{\text{comp}}$ formed by composing k plain polynomials of degree m has degree m^k, yet only $k(m+1)$ degrees of freedom.

2.2 Best approximation and equioscillation

To analyse how well a polynomial $p \in \mathcal{P}_n$ is approximating a function $f \in C(I)$, where I is a closed, bounded interval, we can consider for $p_n \in \mathcal{P}_n$ the error

$$\|(f - p_n)w\|_{\infty,I}\,, \tag{2.2}$$

where $w \in C(I)$ is a positive *weight function*. In the best-case scenario, p will minimise (2.2) over all $p_n \in \mathcal{P}_n$. Such an approximation is known as the *best approximation*, or *minimax*, to f with respect to w. Common weights include $w(x) \equiv 1$ (corresponding to the uniform error) and $w(x) = 1/f(x)$ (corresponding to the relative error, assuming that $f(x) > 0$ on I). The best approximation exhibits a remarkable equioscillation property, as seen in e.g. [1, Chapter II].

Definition 2.4. A function $f \in C(I)$ on a closed, bounded interval I *equioscillates* between k extreme points if there exist $\{x_i\}_{i=1}^{k} \subset I$ such that $x_1 < x_2 < \cdots < x_k$ and for some $j \in \{0, 1\}$,

$$f(x_i) = (-1)^{i+j} \|f\|_{\infty,I}\,, \qquad i = 1, \ldots, k.$$

Theorem 2.5 (Equioscillation Theorem)**.** *Let $f \in C(I)$ be a real function on a closed, bounded interval I. For every $n \in \mathbb{N}$ and weight function $w \in C(I)$, there is a unique best approximation $p^* \in \mathcal{P}_n$ to f with respect to w. Moreover, if $p \in \mathcal{P}_n$, then $p = p^*$ if and only if $(f - p)w$ equioscillates between at least $n+2$ extrema.*

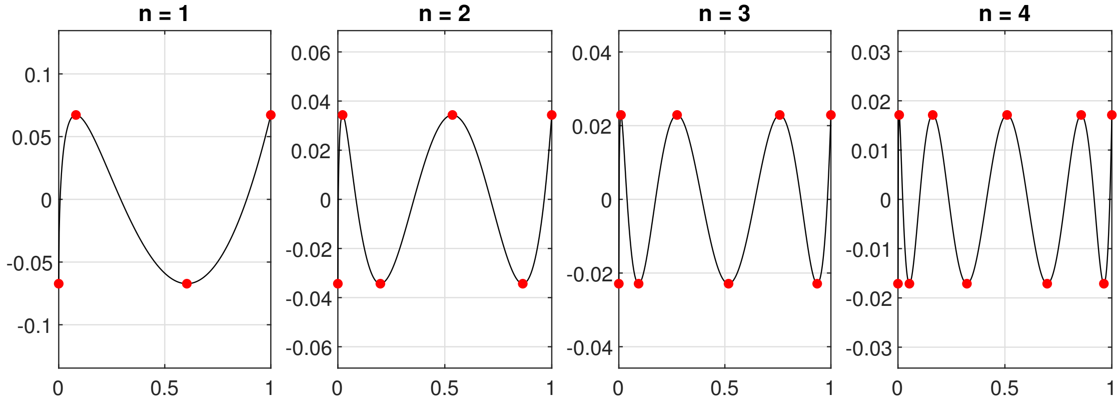

Figure 2.1: The equioscillation property demonstrated: error $E_{2n}(x) = \sqrt{x} - p^*_{2n}(x)$ in the uniform best approximations $p^*_{2n} \in \mathcal{P}_{2n}$ to $\sqrt{x}$ on $[0,1]$.

In [19, Lemma 2], Nakatsukasa and Freund provide an equioscillation result for *rational* best approximations on a disjoint union of closed, bounded intervals in the uniform sense. In Theorem 2.9, we will generalise and prove their result to include errors equipped with a positive weight function, allowing for the same result in the relative sense, for example. We highlight the result in the case of polynomials, since it will be crucial in our construction of a composite polynomial approximation to the sign function in the next chapter.

Corollary 2.6. *Suppose that $f \in C(I \cup J)$ is a real function on disjoint, closed, bounded intervals I, J. Then $p \in \mathcal{P}_n$ is the unique best approximation to f with respect to w if $(f - p)w$ equioscillates between at least $n + 3$ extrema.*

One final result on best approximations, mentioned briefly in [5, p. 2] as a remark, is the following. We present it as a lemma with a full proof.

Lemma 2.7. *Let $n \in \mathbb{N}$, and suppose $f \in C(I)$ is an odd function on an interval I symmetric about the origin. Then the best uniform approximation $p^* \in \mathcal{P}_{2n}$ is an odd function. Hence $p^*(x) = xh(x^2)$ for some $h \in \mathcal{P}_{n-1}$.*

Proof. Define $q(x) = (p^*(x) - p^*(-x))/2$. Then $q(-x) = -q(-x)$, so q is odd. As a result, if we write $q(x) = \sum_{k=0}^{2n} \alpha_k x^k$ for some $\alpha_k \in \mathbb{R}$ and compare the coefficients of the equation $q(x) = -q(-x)$, we find that $\alpha_i = 0$ for $i = 0, 2, \dots, 2n$. Hence $q(x) = \sum_{k=0}^{n-1} \alpha_{2k+1} x^{2k+1}$, so we can write $q(x) = xh(x^2)$ where

$$h(x) = \sum_{k=0}^{n-1} \alpha_{2k+1} x^k \in \mathcal{P}_{n-1}.$$

[1] Dependence on x was crucial for the square root iteration (1.2) in Chapter 1, for example.

We show that $p^* = q$. To see this, we use that f is an odd function and I is symmetric about the origin, so that $x \in I$ if and only if $-x \in I$, to obtain

$$|q(x) - f(x)| = \left| \frac{p^*(x) - f(x)}{2} - \frac{p^*(-x) - f(-x)}{2} \right| \leqslant \|p^* - f\|_{\infty,I} ,$$

by the triangle inequality. Hence $\|q - f\|_{\infty,I} \leqslant \|p - f\|_{\infty,I}$ for all $p \in \mathcal{P}_{2n}$. By uniqueness of the best approximation, we must have $p^* = q$. □

Remark. In the context of Lemma 2.7, the maximum uniform error of p^* on $I^+ = \{x \in I : x \geqslant 0\}$ will be identical to that on $I^- = \{x \in I : x \leqslant 0\}$ by symmetry, so the problem of finding the best approximation $p^* \in \mathcal{P}_{2n}$ to f on I is equivalent to finding the best odd approximation on $f|_{I^+}$, and such an approximation will be characterised by $f - p^*$ having at least $n+1$ equioscillation points in I^+.

2.2.1 Rational best approximation and equioscillation

We occasionally discuss rational functions in this dissertation, and use the results of composite rational functions[2] as motivation for considering similar composite polynomial approximations. As such, we shall briefly discuss the analogues of this section for rational functions.

Suppose we approximate a continuous function $f \in C(I)$ with a rational function $r_{m,n} \in \mathcal{R}_{m,n}(I)$. A *rational best approximation* to f with respect to a positive weight function w minimises an error of the form

$$\|(f - r)w\|_{\infty,I}$$

over all $r \in \mathcal{R}_{m,n}(I)$. A characterisation of the best approximations in terms of an equioscillating error curve also exists for rational functions [1, Chapter II].

Theorem 2.8 (Equioscillation theorem for rational functions). *Let $f \in C(I)$ be a real function on a closed, bounded interval I. For every $m, n \in \mathbb{N}$ and weight function $w \in C(I)$, there is a unique rational best approximation $r^* \in \mathcal{R}_{m,n}$ to f with respect to w. Moreover, if $r = P/Q$ for $P \in \mathcal{P}_m$, $Q \in \mathcal{P}_n$, then $r = r^*$ if and only if $(f - r)w$ equioscillates between at least $m + n + 2 - d_r$ extrema, where*

$$d_r = \min\{m - deg P, n - deg Q\}$$

is the defect *of r.*

[2] Analogous to Definition 2.1, a composite rational function is defined using bivariate rational functions, which we include in Appendix A. This is the same definition as used in [8].

Finally, we include the equioscillation result of Nakatsukasa and Freund, which we have generalised to include a weight function. This will later be used to verify the optimality of rational approximations to the sign function.

Theorem 2.9. *Let $m, n \in \mathbb{N}$, $r \in \mathcal{R}_{m,n}$ and $f \in C(I \cup J)$, where I, J are disjoint closed, bounded intervals. If $(f - r)w$ equioscillates between at least $m + n + 3$ extrema, then r is the (unique) rational best approximation to f in $\mathcal{R}_{m,n}$ with respect to w.*

Proof. By contradiction. If $(f - r)w$ equioscillates between $m + n + 3$ extrema $\{x_i\}_{i=1}^{m+n+3}$ but there exists $r' \in \mathcal{R}_{m,n}$ such that

$$\|(f - r')w\|_{\infty, I \cup J} < \|(f - r)w\|_{\infty, I \cup J},$$

then $(r - r')w$ alternates in sign at the x_i, hence $(r - r')w$ vanishes at a minimum of $m + n + 1$ points (this would be $m + n + 2$ points if we were not considering a disjoint union). Since $w > 0$ and $r - r' \in \mathcal{R}_{m+n,2n}$, we must have $r - r' = 0$, a contradiction. □

2.3 Matrix iterations and Newton's Method

In the introduction, we saw how matrix iterations can be interpreted as composite polynomial or rational approximations to a matrix function $f(A)$ in terms of the initial guess X_0 (usually $X_0 = A$). When deriving matrix iterations, it is common to first apply Newton's Method to an algebraic equation which has $f(A)$ as a solution, then choose an initial guess X_0 that accelerates the algorithm. Recall Newton's Method for finding a root x^* of $f(x) = 0$, where f is a differentiable scalar function: for an initial guess x_0, we define the sequence

$$x_{k+1} = g(x_k) := x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \dots, \tag{2.3}$$

provided that $f'(x_k) \neq 0$ for all k. Newton's Method converges if x_0 is sufficiently close to x^*, in this case converging quadratically [28, Theorem 1.8]. Due to the division by $f'(x_k)$ in (2.3), we often find that scalar Newton iterations contain inverses of the previous iterate. This is unfavourable if the underlying iteration function g is then used to derive matrix iterations $X_{k+1} = g(X_k)$, since each iteration requires a new inverse to be computed, such as in the Newton iteration for the matrix square root (1.1).

The following example demonstrates this pitfall, and shows how considering a different algebraic equation can result in a sequence that doesn't require a new inverse to be computed at each iteration.

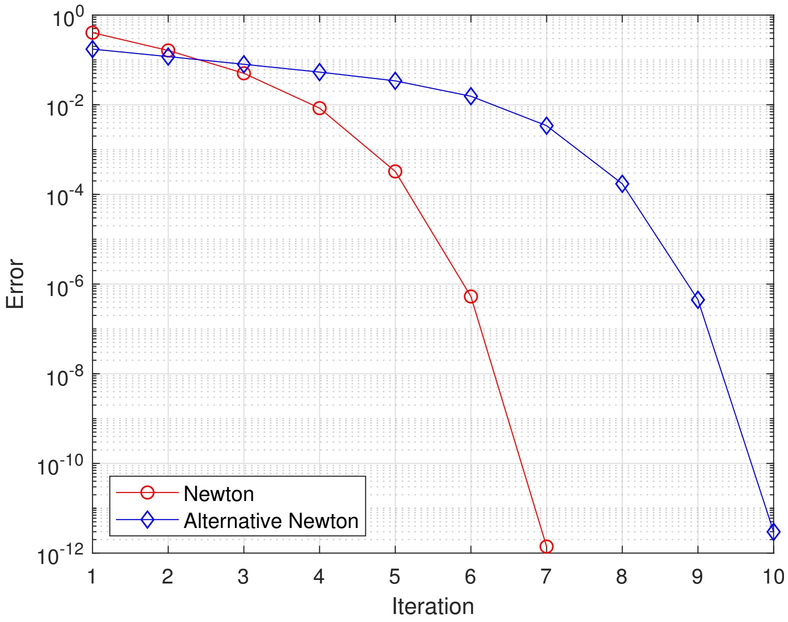

Figure 2.2: Comparison of the maximum absolute uniform error in the standard Newton (1.2) and Alternative Newton (2.5) iterates to $\sqrt{x}$ on $[\delta^2, 1]$, where $\delta = 0.1$.

Example 2.10. To obtain an approximation to $\sqrt{x}$ on some interval I, we can fix $\lambda \in I$ and apply Newton's Method to functions for which $\sqrt{\lambda}$ is a solution. This generates iterates $x_k(\lambda) \approx \sqrt{\lambda}$ which can then be generalised to a function over the whole interval by defining $f_k(\lambda) := x_k(\lambda)$. Applying Newton's Method to the equation $f(x) = x^2 - \lambda$, we obtain the iteration

$$f_{k+1}(x) = \frac{1}{2}\left(f_k(x) + \frac{x}{f_k(x)}\right) =: g(x, f_k(x)), \tag{2.4}$$

which is the scalar Newton iteration seen in (1.2). Alternatively, we could apply Newton's Method to $f(x) = 1 - \lambda/x^2$, which also has $\sqrt{\lambda}$ as a root. This gives

$$f_{k+1}(x) = \frac{f_k(x)}{2}\left(3 - \frac{f_k(x)^2}{x}\right) =: \tilde{g}(x, f_k(x)). \tag{2.5}$$

Remark. While the iteration function $\tilde{g}$ in (2.5) is not a polynomial, the iterative process never divides by the previous iterate. Furthermore, it is clear that any initial guess f_0 that is a multiple of x will produce polynomial approximations. We will refer to the iterations (2.5), with $f_0(x) = x$, as *Alternative Newton* iterates.

Figure 2.2 compares the maximum uniform error in the Newton and Alternative Newton approximations to $\sqrt{x}$. Clearly, the alternative iterates underperform the standard iterates, which is to be expected since we are comparing the convergence of polynomial approximations to rational approximations.

2.3.1 Newton-Schulz iterations and the matrix sign function

The conventional approach for removing inverses in matrix iterations of the form $X_{k+1} = g(X_k)$, where g is a rational function, is to replace all instances of X_k^{-1} with inversion-free approximations to them. This can be achieved by considering Newton's Method for the inverse of an invertible matrix B, given by

$$Y_{k+1} = Y_k(2I - BY_k), \tag{2.6}$$

as proposed by Schulz [25]. Computing one step of Newton's Method for X_k^{-1} by taking $Y_k = B = X_k$ in (2.6), we obtain the approximation $X_k^{-1} \approx X_k(2I - X_k^2)$. Newton iterations that approximate appearances of inverses in this way are referred to as *Newton-Schulz* iterations.

Let us find the Newton-Schulz iteration for the *matrix sign function*. Analogous to the scalar sign function $\text{sgn}(z) = z/\sqrt{z^2}$, where $\text{Re}(z) \neq 0$, we define[3] the sign of a matrix $A \in \mathbb{C}^{n\times n}$ with eigenvalues away from the imaginary axis by

$$\text{sgn}(A) = A(A^2)^{-1/2},$$

where $B^{-1/2}$ denotes the inverse of the principal square root of B. In particular, this is well-defined since A having no eigenvalues on the imaginary axis implies that A^2 has no eigenvalues on the closed negative real axis. In [23, Section 1.3], Roberts shows that the iteration

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \qquad X_0 = A, \tag{2.7}$$

converges to $\text{sgn}(A)$. The iteration is precisely Newton's Method applied to the equation $X^2 = I$, for which $\text{sgn}(A)$ is clearly a solution, hence convergence is quadratic. Replacing X_k^{-1} in (2.7) with the approximation $X_k(2I - X_k^2)$, we obtain the Newton-Schulz iteration

$$X_{k+1} = \frac{X_k}{2}(3I - X_k^2), \qquad X_0 = A.$$

[3] There are several equivalent definitions for matrix functions, which are discussed in detail in [9, Chapter 1].

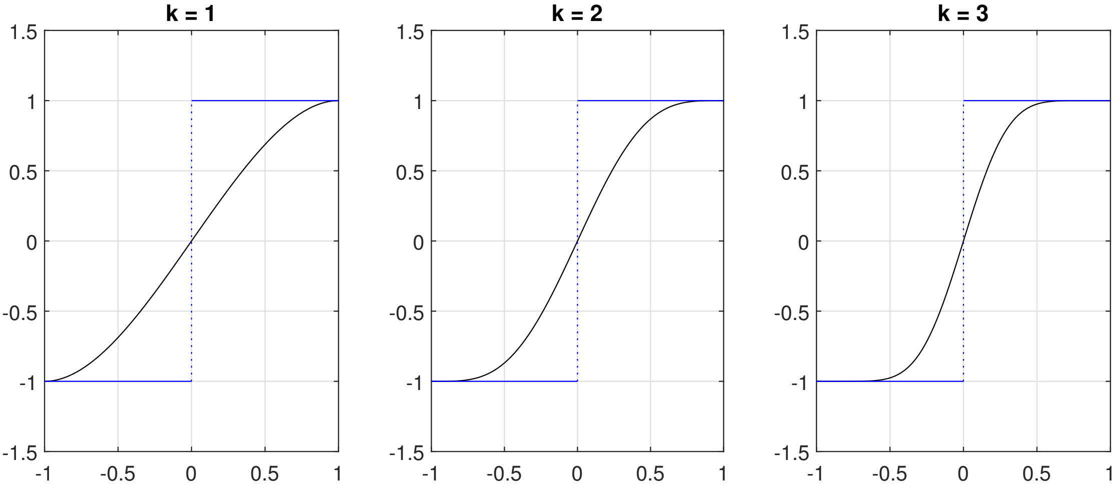

Figure 2.3: Newton-Schulz iterates $f_{k+1}(x) = f_k(x)(3 - f_k(x)^2)/2$ (black) to $\mathrm{sgn}(x)$ (blue) on $[-1, 1]$.

This sequence also converges quadratically—provided that $\|I - A^2\|_2 < 1$, where $\|\cdot\|_2$ denotes the matrix 2-norm[4]—though many iterations are required before quadratic convergence is seen [4, Theorem 1.1]. Similarly, in the scalar case we could approximate $\mathrm{sgn}(x)$ using the functions

$$f_{k+1}(x) = g(f_k(x)), \qquad g(x) = \frac{x}{2}(3 - x^2), \tag{2.8}$$

with $f_0(x) = x$. We shall refer to (2.8) as the *(unscaled) Newton-Schulz* iteration for $\mathrm{sgn}(x)$. The first few iterations are illustrated in Figure 2.3: the uniform error of the Newton-Schulz iteration is always 1 at the origin, so in practice we will consider the iteration on $X(\delta) = [-1, -\delta] \cup [\delta, 1]$ for some $\delta \in (0, 1)$. The maximum uniform error is then given by

$$\|f_k - \mathrm{sgn}\|_{\infty, X(\delta)} = f_k(\delta).$$

Remark. The Newton-Schulz iteration is one of a family of iterations of the form $X_{k+1} = g_{m,n}(X_k)$, where $g_{m,n}(x) = xp_m(1 - x^2)/q_n(1 - x^2)$, used for computing $\mathrm{sgn}(A)$. Here p_m/q_n represents the $[m/n]$ *Padé approximant* of $(1-x)^{-1/2}$ (see e.g. [16, Section 3]). When $n = 0$, we obtain composite polynomial approximations; in particular we recover the Newton-Schulz iteration for $(m, n) = (1, 0)$.

[4] It can be shown that the 2-norm of a matrix B is equal to $\sigma_{\max}(B)$, the largest singular value of B. In the case of Hermitian matrices, the singular values are equal to the absolute values of the eigenvalues (see e.g. [30]), hence the Newton-Schulz iterates for A are quadratically convergent when the eigenvalues of A have magnitude strictly less than $\sqrt{2}$. It may thus be required to scale A down to a matrix with suitably small eigenvalues before proceeding.

2.3.2 Improved Newton's Method

Despite being a quadratically converging sequence, the Newton-Schulz iteration (2.8) can take a large number of iterations to reach a suitably small error. For example, when $\delta = 10^{-2}$, it takes 14 iterations to obtain an error $\varepsilon < 10^{-2}$ on $X(\delta)$. This is because the order of convergence is an *asymptotic* property, and in practice, convergence breaks down into two phases:

- the number of iterations required to reduce the error to some small value;
- the asymptotic behaviour given by the order of convergence.

Crucially, this means that we cannot guarantee that a quadratically convergent sequence (such as Newton's Method, or the Newton-Schulz iterates) will converge in a small number of iterations.

Many contributions have been made to reduce the number of iterations required in the initial phase of convergence by considering scaled iterative processes, such as in [21, 24]. These processes generally take the form $f_{k+1}(x) = g_{k+1}(x, f_k(x))$, where the iteration functions g_k suitably scale each new iteration based on the previous one. For example, Ninomiya provides an improved version of Newton's Method for $\sqrt{x}$ on $I = [a, b]$ where $0 < a < b$ in [21], after noting that each iteration of (2.4) has an upward bias. In this subsection, we will show how he proved this observation, and discuss the scaled iteration he proposed. We begin by defining

$$N^*(x) = \frac{1}{2}(x + x^{-1}), \qquad x > 0,$$

and observe the following properties:

(1) $N^*(x) \geqslant N^*(1) = 1$;

(2) $xy = 1 \Longrightarrow N^*(x) = N^*(y)$;

(3) $(xy - 1)(x - y) > 0 \Longleftrightarrow N^*(x) > N^*(y)$.

With the f_k as in (2.4), we can write

$$s_{k+1}(x) = N^*(s_k(x)), \qquad s_k(x) = \frac{f_k(x)}{\sqrt{x}}.$$

Writing the Newton iterates in this way, and defining

$$s'_k = \min_{x \in I} s_k(x), \qquad s''_k = \max_{x \in I} s_k(x),$$

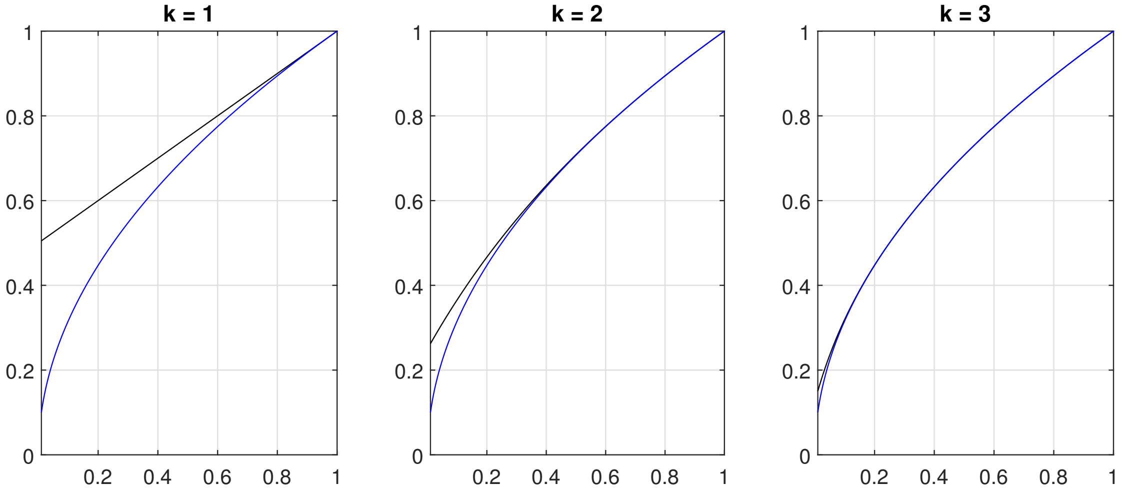

Figure 2.4: Newton iterates $f_{k+1}(x) = (f_k(x) + x/f_k(x))/2$ (black), plotted with $\sqrt{x}$ (blue) on $[\delta^2, 1]$, where $\delta = 0.1$. The f_k always overestimate f in the sense of (2.9).

we see that $s'_k = \min_{x\in I} N^*(s_{k-1}(x)) \geqslant 1$ for $k = 1, 2, \ldots,$ from condition (1). Moreover, by condition (3) we have

$$s'_k s''_k > 1. \tag{2.9}$$

Figure 2.4 demonstrates this effect, which is problematic because it means that the error in iterations can never equioscillate, for instance. With the aim of finding a sequence that satisfies $s'_k s''_k = 1$, Ninomiya proposes the improved Newton iterates

$$F_{k+1}(x) = \frac{C_{k+1}}{2}\left(F_k(x) + \frac{x}{F_k(x)}\right), \qquad k = 0, 1, 2, \ldots, \tag{2.10}$$

with $F_0 = f_0/\sqrt{s'_0 s''_0}$. The constants C_{k+1} are given by

$$C_{k+1} = \frac{1}{\sqrt{N^*(S''_k)}}, \qquad S''_k = \max_{x\in I} \frac{F_k(x)}{\sqrt{x}}.$$

Assuming that we have optimised our initial guess so that $s'_0 s''_0 = 1$, Ninomiya analyses the convergence of the F_k using the following recursions, whose proofs were omitted from his paper. We prove them here as a technical lemma.

Lemma 2.11. *The maximum relative errors $e''_k = s''_k - 1$ and $E''_k = S''_k - 1$ satisfy*

$$e''_{k+1} = \frac{e''^2_k}{2(1 + e''_k)}, \qquad E''_{k+1} = \sqrt{\frac{E''^2_k}{2(1 + E''_k)} + 1} - 1.$$

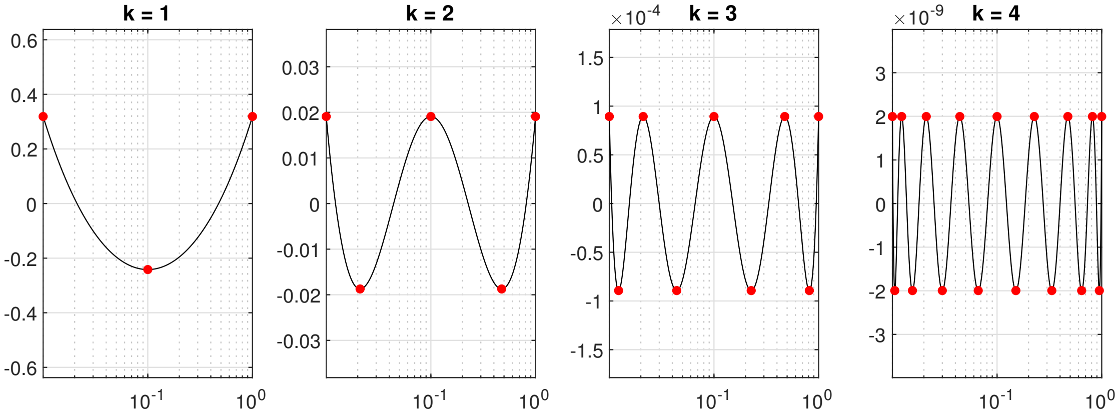

Figure 2.5: Relative error $E_k(x) = F_k(x)/\sqrt{x} - 1$ in Ninomiya's improved Newton iterates (2.10) to $\sqrt{x}$ on $[\delta^2, 1]$, where $\delta = 0.1$.

Proof. Let $e_k(x) = s_k(x) - 1$ and $E_k(x) = S_k(x) - 1$ be the relative errors of $f_k(x)$, $F_k(x)$ respectively. Then

$$\begin{aligned} e_{k+1}(x) &= s_{k+1}(x) - 1 \\ &= \frac{1}{2}\left(s_k(x) + \frac{1}{s_k(x)}\right) - 1 \\ &= \frac{1}{2}\left(e_k(x) + 1 + \frac{1}{e_k(x)+1}\right) - 1 \\ &= \frac{e_k(x)^2}{2(1+e_k(x))}. \end{aligned}$$

Since $x \mapsto x^2(1+x)^{-1}$ is an increasing function on $[0,1]$, $e_{k+1}(x)$ obtains its maximum when $e_k(x)$ is maximised, i.e. $e''_{k+1} = e''^2_k/(2(1+e''_k))$. Similarly we have

$$\begin{aligned} E_{k+1}(x) &= S_{k+1} - 1 \\ &= \frac{C_{k+1}}{2}\left(S_k(x) + \frac{1}{S_k(x)}\right) - 1 \\ &= \frac{1}{2\sqrt{N^*(E''_k+1)}}\left(E_k(x) + 1 + \frac{1}{E_k(x)+1}\right) - 1 \\ &= \frac{1}{\sqrt{N^*(E''_k+1)}}\left(\frac{E_k(x)^2}{2(1+E_k(x))} + 1\right) - 1. \end{aligned}$$

As before, $E_{k+1}(x)$ obtains its maximum when $E_k(x)$ is maximised, giving

$$E''_{k+1} = \frac{1}{\sqrt{N^*(E''_k+1)}}\left(\frac{E''^2_k}{2(1+E''_k)} + 1\right) - 1$$

$$= \sqrt{\frac{E_k''^2}{2(1+E_k'')} + 1} - 1,$$

by definition of N^*, as required. □

By Lemma 2.11 and a Taylor expansion, we obtain the approximate relations

$$e_{k+1}'' \sim \frac{e_k''^2}{2}, \qquad E_{k+1}'' \sim \frac{E_k''^2}{4}$$

as $k \to \infty$. Combining with $E_0'' = e_0''$, we obtain the approximate relation

$$E_k'' \sim \frac{e_k''}{2^{2^k - 1}},$$

demonstrating that the iterates (2.10) improve convergence. Figure 2.5 sketches the relative error of the improved Newton iterates (2.10), which appear to be the best rational approximants in the relative sense as they have the correct number of equioscillation points[5]. However, they do not equioscillate between $\pm E_k''$ upon closer inspection. Ninomiya shows that the F_k are best approximations over all R of the same order with respect to the error of $g(x, R(x))$ relative to $\sqrt{x}$, where g is the standard Newton iteration (2.4) for $\sqrt{x}$ (see [21, Theorem 4] for details).

While Ninomiya's improved Newton iteration (2.10) is not a composite polynomial approximation to $\sqrt{x}$, we can consider his work as motivation for constructing such an approximation. Ideally, the error of our approximation should not display an upward bias, and have a similar error curve to that of Figure 2.5. In general, this dissertation is concerned with constructing polynomial approximations, such as to the Newton-Schulz approximation (2.8) to $\operatorname{sgn}(x)$ and the Alternative Newton approximation (2.5) to $\sqrt{x}$, for which the convergence is accelerated. From the discussion of this section, it seems reasonable that suitably scaling the existing methods at each iteration is a sure-fire way to proceed. In the next chapter, we take a different approach to constructing an approximation to the sign function, yet discover that our method is in fact equivalent to a scaled Newton-Schulz method.

[5] To see this, we note that F_k is of type $(2^{k-1}, 2^{k-1} - 1)$ by induction, hence the best rational approximant equioscillates between at least $2^k + 1 - d_{F_k}$ extrema by Theorem 2.8.

Chapter 3

Composite polynomial approximation to sgn(x)

3.1 Zolotarev functions

To motivate the work of this chapter, we first discuss a notable result in rational approximation theory, concerning the rational best approximations to sgn(x) on

$$X(\delta) = [-1, -\delta] \cup [\delta, 1],$$

for some $\delta \in (0, 1)$. Such approximations are referred to as the *Zolotarev functions*; in particular, we write $Z_{2r+1}(x; \delta)$ for type $(2r+1, 2r)$ approximations. The explicit form of $Z_{2r+1}(\cdot; \delta)$ was found by Zolotarev [31] in terms of Jacobi elliptic functions

$$\operatorname{sn}(u; \delta) = \sin\varphi, \qquad \operatorname{cn}(u; \delta) = \cos\varphi;$$

here φ denotes the Jacobi amplitude, obtained from the inversion of the *incomplete elliptic integral of the first kind* (see e.g. [2, Chapter 5])

$$u = F(\varphi; \delta) := \int_0^\varphi \frac{d\theta}{\sqrt{1 - \delta^2 \sin^2\theta}}.$$

The Zolotarev functions are then given by

$$Z_{2r+1}(x; \delta) = Mx \prod_{i=1}^{r} \frac{x^2 + c_{2i}}{x^2 + c_{2i-1}}, \tag{3.1}$$

where for $i = 1, \ldots, 2r$ the constants c_i are defined by

$$c_i = \delta^2 \frac{\operatorname{sn}^2\left(\frac{iK'}{2r+1}; \delta'\right)}{\operatorname{cn}^2\left(\frac{iK'}{2r+1}; \delta'\right)},$$

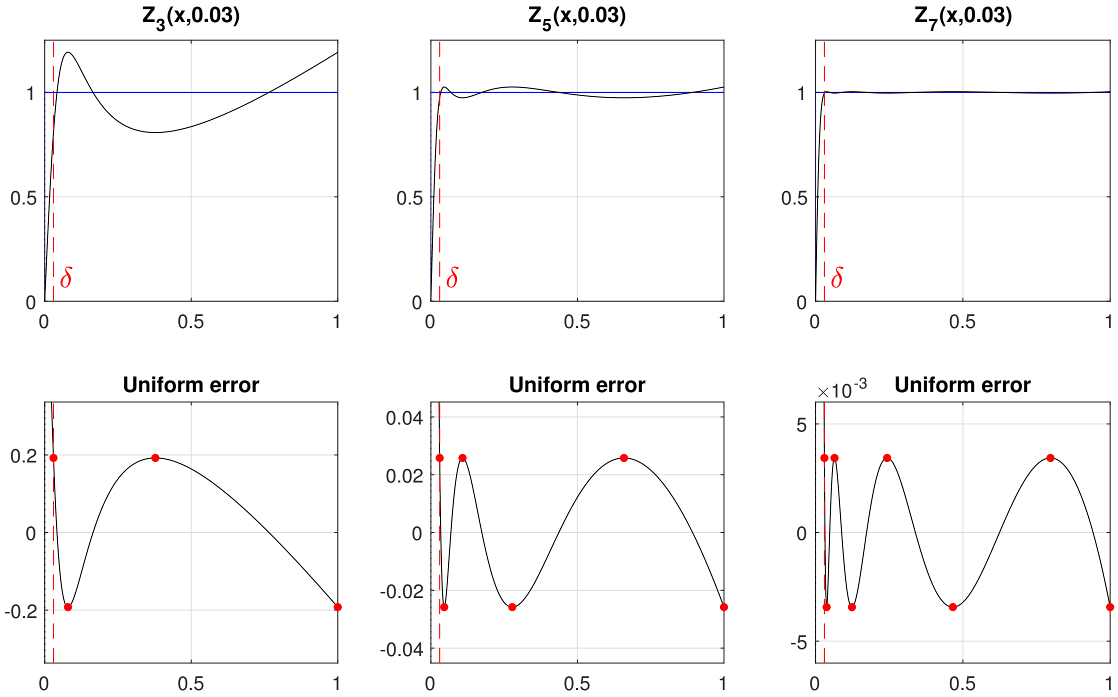

Figure 3.1: Zolotarev functions $Z_{2r+1}(x;\delta)$ for $\delta = 0.03$, constructed using `ellipke` and `ellipj`. The corresponding uniform error $\operatorname{sgn}(x) - Z_{2r+1}(x;\delta)$ is sketched below each plot, which equioscillates between $4r+4$ extrema in each case.

with $\delta' = \sqrt{1-\delta^2}$ and $K' = F(\pi/2;\delta')$ respectively denoting the *complementary modulus* and *complete elliptic integral of the first kind* for δ'. Here, $M > 0$ is a scalar determined by the equioscillation property of the best approximation, which ensures that $1 - Z_{2r+1}(\delta;\delta) = -(1 - Z_{2r+1}(1;\delta))$. Hence

$$M = 2\left(\prod_{i=1}^{r}\frac{1+c_{2i}}{1+c_{2i-1}} + \delta\prod_{i=1}^{r}\frac{\delta^2+c_{2i}}{\delta^2+c_{2i-1}}\right)^{-1}.$$

Nakatsukasa and Freund showed that normalised Zolotarev functions $\hat{Z}_{2r+1}(\cdot;\delta)$, which take the form (3.1) multiplied by $1/Z_{2r+1}(1;\delta)$, satisfy a recursive optimality property. Explicitly, we can write

$$\hat{Z}_{2r'+1}(\hat{Z}_{2r+1}(x;\delta_1);\delta_2) = \hat{Z}_{(2r+1)(2r'+1)}(x;\delta_1),$$

where δ_2 is such that $X(\delta_2) = \hat{Z}_{2r+1}(X(\delta_1);\delta_1)$ [19, Theorem 3]. The proof involves counting equioscillation points and invoking the minimax equioscillation property. We use this as motivation for constructing a composite polynomial approximation consisting of the composition of low-order best approximations to the sign function.

3.2 Greedy approximation to sgn(x) in $\mathcal{P}^{\mathbf{comp}}_{(k,3)}$

In this section, we construct a pure composite polynomial approximation to $\operatorname{sgn}(x)$ on $X(\delta)$ using a greedy iterative process, namely an approximation of the form

$$f_{k+1}(x) = g_{k+1}(f_k(x)), \qquad k = 1, 2, \dots,$$

where each $g_{k+1} \in \mathcal{P}_3$ is the best approximation to $\operatorname{sgn}(x)$ on its domain $f_k(X(\delta))$. We start by finding the best polynomial approximant $p \in \mathcal{P}_3$ to $\operatorname{sgn}(x)$ on $X(\delta)$, which is an odd function by Lemma 2.7, hence has the form

$$p(x) = x(A + Bx^2),$$

for some constants A, B to be determined. For simplicity, we normalise p such that $\|p\|_{\infty,X(\delta)} = 1$, and as such approximate a scaled version of the sign function $C\operatorname{sgn}(x)$, for some constant $C \in (0, 1)$ determined by the normalisation of p. By Corollary 2.6, we know that $C\operatorname{sgn} - p$ equioscillates between at least 6 extrema: to obtain equioscillation, we can impose

$$p(\delta) = p(1) = 2C - 1, \tag{3.2}$$

so that the maximum error $1 - C$ is obtained at the points $\pm 1, \pm\delta$, $\pm\xi$, where $\xi \in (\delta, 1)$ is the extremal point such that $p(\xi) = 1$. This gives us three conditions:

- $p(\delta) = p(1) \implies (1 - \delta)A + (1 - \delta^3)B = 0;$
- $p'(\xi) = 0 \implies A + 3\xi^2 B = 0;$
- $p(\xi) = 1 \implies \xi A + \xi^3 B - 1 = 0.$

Solving these equations, we find that the best approximation is

$$p(x) = \frac{x}{2\xi}\left(3 - \frac{x^2}{\xi^2}\right), \qquad \xi = \sqrt{\frac{1 + \delta + \delta^2}{3}}, \tag{3.3}$$

By the symmetry of our construction, and using that $C = (1 + p(\delta))/2$ by (3.2), we can obtain the maximum error E in the approximation

$$E = 1 - C = \frac{1 - p(\delta)}{2}.$$

Figure 3.2 sketches the approximation for $\delta = 0.1$, for which the normalisation of p gives $C \approx 0.6222$ and $E \approx 0.3778$.

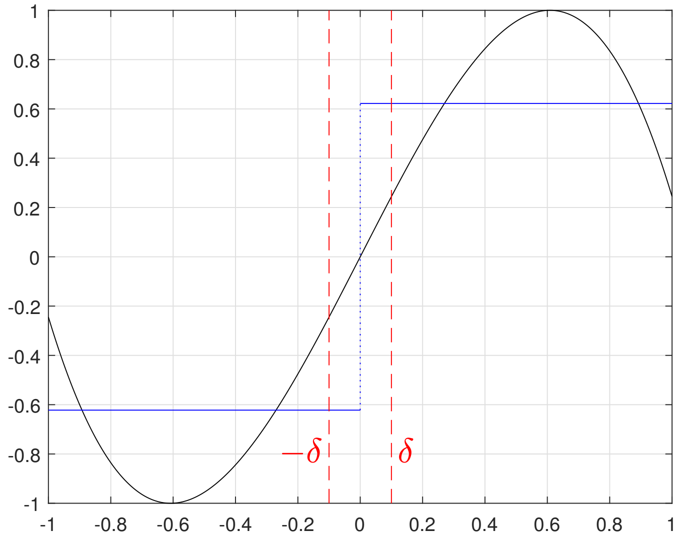

Figure 3.2: Best $\mathcal{P}_3$-approximation (black) to $C\mathrm{sgn}(x)$ (blue) on $X(\delta)$, where $\delta = 0.1$.

Remark. By symmetry, the range of p is $X(p(\delta))$. In particular, the containment $X(p(\delta)) \subset X(\delta)$ is strict, thanks to the following lemma.

Lemma 3.1. *Given $\delta \in (0,1)$ and $p \in \mathcal{P}_3$ defined by* (3.3)*, we have $p(\delta) > \delta$.*

Sketch of proof. There holds

$$\frac{p(\delta)}{\delta} = \frac{1}{2\xi}\left(3 - \frac{\delta^2}{\xi^2}\right) = \frac{3\sqrt{3}}{2}\left(\frac{1+\delta}{(1+\delta+\delta^2)^{3/2}}\right),$$

so it remains to show that

$$\frac{3\sqrt{3}}{2} > \frac{(1+\delta+\delta^2)^{3/2}}{1+\delta} =: h(\delta).$$

But h is strictly increasing, therefore $h(\delta) < h(1) = 3\sqrt{3}/2$. □

Fixing p as in (3.3), we can similarly find the polynomial

$$q \in \{f \in \mathcal{P}_3 : \|f\|_{\infty, p(X(\delta))} = 1\}$$

such that $q(p(x))$ is an optimal approximation for $D\mathrm{sgn}(x)$, for some scale factor D determined by the normalisation of q. To see this, note that by the above remark, the domain of q is $p(X(\delta)) = X(p(\delta))$. Hence q is obtained in the same manner as p, except with δ replaced by $p(\delta)$. Repeating this process, we obtain a sequence

$$f_{k+1}(x) = g_{k+1}(f_k(x)), \qquad k = 0, 1, 2, \ldots,$$

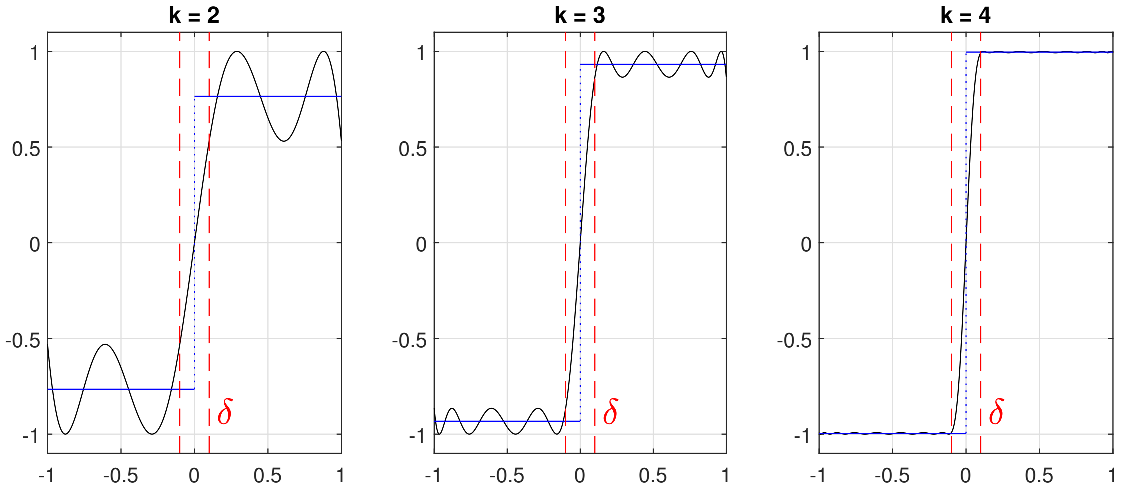

Figure 3.3: Scaled Newton-Schulz iterates $f_k(x)$ (blue) with $C_k\text{sgn}(x)$ (black) on $X(\delta)$, where $\delta = 0.1$. The iterate for $k = 1$ is shown in Figure 3.2.

where $f_0(x) = x$ and

$$g_{k+1}(x) = \frac{x}{2\xi_{k+1}}\left(3 - \frac{x^2}{\xi_{k+1}^2}\right), \qquad \xi_{k+1} = \sqrt{\frac{1 + f_k(\delta) + f_k(\delta)^2}{3}}. \tag{3.4}$$

By induction, it follows that $f_k \in \mathcal{P}^{\text{comp}}_{(k,3)}$, and is an approximation for $C_k\text{sgn}(x)$ with maximum uniform error E_k, where

$$C_k = \frac{1 + f_k(\delta)}{2}, \qquad E_k = \frac{1 - f_k(\delta)}{2}. \tag{3.5}$$

Remark. We note that $g_k(x) = g(x/\xi_k)$, where g is the iteration function for the unscaled Newton-Schulz approximation to the sign function (2.8). Hence the iteration (3.4) is equivalent to a scaled version of the of (2.8), with x replaced by x/ξ_k at the k^{th} iteration. As such, we will refer to (3.4) as *scaled Newton-Schulz* iterates to the sign function.

3.3 Equioscillation behaviour

Since the scaled Newton-Schulz iterate f_k has degree 3^k, it follows by Corollary 2.6 that it is the best $\mathcal{P}_{3^k}$-approximation if $f_k - C_k f$ equioscillates between at least $3^k + 3$ extrema on $X(\delta)$. By symmetry, this is the case when $f_k - C_k$ equioscillates between $(3^k + 3)/2$ extrema on $[\delta, 1]$. In the following technical lemma, we show that $f_k - C_k$ falls short of this number of equioscillation points.

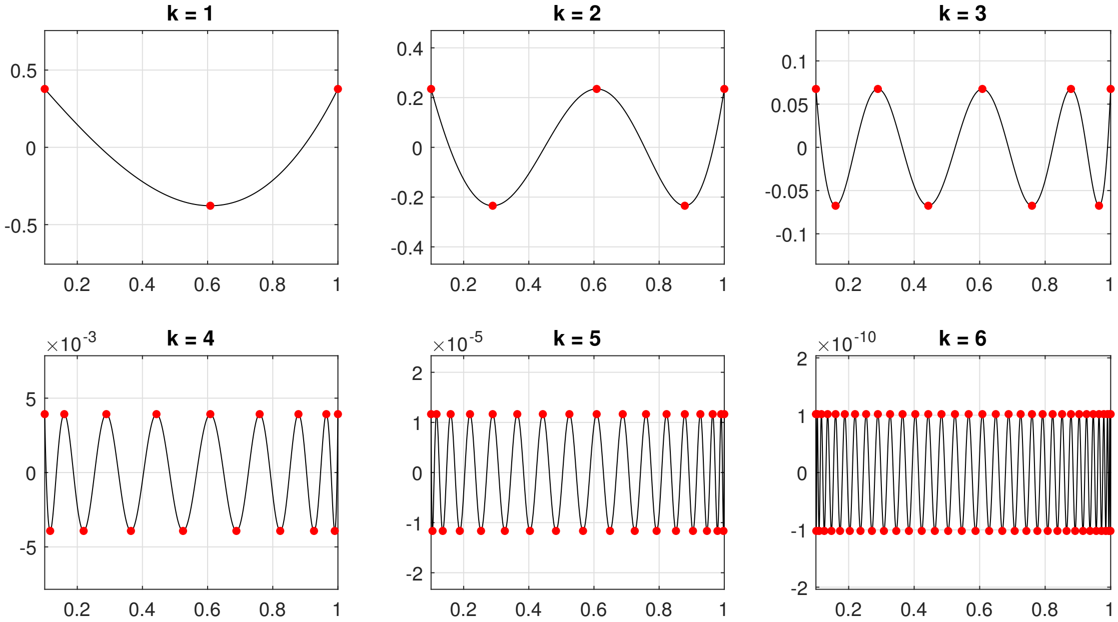

Figure 3.4: Error $E_k(x) = C_k\text{sgn}(x) - f_k(x)$ on $[\delta, 1]$, where $\delta = 0.1$. The error curve equioscillates between $2^k + 1$ extrema in each case, as shown in Lemma 3.2.

Lemma 3.2. *Let $\delta \in (0, 1)$ and $f_k \in \mathcal{P}^{\text{comp}}_{(k,3)}$ be defined by (3.4). Then the error curve $f_k - C_k\text{sgn}$ equioscillates between $2^{k+1} + 2$ extrema on $X(\delta)$, or equivalently, $f_k - C_k$ equioscillates between $2^k + 1$ extrema on $[\delta, 1]$.*

Proof. By induction. The case for $k = 1$ is clear, as we constructed f_1 to have 6 equioscillation points on $X(\delta)$.

Now assume that $f_k - C_k\text{sgn}$ equioscillates between $2^{k+1} + 2$ extrema on $X(\delta)$, or equivalently between $2^k + 1$ points $\{x_j\}_{j=1}^{2^k+1}$ on $[\delta, 1]$. Then there are 2^k intervals $I_j = [x_j, x_{j+1}]$, $j = 1, \dots, 2^k$ such that $f_k(I_j) = [f_k(\delta), 1]$.

Consider $g_{k+1}(x)$ on $[f_k(\delta), 1]$. By construction, g_{k+1} equioscillates between three points on this interval, namely $f_k(\delta)$, ξ_{k+1} and 1. Hence $g_{k+1}(f_k(x))$ equioscillates between 3 points on each I_j. Since we have 2^k such intervals, and $2^k - 1$ points $\{x_j\}_{j=2}^{2^k}$ where equioscillation points overlap, the total number of equioscillation points of $g_{k+1}(f_k(x))$ on $[\delta, 1]$ is

$$3(2^k) - (2^k - 1) = 2^{k+1} + 1,$$

hence $2^{k+2} + 2$ points on $X(\delta)$ by symmetry. But $g_{k+1}(f_k(x)) = f_{k+1}(x)$, so this completes the inductive step of the theorem. □

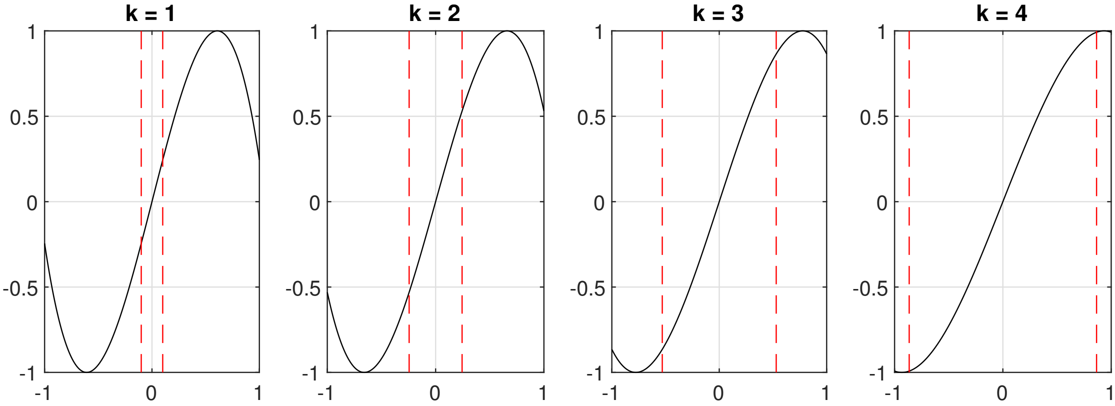

Figure 3.5: Iteration functions g_k for the scaled Newton-Schulz iterates (3.6). In each plot, the dotted red lines show the values of $\pm\delta_{k-1}$, starting with $\delta_0 = \delta = 0.1$.

3.4 Convergence analysis of the scaled Newton-Schulz iterates

We know from Section 3.3 that the recursive optimality property of the Zolotarev functions (3.1) does not hold in the polynomial setting: composing normalised best cubic approximations to the sign function on $X(\delta)$ does not result in best approximations of higher order. However, we shall compare the convergence of the scaled Newton-Schulz iterates to the unscaled iterates and the minimax. In what follows, it will be useful to write $\delta_k := f_k(\delta)$, so that (3.4) reads

$$f_{k+1}(x) = \frac{f_k(x)}{2\xi_{k+1}}\left(3 - \frac{f_k(x)^2}{\xi_{k+1}^2}\right), \qquad \xi_{k+1} = \sqrt{\frac{1+\delta_k+\delta_k^2}{3}}. \tag{3.6}$$

In particular, δ_k represents the value of δ used at the $(k+1)^{\mathrm{st}}$ iteration to find the best cubic approximation to $C_{k+1}\mathrm{sgn}(x)$ on $X(\delta_k)$, as demonstrated in Figure 3.5. As a starting point, we have the formula

$$\varepsilon_k := \frac{E_k}{C_k} = \frac{1-\delta_k}{1+\delta_k},$$

which is the maximum uniform error $C_k^{-1}f_k - \mathrm{sgn}$. A comparison in the error of the scaled/unscaled Newton-Schulz iterates and the minimax approximation with respect to the degrees of freedom yields a striking result. Recall that the k^{th} Newton-Schulz iterate (scaled or unscaled) has degree 3^k, yet only $2k$ degrees of freedom[1]. However, the best approximation of degree 3^k requires $(3^k+1)/2$ degrees

[1] Each iteration function g_k is an odd cubic polynomial, needing only two coefficients in terms of ξ_k to be defined.

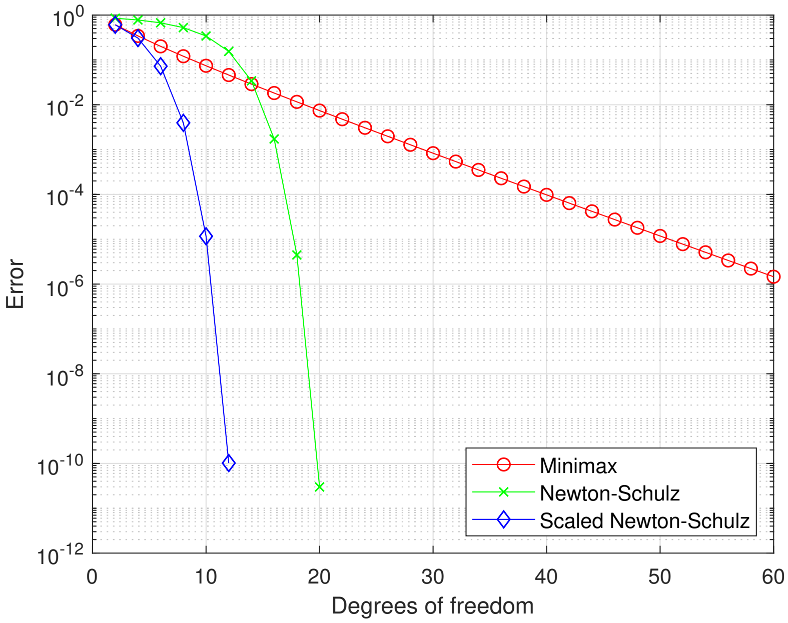

Figure 3.6: Comparison in the error of the Newton-Schulz, scaled Newton-Schulz and best approximants to $\operatorname{sgn}(x)$ on $X(\delta)$, where $\delta = 0.1$, with respect to the degrees of freedom. The minimax is approximated using `polyvalA` and `polyfitA_Lawson`[2].

of freedom—half of a plain polynomial of degree 3^k, since the best approximation is also an odd function). Therefore, the k^{th} scaled or unscaled Newton-Schulz iterate has the same number of degrees of freedom as the minimax approximation of degree $4k - 1$.

Figure 3.6 shows that the composite polynomial approximations are superior to the minimax with respect to degrees of freedom. Compared to [5, Theorem 1], which shows that the minimax $p_m^* \in \mathcal{P}_{2m+1}$ to $\operatorname{sgn}(x)$ on $X(\delta)$ satisfies

$$\|p_m^* - \operatorname{sgn}\|_{\infty, X(\delta)} \sim \frac{1-\delta}{\sqrt{\pi\delta}} m^{-1/2} \left(\frac{1-\delta}{1+\delta}\right)^m \tag{3.7}$$

as $m \to \infty$, we gain insight into why convergence of the minimax is much slower: despite being exponentially convergent, the term $\frac{1-\delta}{1+\delta}$ is very close to 1 for small values of δ. A further observation from Figure 3.6 is that, whereas the standard Newton-Schulz iterates perform worse than the minimax in the initial convergence phase, the scaled Newton-Schulz iterates are quick to outperform the minimax.

[2] Introduced by Brubeck, Nakatsukasa and Trefethen, these functions are adaptations of the standard `polyval` and `polyfit` methods—which fit polynomials to data using Vandermonde matrices—made stable by means of Arnoldi orthogonalisation. For details, see [3].

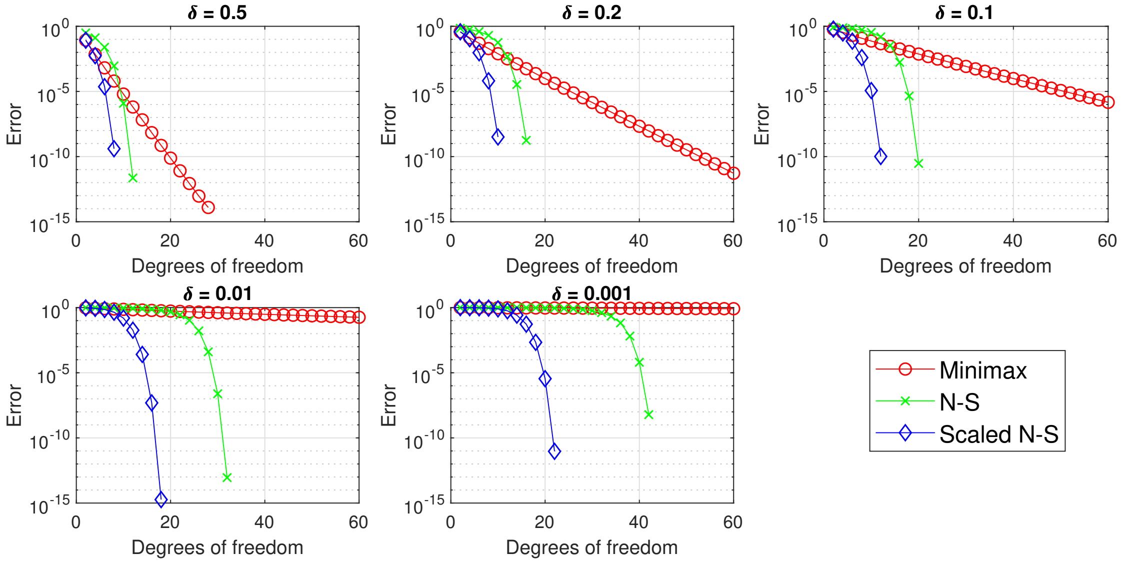

Figure 3.7: Error of the unscaled/scaled Newton-Schulz iterates and best approximants to $\operatorname{sgn}(x)$ on $X(\delta)$, with respect to the degrees of freedom, for different values of δ.

Comparisons in the error of all the approximations for different values of δ are shown in Figure 3.7. For increasingly small values of δ, we observe that the number of iterations required by the unscaled Newton-Schulz iteration in the initial phase of convergence grows faster than that required by the scaled Newton-Schulz iteration.

3.4.1 Deriving a lower bound for the number of iterations

To obtain a lower bound for the number of iterations needed to obtain a given accuracy $\varepsilon > 0$ for arbitrary δ, it thus seems reasonable to split the convergence analysis into the two phases mentioned in Section 2.3.2: the initial convergence phase where we reduce the error to a suitably small value, and the asymptotic phase thereafter. We do this by taking inspiration from the analysis conducted by Gawlik and Nakatsukasa (see [8], or Appendix A).

Let $\varepsilon > 0$ be the required accuracy. In this subsection, we appropriately select a constant $\delta^* \in (1/e, 1)$ and split the convergence into the the following three steps:

(1) find k_1 such that $\delta_{k_1} \geqslant 1/e$;

(2) find k_2 such that $\delta_{k_1+k_2} \geqslant \delta^*$;

(3) find k_3 such that $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$.

Step 1. Let k_1 be the smallest k such that $\delta_k \geqslant 1/e$. By substituting $x = \delta$ in (3.6), we obtain the recursion $\delta_{k+1} = H(\delta_k)$, where

$$H(x) = \frac{3\sqrt{3}}{2} x \left(\frac{1+x}{(1+x+x^2)^{3/2}} \right). \tag{3.8}$$

A lower bound on k_1 is then readily obtained using the following observation.

Lemma 3.3. *Let $\delta = \delta_0 \in (0,1)$, and define $\delta_{k+1} = H(\delta_k)$ using* (3.8)*. Then*

$$\delta_k > \Delta^k \delta$$

for all $k \leqslant k_1$, where

$$\Delta = \frac{H(1/e)}{1/e} \approx 1.928.$$

Proof. Since $H(x)/x$ is strictly decreasing for $x > 0$, and $\{\delta_k\}_{k \geqslant 0}$ is an increasing sequence, it follows that the ratio δ_{k+1}/δ_k is also strictly decreasing. In particular, for all $k < k_1$, we have

$$\frac{\delta_{k+1}}{\delta_k} > \frac{H(1/e)}{1/e} = \Delta.$$

Inductively, it follows that

$$\delta_k > \Delta^k \delta_0 = \Delta^k \delta$$

for all $k \leqslant k_1$, as required. □

Choosing $k = k_1$ in Lemma 3.3, we find that $\delta_{k_1} \geqslant 1/e$ whenever $\Delta^{k_1}\delta \geqslant 1/e$. As a result, we obtain the lower bound

$$k_1 \geqslant \frac{\log \frac{1}{\delta} - 1}{\log \Delta}.$$

Step 2. This step no longer depends on ε or δ, so k_2 is a constant. We select δ^* such that the following lemma holds, as this will help us in step 3. Firstly, define

$$G(x) = \frac{2\sqrt{x}}{1+x}.$$

Lemma 3.4. *There is $\delta^* \in (0,1)$ such that for every $\delta \in [\delta^*, 1]$, there holds*

$$\frac{1 - G(\delta)}{1 + G(\delta)} \leqslant \left(\frac{1-\delta}{1+\delta} \right)^2.$$

Proof. As in [8, Lemma 4.1]. It is shown in [6, Theorem 3.2] that the sequence defined by $\alpha_{k+1} = G(\alpha_k)$, where $\alpha_0 \in (0,1)$, is increasing to 1 as $k \to \infty$, and

$$\frac{1 - G(\alpha_k)}{1 + G(\alpha_k)} = \frac{1 - \alpha_{k+1}}{1 + \alpha_{k+1}} = \frac{1}{4}\left(\frac{1-\alpha_k}{1+\alpha_k}\right)^2 + o\left(\left(\frac{1-\alpha_k}{1+\alpha_k}\right)^2\right).$$

It follows that

$$\frac{1 - G(\alpha)}{1 + G(\alpha)}\bigg/\left(\frac{1-\alpha}{1+\alpha}\right)^2 \to \frac{1}{4}$$

as $\alpha \to 1^-$. Hence $\frac{1-G(\alpha)}{1+G(\alpha)}\big/\left(\frac{1-\alpha}{1+\alpha}\right)^2$ is bounded by 1 for α sufficiently close to 1. □

Step 3. Let k_3 be such that $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$. To find a lower bound on k_3, we will use one final technical lemma.

Lemma 3.5. *For every $x \in [0,1]$, we have*

$$H(x) \leqslant G(x).$$

Proof. Re-arranging the desired inequality, it suffices to show that

$$J(x) := \frac{x(1+x)^4}{(1+x+x^2)^3} \leqslant \frac{16}{27}$$

for all $x \in [0,1]$. Differentiating J, we find that

$$J'(x) = \frac{(1-x)(x^2+4x+1)(1+x)^3}{(1+x+x^2)^4} \geqslant 0$$

on $[0,1]$, with equality only when $x = 1$. Hence J is bounded by $J(1) = 16/27$. □

It follows by Lemma 3.5 that

$$\frac{1 - H(\delta_k)}{1 + H(\delta_k)} \leqslant \frac{1 - G(\delta_k)}{1 + G(\delta_k)},$$

since $x \mapsto \frac{1-x}{1+x}$ is decreasing. By our choice of δ^*, we apply Lemma 3.4 to obtain

$$\frac{1 - \delta_{k+1}}{1 + \delta_{k+1}} = \frac{1 - H(\delta_k)}{1 + H(\delta_k)} \leqslant \frac{1 - G(\delta_k)}{1 + G(\delta_k)} \leqslant \left(\frac{1-\delta_k}{1+\delta_k}\right)^2,$$

for $k \geqslant k_1 + k_2$. Inductively, we obtain

$$\frac{1 - \delta_{k_1+k_2+k}}{1 + \delta_{k_1+k_2+k}} \leqslant \left(\frac{1-\delta^*}{1+\delta^*}\right)^{2^k},$$

so we have $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$ if

$$\left(\frac{1-\delta^*}{1+\delta^*}\right)^{2^{k_3}} \leqslant \varepsilon.$$

Taking logarithms twice, we find

$$k_3 \geqslant \frac{\log\log\frac{1}{\varepsilon} - \log\log\frac{1+\delta^*}{1-\delta^*}}{\log 2}.$$

Combining the all steps, we have that $\varepsilon_k \leqslant \varepsilon$ when

$$k \geqslant \frac{\log\frac{1}{\delta}}{\log\Delta} + \tilde{k}_2 + \frac{\log\log\frac{1}{\varepsilon}}{\log 2}, \tag{3.9}$$

where $\tilde{k}_2$ is a constant such that

$$\tilde{k}_2 = k_2 - \frac{1}{\log\Delta} - \frac{\log\log\frac{1+\delta^*}{1-\delta^*}}{\log 2}.$$

3.4.2 Convergence with respect to degrees of freedom

In Figure 3.7, we saw how the scaled Newton-Schulz approximation outperformed the minimax with respect to degrees of freedom. While (3.7) shows that the minimax converges exponentially with respect to degrees of freedom, we can use (3.9) to prove that, for any fixed value of δ, the convergence of the scaled Newton-Schulz approximation is *doubly exponential* with respect to degrees of freedom.

Theorem 3.6. *Let $\delta \in (0,1)$, and $f_k \in \mathcal{P}^{comp}_{(k,3)}$ be the scaled Newton-Schulz iteration defined by* (3.6)*. Then there exist $C_1, C_2 > 0$ such that*

$$\|f_k - C_k\mathrm{sgn}\|_{\infty,X(\delta)} = O(\exp(-C_1\exp(C_2 d))),$$

where $d = 2k$ denotes the number of degrees of freedom of f_k.

Proof. Equivalently we can show that ε_k, the maximum uniform error of $C_k^{-1}f_k$ to sgn on $X(\delta)$, is $O(\exp(-C_1\exp(C_2 d)))$. Increasing $\tilde{k}_2$ until the right-hand side of (3.9) is an integer, and recalling that f_k has degree 3^k, we find that the degree n of f_k achieving accuracy $\varepsilon_k \leqslant \varepsilon$, where $\varepsilon > 0$, satisfies

$$\frac{\log n}{\log 3} = \frac{\log\frac{1}{\delta}}{\log\Delta} + \tilde{k}_2 + \frac{\log\log\frac{1}{\varepsilon}}{\log 2}.$$

We rearrange this to obtain a bound on ε as follows:

$$\log\log\frac{1}{\varepsilon} = \log 2\left(\frac{\log n}{\log 3} - \frac{\log\frac{1}{\delta}}{\log\Delta} - \tilde{k}_2\right)$$

$$= C \log n + D,$$

where $C = \log_3 2$ and $D = \log 2 \left(\frac{\log \delta}{\log \Delta} - \tilde{k}_2 \right)$. Then

$$\log \frac{1}{\varepsilon} = \exp(C \log n + D) = \tilde{D} n^C,$$

where $\tilde{D} = e^D$. Finally, we obtain

$$\begin{aligned} \varepsilon &= \exp(-\tilde{D} n^C) \\ &= \exp(-\tilde{D} \exp(C \log n)) \\ &= \exp(-\tilde{D} \exp(\tilde{C} d)), \end{aligned}$$

where $\tilde{C} = \frac{1}{2} C \log 3$, since $d = 2k = 2 \log n / \log 3$. Hence the convergence of the scaled Newton-Schulz iteration is doubly exponential with respect to d. □

Chapter 4

Composite polynomial approximation to $\sqrt{x}$

4.1 Zolotarev functions and matrix roots revisited

The Zolotarev functions (3.1) have a diverse range of applications beyond the sign function, as seen in [19, Section 3.7] and [2, Chapter 9]. One particularly interesting feature is that best rational approximations to the square root on $[\delta^2, 1]$, in the relative sense, are closely related to the Zolotarev functions: if

$$Z_{2r+1}(x;\delta) = x\frac{P_r(x^2)}{Q_r(x^2)}, \qquad P_r, Q_r \in \mathcal{P}_r,$$

then $Q_r(x)/P_r(x)$ is a rational best approximation to $\sqrt{x}$. Figure 4.1 illustrates the first few iterates for $\delta = 0.1$.

In recent literature, the Zolotarev functions have been used to construct iterations for computing the principal square root of a matrix [6, 7]. The matrix square root is perhaps the most widely studied matrix function—a range of iterative procedures exist for the approximation of the matrix square root, and they are well-documented in [9, Section 6.3]. Beyond Newton's Method, we can also obtain a large class of iterations[1] for the principal square root from the matrix sign function, by means of the identity

$$\operatorname{sgn}\left(\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix},$$

first noted by Higham in [11]. Clearly there are deep connections between the sign and square root functions, and we use the above observations as motivation

[1] Details can be found in the paper of Higham, Mackey, Mackey & Tisseur [14, Theorem 4.5].

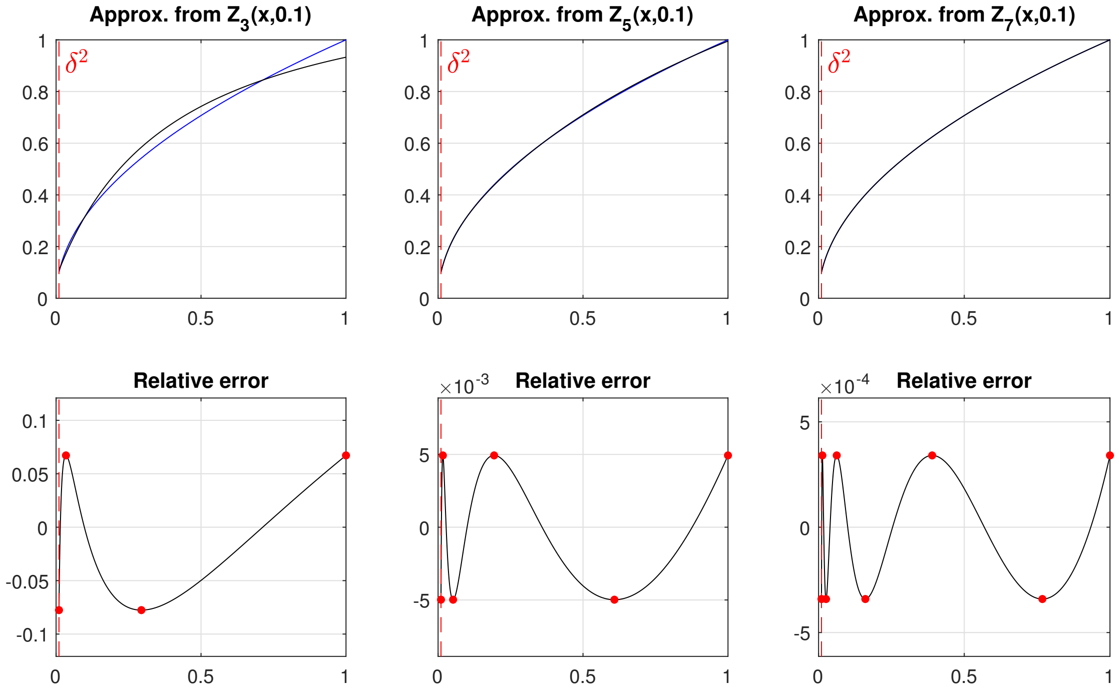

Figure 4.1: Approximations to $\sqrt{x}$ on $[\delta^2, 1]$ for $\delta = 0.03$, derived from the Zolotarev functions $Z_{2r+1}(x;\delta)$. The corresponding relative error is sketched below each plot, which equioscillates between $2r+2$ extrema in each case.

for assesses the connection between $\mathrm{sgn}(x)$ and $\sqrt{x}$ in the setting of polynomial approximations. The following example relates their best approximations over previously considered intervals.

Example 4.1. Consider the function $\mathrm{sgn}(x)$ on $X(\delta)$. Since sgn is odd, and $X(\delta)$ is symmetric about the origin, it follows by the remark after Lemma 2.7 that the best uniform polynomial approximation to $\mathrm{sgn}(x)$ on $X(\delta)$ is the best odd approximation $p^*(x) = xp(x^2)$ on $[\delta, 1]$. Moreover, since $\mathrm{sgn}(x) = x/\sqrt{x^2}$ we can write the uniform error $\|\mathrm{sgn} - p^*\|_{\infty, X(\delta)}$ as

$$\sup_{x\in[\delta,1]} |1 - xp(x^2)| = \sup_{x\in[\delta^2,1]} \left| \frac{xp(x) - \sqrt{x}}{\sqrt{x}} \right| .$$

That is, if the best uniform approximation to $\mathrm{sgn}(x)$ on $X(\delta)$ is $xp(x^2)$, then the best relative approximation to $\sqrt{x}$ on $[\delta^2, 1]$ is $xp(x)$.

Motivated by the connections between best approximations to $\mathrm{sgn}(x)$ and $\sqrt{x}$, this chapter attempts to use our scaled Newton-Schulz approximation to $\mathrm{sgn}(x)$ to generate a composite approximation to $\sqrt{x}$.

4.2 Deriving an approximation from scaled Newton-Schulz iterates

Recall the scaled Newton-Schulz approximation to $\operatorname{sgn}(x)$ on $X(\delta)$, given by

$$f_{k+1}(x) = \frac{f_k(x)}{2\xi_{k+1}}\left(3 - \frac{f_k(x)^2}{\xi_{k+1}^2}\right), \qquad \xi_{k+1} = \sqrt{\frac{1 + f_k(\delta) + f_k(\delta)^2}{3}} \tag{4.1}$$

for $k \geqslant 0$, with initial guess $f_0(x) = x$. Here the f_k approximate $C_k\operatorname{sgn}(x)$, where $C_k = (1 + f_k(\delta))/2$. Each f_k is an odd function, hence by Lemma 2.7 we can write $f_k(x) = xh_k(x^2)$ for some polynomial h_k. This observation allows us to derive a square root approximation, which we illustrate in Figure 4.2.

Theorem 4.2. *Let $\delta \in (0, 1)$, $f_k \in \mathcal{P}_{(k,3)}^{comp}$ be the scaled Newton-Schulz iteration to $C_k sgn(x)$ defined by (4.1), and E_k be the maximum uniform error. Writing $f_k(x) = xh_k(x^2)$, the iterates $F_k(x) = xh_k(x)$ provide a composite approximation*

$$F_{k+1}(x) = \frac{F_k(x)}{2\xi_{k+1}}\left(3 - \frac{F_k(x)^2}{x\xi_{k+1}^2}\right), \qquad \xi_k = \sqrt{\frac{\delta^2 + \delta F_{k-1}(\delta^2) + F_{k-1}(\delta^2)^2}{3\delta^2}}$$

to $\sqrt{x}$, for which

$$\left\|F_k - C_k\sqrt{x}\right\|_{\infty,[\delta^2,1]} \leqslant E_k.$$

Proof. The error bound follows from the fact that

$$\begin{aligned}\left\|\operatorname{sgn} - C_k^{-1}f_k\right\|_{\infty,X(\delta)} &= \left\|\frac{C_k x\operatorname{sgn} - xf_k}{C_k x}\right\|_{\infty,X(\delta)} \\ &= \left\|\frac{C_k\sqrt{x} - F_k}{C_k\sqrt{x}}\right\|_{\infty,[\delta^2,1]} \\ &\geqslant C_k^{-1}\left\|C_k\sqrt{x} - F_k\right\|_{\infty,[\delta^2,1]},\end{aligned}$$

similar to Example 4.1. To obtain a recursion for the F_k, we first note that

$$F_k(\delta^2) = \delta^2 h_k(\delta^2) = \delta f_k(\delta),$$

so we can rewrite ξ_k in terms of the F_{k-1}, namely

$$\xi_k = \sqrt{\frac{\delta^2 + \delta F_{k-1}(\delta^2) + F_{k-1}(\delta^2)^2}{3\delta^2}}.$$

We can derive a sequence for the F_k, since by (4.1) we have

$$xh_{k+1}(x^2) = \frac{xh_k(x^2)}{2\xi_{k+1}}\left(3 - \frac{x^2h_k(x^2)^2}{\xi_{k+1}^2}\right),$$

hence a recursion for the h_k is given by

$$h_{k+1}(x) = \frac{h_k(x)}{2\xi_{k+1}} \left(3 - \frac{x h_k(x)^2}{\xi_{k+1}^2} \right).$$

Multiplying by x gives

$$F_{k+1}(x) = \frac{F_k(x)}{2\xi_{k+1}} \left(3 - \frac{F_k(x)^2}{x\xi_{k+1}^2} \right), \tag{4.2}$$

as required. □

Remark. For any initial guess F_0 divisible by x, it follows inductively that F_k is a polynomial for all k. This is a non-pure composite approximation to $C_k\sqrt{x}$ on $[\delta^2, 1]$ in the relative sense, and the error analysis will thus fall largely in line with that of the previous chapter. Furthermore, the iteration (4.2) can be seen as a scaled version of the Alternative Newton iterates (2.5) considered in the preliminary discussion, where once again the iteration function G_k is such that x is replaced by $x\xi_k^{-1}$. As such, we will refer to these as *scaled Alternative Newton* iterates to $\sqrt{x}$.

4.3 Observations of the scaled Alternative Newton iterates

On the interval $[\delta^2, 1]$, our convergence analysis follows largely as in the previous chapter due to Theorem 4.2. A natural question to ask is how convergence differs when considering the F_k on the whole interval $[0, 1]$. In particular, if we are allowed at most k iterations, what is the value of δ that we should choose to minimise the error? The answer to this question is far from obvious, as Figures 4.3-4.6 illustrate. These figures compare the maximum error in the unscaled and scaled Alternative Newton iterates, alongside the standard Newton iteration. As we might have expected, the Newton iteration eventually outperforms the alternative iterations for all choices of δ. What is unexpected is that for sufficiently small values of δ, the scaled Alternative Newton approximation can even perform better than Newton's method. For instance, if we allow ourselves 10 iterations, it is more efficient to use our approximation with $\delta = 0.001$ with the standard Newton iterates.

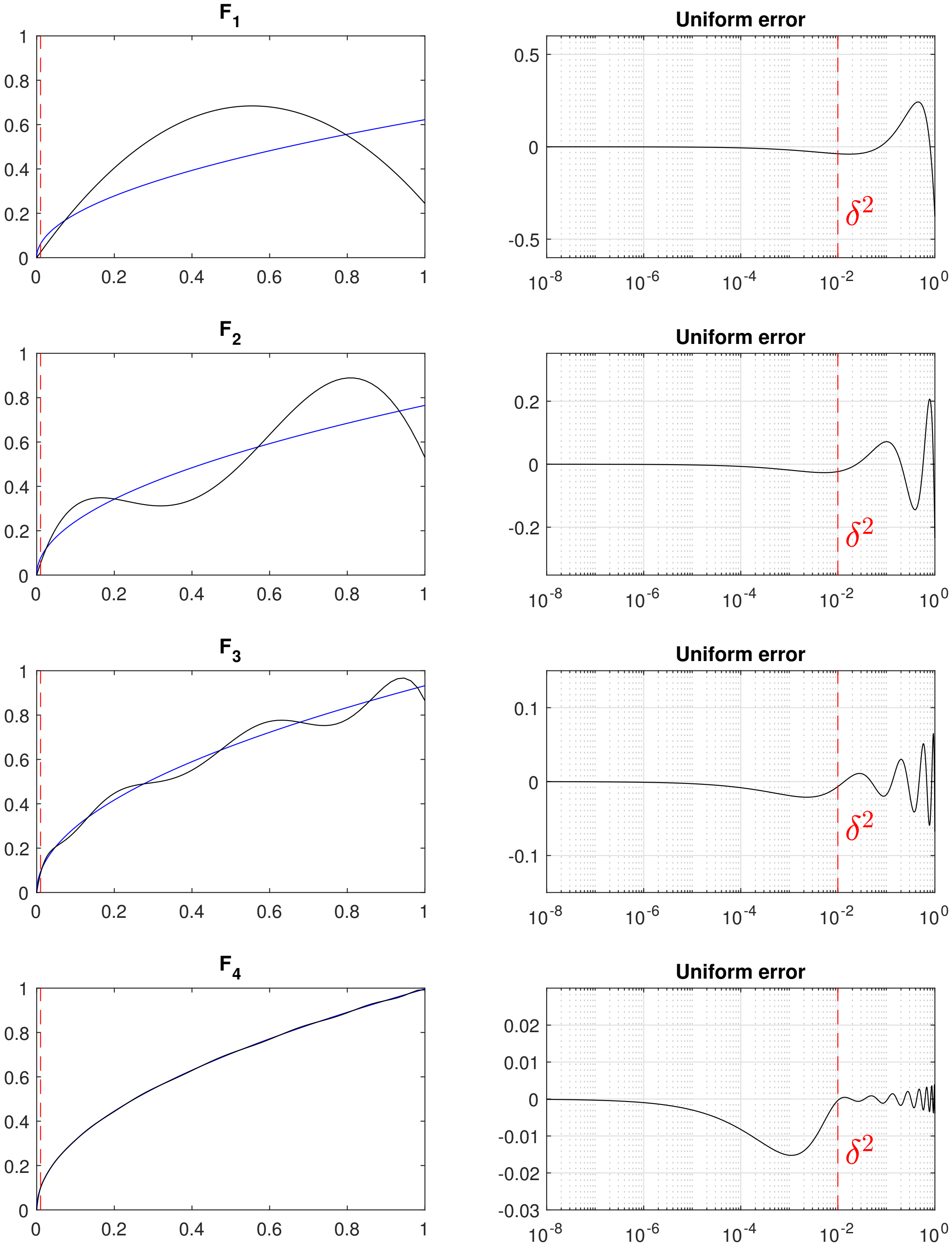

Figure 4.2: Composite polynomial approximations F_k (black) to $C_k\sqrt{x}$ (blue) on $[0,1]$, choosing $\delta = 0.1$ (red dashed line). The uniform errors are plotted beside each iteration. On $[\delta^2, 1]$, the relative error will be identical to the uniform error the scaled Newton-Schulz iterates to the sign function on $X(\delta)$.

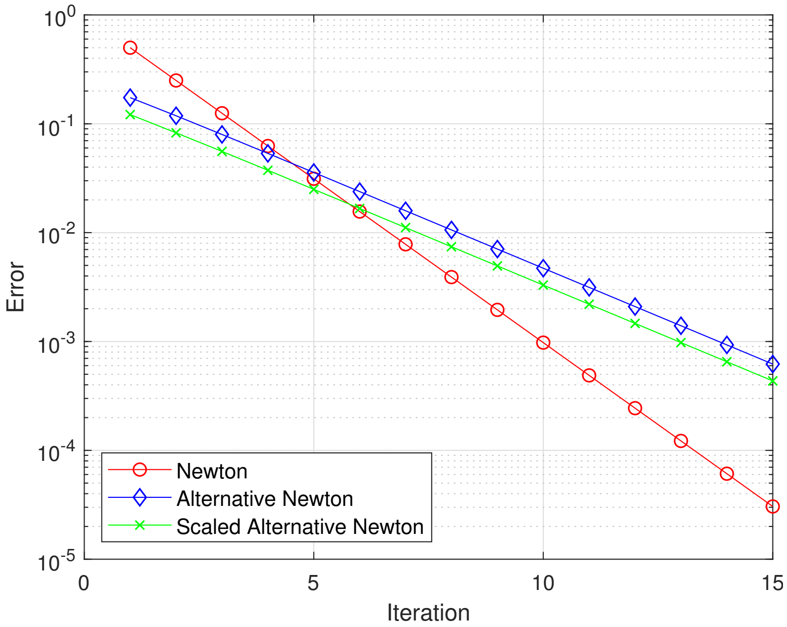

Figure 4.3: Comparison of the error in Newton, Alternative Newton and scaled Alternative Newton iterates to $\sqrt{x}$ on $[0, 1]$, with a choice of $\delta = 0.5$.

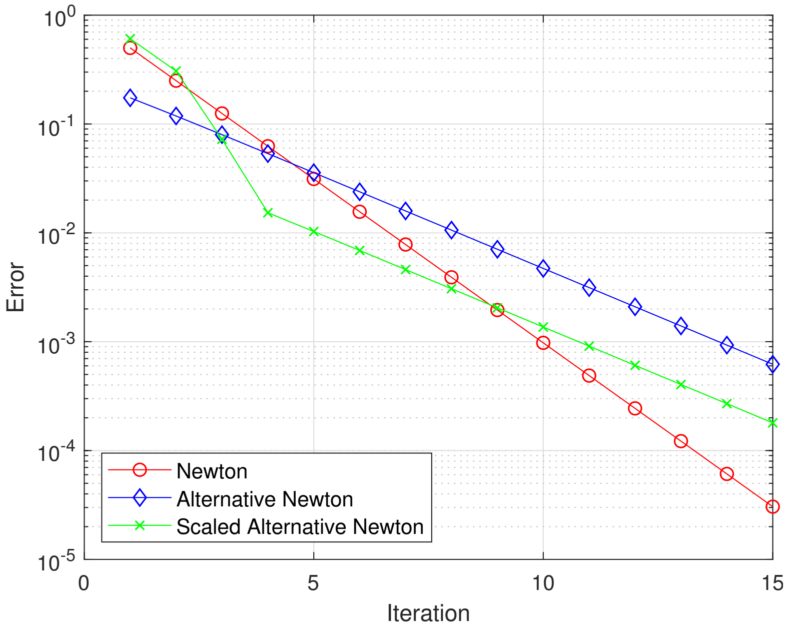

Figure 4.4: Comparison of the error in Newton, Alternative Newton and scaled Alternative Newton iterates to $\sqrt{x}$ on $[0, 1]$, with a choice of $\delta = 0.1$.

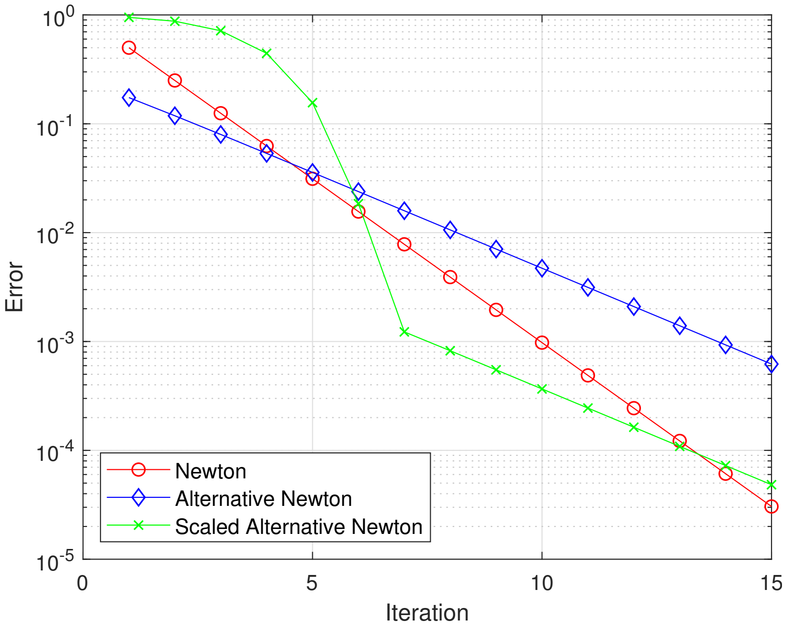

Figure 4.5: Comparison of the error in Newton, Alternative Newton and scaled Alternative Newton iterates to $\sqrt{x}$ on $[0, 1]$, with a choice of $\delta = 0.01$.

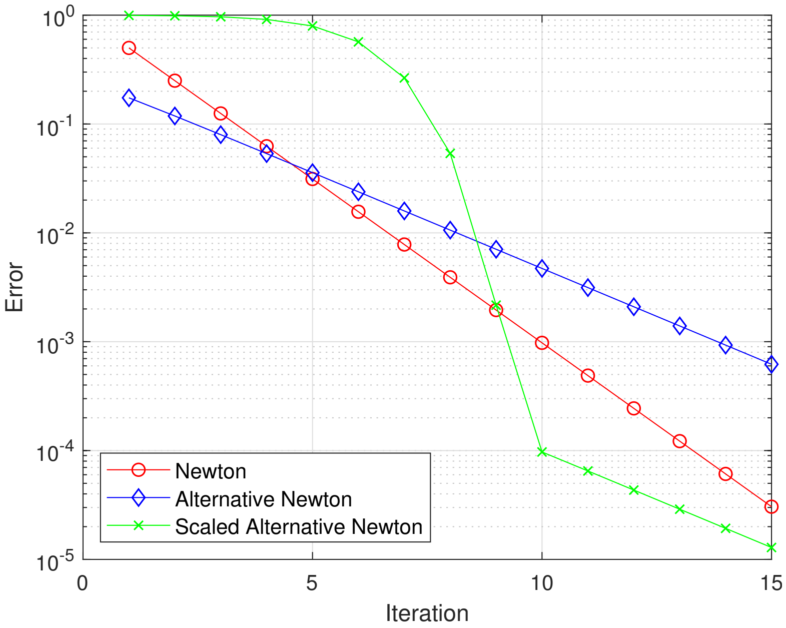

Figure 4.6: Comparison of the error in Newton, Alternative Newton and scaled Alternative Newton iterates to $\sqrt{x}$ on $[0, 1]$, with a choice of $\delta = 0.001$.

Chapter 5

Conclusion

Motivated by the recursive optimality property of the Zolotarev functions, we constructed a composite polynomial approximation to the scalar sign function on $[-1,-\delta]\cup[\delta,1]$ using a greedy iterative process. Although our construction was not optimal, we showed that it was equivalent to a scaled variant of the Newton-Schulz method for approximating the sign function, where the scaling at each iteration depended on the value of the extremal point of the iteration function. We noted that, when compared to the minimax approximation with respect to degrees of freedom, the scaled Newton-Schulz iteration was far superior. For a given δ, a lower bound number of iterations needed to obtain a given accuracy $\varepsilon > 0$ was derived, in a similar manner to the work of Nakatsukasa and Gawlik [8], and this bound was used to show that the converge of the scaled-Newton Schulz iteration is doubly exponential with respect to the degrees of freedom.

We also proved that the scaled Newton-Schulz iteration can be transformed into a composite approximation for the square root function whose error on $[\delta^2,1]$ is bounded by the error in the sign functino approximation; the iteration we derived produces polynomial iterates for any initial guess divisible by x. This approximation was shown to be equivalent and superior to a scaled variant of Newton's Method—one in which a different algebraic equation was initially considered, in order to avoid iterates dividing by their predecessor and hence generating rational approximations.

5.1 Where does this project lead?

Throughout this dissertation, many potential avenues were unlocked for further research. On multiple occasions, we drew attention to how composite polynomial and rational approximations are used to compute matrix functions, yet this is an area which we purposefully avoided to focus our work on gleaning results at a scalar level. The power and limitations of our constructed approximations when evaluated at a matrix argument are yet to be fully explored.

With regards to the scaled Newton-Schulz iteration, the following two questions were left unsolved. Firstly, while the iteration appears to perform well in the space of composite polynomials, it is unclear whether or not our construction is indeed the best approximation with respect to the space $\mathcal{P}^{\text{comp}}_{(k,3)}$. For example, can we find suboptimal iteration functions $\tilde{g}_i$ such that, when composed, $\tilde{g}_2(\tilde{g}_1(x))$ is a better approximation to the sign function than $g_2(g_1(x))$? Moreover, what if we considered greedy approximations comprising of higher-degree polynomials, such as $\mathcal{P}^{\text{comp}}_{(k,5)}$ or $\mathcal{P}^{\text{comp}}_{(k,7)}$? At what point does the computational cost arising from the degree of the composed polynomials cease to be optimal?

Finally, and most importantly, we wish to continue the analysis of the scaled Alternative Newton iteration to $\sqrt{x}$ beyond the submission of this dissertation. From the perspective of rational functions, we know that there is a composite approximation whose convergence is *doubly exponential* with respect to the degrees of freedom [8], and Figures 4.3-4.6 seem to suggest that convergence will be at least exponential in the polynomial setting.

Appendix A

Results for composite rational functions

This appendix will contain results relating to composite rational functions. As such, we will briefly discuss some definitions and examples—similar to those seen in the preliminary discussion in a polynomial setting—involving rational functions.

A.1 Composite rational functions and degrees of freedom

Definition A.1. The *degrees of freedom* of a rational function $r_{m,n} \in \mathcal{R}_{m,n}$ is the number of parameters required to completely determine $r_{m,n}$. We say that $r_{m,n}$ is a *plain rational function* if $r_{m,n}$ has $m+n+1$ degrees of freedom, i.e.

$$r_{m,n}(x) = \frac{\sum_{k=0}^{m} \alpha_k x^k}{1 + \sum_{k=1}^{n} \beta_k x^k},$$

for some independent scalars $\alpha_k, \beta_k \in \mathbb{R}$.

Definition A.2. A bivariate rational function $r(x, y)$ is said to be of type (m, n) if $r(x, x) \in \mathcal{R}_{m,n}$. A rational function $r(x)$ is said to be (k, m, n)-*composite* if

$$r(x) = r_k(x, r_{k-1}(x, r_{k-2}(\cdots (x, r_1(x, 1))))) \tag{A.1}$$

for bivariate rational functions $r_i(x, y)$, $i = 1, \ldots, k$, each of type (m, n). We say that a (k, m, n)-composite rational function is *pure* if the $r_i(x, y)$ in (A.1) are univariate, i.e. $r(x) = r_k(r_{k-1}(\cdots (r_1(x))))$.

Example A.3. A rational function of the form (A.1) has at most $k(m + n + 1)$ degrees of freedom. In particular, a pure composite rational function of the form

$r(x) = r_k(\cdots r_2(r_1(x)))$ with each r_i of type (n,n) is of type (n^k, n^k), but has $d \approx 2kn$ degrees of freedom.

Examples of composite rational functions we have encountered throughout this dissertation include the Newton iterates for the sign and square root functions, and more subtly the Zolotarev functions.

A.2 Composite rational approximation to $\sqrt{x}$

The motivation for the novel part of this dissertation, namely the construction of a composite polynomial approximation to $\sqrt{x}$, stemmed from the work of Gawlik and Nakatsukasa [8] concerning the composite rational approximation of $x^{1/p}$ on $[\alpha^p, 1]$ for some $\alpha \in (0,1)$. In this appendix, we present their results for $p = 2$, and simplify proofs where possible.

Theorem A.4. *There exists $N \in \mathbb{N}$ such that for every integer $n \geqslant N$, there is a rational function $r \in \mathcal{R}_{n,n-1}([0,1])$, which is a composition of $\lfloor \log_2 n \rfloor + 1$ type-$(2,1)$ rational functions, and a constant $C > 0$ such that*

$$\max_{x \in [0,1]} |r(x) - \sqrt{x}| \leqslant \exp(-C\sqrt{n}).$$

Theorem A.4 implies that the convergence of r to $\sqrt{x}$ is *doubly exponential* with respect to d, the number of degrees of freedom, as $n \to \infty$. That is, for constants $C_1, C_2 > 0$, the error is $O(\exp(-C_1 \exp(C_2 d)))$, where $d \approx 4 \log_2 n$. To prove Theorem A.4, we consider

$$f_{k+1}(x) = \frac{1}{2}\left(\sqrt{\alpha_k} f_k(x) + \frac{x}{\sqrt{\alpha_k} f_k(x)}\right), \qquad f_0(x) = 1, \tag{A.2}$$

where throughout this appendix we define the sequence (α_k) by

$$\alpha_{k+1} = \frac{2\sqrt{\alpha_k}}{1+\alpha_k}, \qquad \alpha_0 = \alpha.$$

Note that f_k is $(k,2,1)$-composite with

$$r_j(x,y) = \frac{1}{2}\left(\frac{\alpha_{j-1} y^2 + x}{\sqrt{\alpha_{j-1}} y}\right), \qquad j = 1, \ldots, k.$$

A simple induction shows that $f_k \in \mathcal{R}_{2^{k-1}, 2^{k-1}-1}$ for each $k \geqslant 1$. In what follows, we will approximate the square root function with the scaled functions

$$\tilde{f}_k(x) = \frac{2\alpha_k}{1+\alpha_k} f_k(x),$$

which in particular have the same number of degrees of freedom of the f_k. We will prove the following result, which will enable us to show that the $\tilde{f}_k$ converge to $\sqrt{x}$ at a doubly exponential rate with respect to the degrees of freedom.

Theorem A.5. *For any $k \in \mathbb{N}$, $\alpha \in (0,1)$, there holds*

$$\max_{x\in[0,1]} |\tilde{f}_k(x) - \sqrt{x}| \leqslant \max\left\{2\alpha, \frac{1-\alpha_k}{1+\alpha_k}\right\}.$$

A.2.1 Bounding the error on $[0,\alpha^2]$

To prove Theorem A.5, we consider the error over $[0,\alpha^2]$ and $[\alpha^2,1]$ separately. In this section, we consider the first interval and prove the following error bound.

Theorem A.6. *For any $k \in \mathbb{N}$, $\alpha \in (0,1)$, there holds*

$$\max_{x\in[0,\alpha^2]} |\tilde{f}_k(x) - \sqrt{x}| \leqslant 2\alpha.$$

To prove this, we need a few lemmas. We first define the functions

$$s_\alpha(x) := \frac{2x\sqrt{\alpha}}{\alpha + x^2}, \qquad H(\alpha) := s_\alpha(\alpha) = \frac{2\sqrt{\alpha}}{1+\alpha}, \qquad g_k(x) := \frac{x}{f_k(x^2)},$$

and note that

$$H(\alpha_k) = \alpha_{k+1}, \qquad g_{k+1}(x) = s_{\alpha_k}(g_k(x)). \tag{A.3}$$

Lemma A.7. *For any $\alpha \in (0,1)$, $x \in [0,\alpha]$, we have*

$$0 \leqslant xs'_\alpha(x) \leqslant s_\alpha(x) \leqslant H(\alpha).$$

Proof. Since s_α is nondecreasing on $[0,\alpha]$, we have

$$0 \leqslant xs'_\alpha(x), \qquad s_\alpha(x) \leqslant H(\alpha).$$

Furthermore, we can compute

$$xs'_\alpha(x) = \left(\frac{\alpha - x^2}{\alpha + x^2}\right)\left(\frac{2x\sqrt{\alpha}}{\alpha + x^2}\right) \leqslant s_\alpha(x)$$

since $0 \leqslant \frac{\alpha - x^2}{\alpha + x^2} \leqslant 1$ for all $x \in [0,\alpha] \subset [0,\sqrt{\alpha}]$. □

Lemma A.8. *For any $k \in \mathbb{N}$, $\alpha \in (0,1)$, $x \in [0,\alpha]$, we have*

$$0 \leqslant xg'_k(x) \leqslant g_k(x) \leqslant \alpha_k. \tag{A.4}$$

Proof. By induction. The case $k = 0$ is clear since $g_0(x) = x$ and $\alpha_0 = \alpha$. Now assume that (A.4) holds for $k = n$. By (A.3), we have

$$xg'_{n+1}(x) = xs'_{\alpha_n}(g_n(x))g'_n(x).$$

As $0 \leqslant g_n(x) \leqslant \alpha_n$ by the induction hypothesis, Lemma A.7 with $\alpha = \alpha_n$ implies that $xg'_{n+1}(x) \geqslant 0$ for $x \in [0, \alpha]$. The lemma further shows that

$$xg'_{n+1}(x) \leqslant s'_{\alpha_n}(g_n(x))g_n(x) \leqslant s_{\alpha_n}(g_n(x)) = g_{n+1}(x).$$

Finally, as noted in Lemma A.7,

$$g_{n+1}(x) = s_{\alpha_n}(g_n(x)) \leqslant H(\alpha_n) = \alpha_{n+1},$$

which completes the inductive step. □

Lemma A.9. *For any* $k \in \mathbb{N}$*,* $\alpha \in (0, 1)$*,* $x \in [0, \alpha^2]$*, we have*

$$0 < \tilde{f}_k(x) \leqslant \alpha(1 + \varepsilon_k), \qquad \textit{where} \qquad \varepsilon_k = \frac{1 - \alpha_k}{1 + \alpha_k}.$$

Proof. A simple induction shows that $f_k(x) > 0$ is non-decreasing on $[0, \alpha^2]$ for every k. Writing $f_k(x^2) = x/g_k(x)$, we differentiate to obtain

$$2xf'_k(x^2) = \frac{g_k(x) - xg'_k(x)}{g_k(x)^2} \geqslant 0$$

by Lemma A.8. Hence $f'_k(x^2) \geqslant 0$ for $x \in [0, \alpha^2]$. Setting (A.2) at $x = 0$ gives

$$f_{k+1}(0) = \frac{\sqrt{\alpha_k}}{2} f_k(0), \qquad f_0(0) = 1,$$

so $0 < f_k(0) \leqslant f_k(x)$ on $[0, \alpha^2]$ for every k, and so $0 < \tilde{f}_k(x) \leqslant \tilde{f}_k(\alpha^2)$ as $\tilde{f}_k$ is a positive multiple of f_k. The upper bound on $\tilde{f}_k$ is obtained by [6, Theorem 3.1], which provides the relative error bound

$$\max_{x \in [\alpha^2, 1]} \frac{\tilde{f}_k(x) - \sqrt{x}}{\sqrt{x}} = \frac{1 - \alpha_k}{1 + \alpha_k} = \varepsilon_k.$$

Taking $x = \alpha^2$ gives $\tilde{f}_k(\alpha^2) \leqslant \alpha(1 + \varepsilon_k)$, as required. □

We are now in a position to prove Theorem A.6, from which we can quickly deduce the result of Theorem A.5.

Proof of Theorem A.6. By Lemma A.9, we have for $x \in [0, \alpha^2]$

$$\begin{aligned} |\tilde{f}_k(x) - \sqrt{x}| &\leqslant \max\{|\tilde{f}_k(x)|, |\sqrt{x}|\} \\ &\leqslant \max\{\alpha(1+\varepsilon_k), \alpha\} \\ &= \alpha(1+\varepsilon_k) \\ &\leqslant 2\alpha, \end{aligned}$$

since $\varepsilon_k \in (0,1)$. Thus $\max_{x\in[0,\alpha^2]} |\tilde{f}_k(x) - \sqrt{x}| \leqslant 2\alpha$, as required. □

Proof of Theorem A.5. If we combine the result of [6, Theorem 3.1] with Theorem A.6, we obtain

$$\begin{aligned} \max_{x\in[0,1]} |\tilde{f}_k(x) - \sqrt{x}| &= \max\left\{2\alpha, \max_{x\in[\alpha^2,1]} |\tilde{f}_k(x) - \sqrt{x}|\right\} \\ &\leqslant \max\left\{2\alpha, \max_{x\in[\alpha^2,1]} \left|\frac{\tilde{f}_k(x) - \sqrt{x}}{\sqrt{x}}\right|\right\} \\ &= \max\left\{2\alpha, \frac{1-\alpha_k}{1+\alpha_k}\right\}, \end{aligned}$$

as required. □

A.2.2 Proof of Theorem A.4

Theorem A.5 is used in [8] as a basis for the analysis of the convergence of $\tilde{f}_k$ to $\sqrt{x}$ on $[0,1]$, and ultimately proving Theorem A.4. Nakatsukasa and Gawlik take a constructive approach to determining, for arbitrary $\varepsilon > 0$, values of α, k such that $\max_{x\in[0,1]} |\tilde{f}_k(x) - \sqrt{x}| < \varepsilon$. By Theorem A.5, we must have $\alpha \leqslant \varepsilon/2$ and k large enough such that $\varepsilon_k \leqslant \varepsilon$. We determine k in three steps:

(1) Find k_1 such that $\alpha_{k_1} \geqslant 1/e$;

(2) Select $\alpha^* \in (1/e, 1)$ and find k_2 such that $\alpha_{k_1+k_2} \geqslant \alpha^*$;

(3) Find k_3 such that $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$.

Then $k \geqslant k_1 + k_2 + k_3$. The choice of α^* in step 2 is given by δ^* in Lemma 3.4.

Step 1. To determine k_1 such that $\alpha_{k_1} \geqslant 1/e$, we first note that $\alpha_{k+1} > \sqrt{\alpha_k}$. To see this, we can argue by contradiction (a much simpler proof than that of Nakatsukasa and Gawlik). If $\alpha_{k+1} \leqslant \sqrt{\alpha_k}$, then

$$\frac{2\sqrt{\alpha_k}}{1+\alpha_k} \leqslant \sqrt{\alpha_k} \qquad \Longrightarrow \qquad 2\sqrt{\alpha_k} \leqslant \sqrt{\alpha_k} + \alpha_k\sqrt{\alpha_k} < 2\sqrt{\alpha_k},$$

since $\alpha_k \in (0,1)$, a contradiction. Hence $\alpha_{k+1} > \sqrt{\alpha_k}$ for all k, so

$$\alpha_k \geqslant \alpha_0^{(1/2)^k} = \alpha^{(1/2)^k}.$$

Thus we will have $\alpha_{k_1} \geqslant 1/e$ if $\alpha^{(1/2)^{k_1}} \geqslant 1/e$, namely

$$k_1 \geqslant \frac{\log\log\frac{1}{\alpha}}{\log 2}.$$

Since $\alpha \leqslant \varepsilon/2$ is assumed, we can write a lower bound in terms of ε, i.e.

$$k_1 \geqslant \frac{\log\log\frac{2}{\varepsilon}}{\log 2}.$$

Step 2. We pick α^* such that the result of Lemma 3.4 holds. Note that k_2 is independent of ε and α, hence constant.

Step 3. To determine k_3 such that $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$, we note that

$$\varepsilon_{k+1} = \frac{1-H(\alpha_k)}{1+H(\alpha_k)} \leqslant \left(\frac{1-\alpha^*}{1+\alpha^*}\right)^2$$

for $k \geqslant k_1 + k_2$ by Lemma 3.4. In particular,

$$\varepsilon_{k_1+k_2+k} \leqslant \varepsilon_{k_1+k_2}^{2^k} \leqslant \left(\frac{1-\alpha^*}{1+\alpha^*}\right)^{2^k},$$

so $\varepsilon_{k_1+k_2+k_3} \leqslant \varepsilon$ if k_3 is chosen such that

$$k_3 \geqslant \frac{\log\log\frac{1}{\varepsilon} - \log\log\frac{1+\alpha^*}{1-\alpha^*}}{\log 2}.$$

Combining the steps, we find that $\max_{x\in[0,1]} |\tilde{f}_k(x) - \sqrt{x}| < \varepsilon$ when

$$k \geqslant \frac{\log\log\frac{2}{\varepsilon}}{\log 2} + k_2 + \frac{\log\log\frac{1}{\varepsilon} - \log\log\frac{1+\alpha^*}{1-\alpha^*}}{\log 2}.$$

We now prove Theorem A.4, simplifying the version provided in [8].

Proof of Theorem A.4. As previously mentioned, $\tilde{f}_k$ will be a type $(2^{k-1}, 2^{k-1}-1)$ rational function. Define

$$\tilde{k}_2 = \left\lfloor k_2 - \frac{\log\log\frac{1+\alpha^*}{1-\alpha^*}}{\log 2} \right\rfloor,$$

so that the degree $n = 2^{k-1}$ of the iteration $\tilde{f}_k$ which approximates $\sqrt{x}$ with accuracy ε can be given by

$$\begin{aligned}\log n = (k-1)\log 2 &= \left(\frac{\log\log\frac{2}{\varepsilon} + \log\log\frac{1}{\varepsilon}}{\log 2} + \tilde{k}_2\right)\log 2 \\ &\leqslant \left(\frac{2\log\log\frac{2}{\varepsilon}}{\log 2} + \tilde{k}_2\right)\log 2,\end{aligned}$$

after absorbing the constant -1 into $\tilde{k}_2$. Hence

$$\log\log\frac{2}{\varepsilon} \geqslant \frac{\log n - \tilde{k}_2 \log 2}{2},$$

which further simplifies to give

$$\varepsilon \leqslant 2\exp(-C\sqrt{n}),$$

where $C = 2^{-\tilde{k}_2/2}$. This proves the theorem if the exact degree n is a sufficiently large power of 2; if $n \notin \{2^k : k \in \mathbb{N}\}$, then we note that $\lfloor \log_2 n \rfloor + 1$ iterations gives a rational function of type $(2^{\lfloor \log_2 n \rfloor}, 2^{\lfloor \log_2 n \rfloor} - 1)$, and for $n \geqslant N$, the error is bounded by

$$\begin{aligned}2\exp(-C(2^{\lfloor \log_2 n \rfloor})^{1/2}) &\leqslant 2\exp(-2^{-1/2}Cn^{1/2}) \\ &\leqslant \exp(-(2^{-1/2}C - N^{-1/2}\log 2)n^{1/2}).\end{aligned}$$

Taking N so that $2^{-1/2}C - N^{-1/2}\log 2 > 0$ proves the theorem. □

Acknowledgements

I would like to thank my supervisor, Prof Yuji Nakatsukasa, for his guidance and enthusiasm throughout the project. I would also like to thank Abinand Gopal for attending my project presentation via Skype at short notice, and his insightful questions.

Thanks to my family and friends for their continuous support. It has been a privilege to have spent time with you, despite things coming abruptly to an end after Hilary term. I hope to see you all soon.

References

[1] N. I. Akhiezer. *Theory of Approximation.* Dover Publications, 1992.

[2] N. I. Akhiezer and H. H. McFaden. *Elements of the Theory of Elliptic Functions.* Translations of Mathematical Monographs. American Mathematical Society, 1990.

[3] P. D. Brubeck, Y. Nakatsukasa and L. N. Trefethen. 'Vandermonde with Arnoldi' (2019). eprint: `1911.09988`.

[4] J. Chen and E. Chow. 'A Newton-Schulz Variant for Improving the Initial Convergence in Matrix Sign Computation'. *Preprint ANL/MCS-P5059-0114, Mathematics and Computer Science Division, Argonne National Laboratory* 60439 (2014).

[5] A. E. Eremenko and P. M. Yuditskii. 'Uniform approximation of $\operatorname{sgn}(x)$ by polynomials and entire functions'. *Journal d'Analyse Mathématique* 101.1 (2007), pp. 313–324.

[6] E. S. Gawlik. 'Rational Minimax Iterations for Computing the Matrix pth Root' (2019). eprint: `1903.06268`.

[7] E. S. Gawlik. 'Zolotarev Iterations for the Matrix Square Root'. *SIAM Journal on Matrix Analysis and Applications* 40.2 (2019), pp. 696–719.

[8] E. S. Gawlik and Y. Nakatsukasa. 'Approximating the pth Root by Composite Rational Functions' (2019). eprint: `1906.11326`.

[9] N. J. Higham. 'Functions of Matrices: Theory and Computation'. Society for Industrial and Applied Mathematics, 2008.

[10] N. J. Higham. 'Newton's Method for the Matrix Square Root'. *Mathematics of Computation* 46.174 (1986), pp. 537–549.

[11] N. J. Higham. 'Stable Iterations For The Matrix Square Root'. *Numerical Algorithms* 15 (1997), pp. 227–242.

[12] N. J. Higham and L. Lin. 'On pth roots of stochastic matrices'. *Linear Algebra and its Applications* 435.3 (2011), pp. 448–463.

[13] N. J. Higham and A. H. Al-Mohy. ‘Computing Matrix Functions’. *Acta Numerica* 19 (2010), pp. 159–208.

[14] N. J. Higham et al. ‘Functions Preserving Matrix Groups and Iterations for the Matrix Square Root’. *SIAM Journal on Matrix Analysis and Applications* 26 (2005).

[15] A. D. Kennedy. ‘Approximation Theory for Matrices’. *Nuclear Physics B - Proceedings Supplements* 128 (2004), pp. 107–116.

[16] C. S. Kenney and A. J. Laub. ‘Rational iterative methods for the matrix sign function’. *SIAM Journal on Matrix Analysis and Applications* 12.2 (1991), pp. 273–291.

[17] C. S. Kenney and A. J. Laub. ‘The Matrix Sign Function’. *IEEE Transactions on Automatic Control* 40.8 (1995), pp. 1330–1348.

[18] A. Krishnamoorthy and D. Menon. ‘Matrix Inversion Using Cholesky Decomposition’ (2011). eprint: `1111.4144`.

[19] Y. Nakatsukasa and R. W. Freund. ‘Computing Fundamental Matrix Decompositions Accurately via the Matrix Sign Function in Two Iterations: The Power of Zolotarev’s Functions’. *SIAM Review* 58.3 (2016), pp. 461–493.

[20] Y. Nakatsukasa and N. J. Higham. ‘Stable and Efficient Spectral Divide and Conquer Algorithms for the Symmetric Eigenvalue Decomposition and the SVD’. *SIAM Journal on Scientific Computing* 35.3 (2013), A1325–A1349.

[21] I. Ninomiya. ‘Best rational starting approximations and improved Newton iteration for the square root’. *Mathematics of Computation* 24.110 (1970), pp. 391–404.

[22] J. Rickards. ‘When Is a Polynomial a Composition of Other Polynomials?’ *The American Mathematical Monthly* 118.4 (2011), pp. 358–363.

[23] J.D. Roberts. ‘Linear model reduction and solution of the algebraic Riccati equation by use of the sign function’. *International Journal of Control* 32.4 (1980), pp. 677–687.

[24] H. Rutishauser. ‘Betrachtungen zur Quadratwurzeliteration’. *Monatshefte für Mathematik* 67 (1963), pp. 452–464.

[25] G. Schulz. ‘Iterative berechung der reziproken matrix’. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 13.1 (1933), pp. 57–59.

[26] F. Soleymani et al. ‘Approximating the matrix sign function using a novel iterative method’. *Abstract and Applied Analysis*. 2014.

[27] F. Soleymani et al. ‘Some matrix iterations for computing matrix sign function’. *Journal of Applied Mathematics* (2014).

[28] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003.

[29] L. N. Trefethen. ‘Approximation Theory and Approximation Practice’. Society for Industrial and Applied Mathematics, 2013.

[30] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

[31] E. I. Zolotarev. ‘Applications of elliptic functions to problems of functions deviating least and most from zero’. *Zapiski St-Petersburg Akad.* 30 (1877), pp. 1–59.