THE TEACHING OF MATHEMATICS

2022, Vol. XXV, 1, pp. 38–52

ROOT FINDING TECHNIQUES THAT WORK

Aaron Melman

Abstract. Several general techniques are described to incorporate the specific structure or properties of a nonlinear equation into a method for solving it. This can mean the construction of a method specifically tailored to the equation, or the transformation of the equation into an equivalent one for which an existing method is well-suited. The techniques are illustrated with the help of several case studies taken from the literature.

MathEduc Subject Classification: N45

AMS Subject Classification: 97N40

Key words and phrases: Nonlinear equation; transformation; multiplier; approximation.

1. Introduction

The numerical solution of a real nonlinear equation is frequently required in many areas of science and engineering, where it often occurs as an important subproblem that needs to be solved repeatedly. There exist many well-known standard methods to achieve this, typically iterative in nature, such as the well-known secant and Newton methods, and many others that can be found in any introductory numerical analysis textbook (Newton's method can even be found in high school calculus books). It is, in general, not easy to fully automate the solution of nonlinear equations because of the many problems that may arise, such as, e.g., singularities or tightly clustered roots, to name but a few. The best situations are those where the properties of the equation to be solved are well-known, as frequently occurs in specific applications. However, the inflexibility of standard methods often prevents efficient use of this information. As a general rule, even though it may sometimes be easier said than done, every attempt should be made to tailor a method to the specific equation at hand, rather than using a general purpose method. The result will be a faster and more accurate method. An alternative approach is to transform the equation into an equivalent one for which an already existing method is appropriate. Such considerations are unfortunately not commonly emphasized in textbooks.

We will address these issues first by showing that several standard methods can be derived from a simple general principle that, unlike these standard methods, has the flexibility to also generate methods that can take into account a particular problem's properties. Secondly, we consider techniques to transform a given equation into an equivalent one with more useful properties, such as those that guarantee the convergence of a particular method. Examples of the aforementioned principle

and techniques can be found scattered throughout the literature, and we will use several of them as illustrative case studies.

There are two main aspects to a numerical method: its construction and its subsequent convergence analysis. *Global* convergence concerns global conditions under which the method is guaranteed to converge. Newton's method, for example, converges from any point to the right of the root of a convex increasing function, as its iterates are the roots of the tangents, which lie below such a convex function. Similar conditions can be obtained for other standard methods, although they are usually more complicated. *Local* convergence of an iterative method is concerned with conditions that guarantee convergence from a point that is sufficiently close to the root and also with the asymptotic rate at which the iterates converge. Denoting by x_k the iterates and by x^* the root to which they converge, a method is said to be of order $q > 0$ if

$$\limsup_{k \to +\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^q} \leq C < +\infty \, .$$

For $q = 1$, it is required that $C < 1$. The higher the value of q, the faster the convergence. For example, when they converge to a simple root, i.e., when $f'(x^*) \neq 0$, then Newton's method [1, Ch. 2], [7, Ch. 3], [18, Ch. 5] is of second order, the secant method [1, Ch. 2], [7, Ch. 3], [18, Ch. 5] is of order $(1 + \sqrt{5})/2$, and Halley's method [16], [17] is of third order. Generally speaking, the more information (such as function and derivative values) is taken into account to construct the approximation the method is based on, the higher the order of convergence. An exhaustive treatment of methods and their convergence order can be found in [19].

2. Adaptive approximation and transformations

Consider the real nonlinear equation $f(x) = 0$ with $x \in \mathbb{R}$. To solve it, we formulate the following general adaptive approximation principle: approximate f by another function g that satisfies certain requirements and for which the equation $g(x) = 0$ is easy to solve. An approximation to the root of f is then given by the (appropriate) root of the approximation g. An iterative method follows naturally from this principle by approximating f at a current iterate and generating the next iterate as a solution of $g(x) = 0$. There are two choices to make: the function g and the way in which it approximates f. The approximation function is often (and naturally) required to mimic the behavior of the function it approximates as much as possible. However, different requirements are sometimes imposed, e.g., when the iterates need to be constrained in a certain way, as will be the case in Example 4.

As an example, we now show how three well-known methods follow from this simple principle. To avoid interrupting the exposition, we will implicitly assume that all expressions are valid, e.g, if a number appears in the denominator, it is assumed to be nonzero. The first method chooses a line as an appropriate approximation function g, i.e., $g(x) = \alpha + \beta x$, and the approximation conditions as requiring that g coincide with f to first order, i.e., in function and first derivative

values, at a given point $\bar{x}$, so that

$$\begin{cases} f(\bar{x}) = g(\bar{x}) = \alpha + \beta\bar{x} \\ f'(\bar{x}) = g'(\bar{x}) = \beta \ , \end{cases}$$

from which we obtain that $g(x) = f(\bar{x}) - \bar{x}f'(\bar{x}) + f'(\bar{x})x$. The next iterate is the root of g, given by $\bar{x} - f(\bar{x})/f'(\bar{x})$, which is exactly one step of Newton's famous method that generates iterates $\{x_k\}$, $k \in \mathbb{N}$, defined by the *iteration formula*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ .$$

Of course, this is not a surprise because the Newton iterate is often explained geometrically as the root of the tangent to f, i.e., the root of the linear approximation to f at a certain point, which is precisely g.

The second method also picks $g(x) = \alpha + \beta x$, but changes the approximation conditions to the requirement that f and g coincide in function value at $\bar{x}$ and at one other (distinct) point $\bar{y}$, i.e.,

$$\begin{cases} f(\bar{x}) = g(\bar{x}) = \alpha + \beta\bar{x} \\ f(\bar{y}) = g(\bar{y}) = \alpha + \beta\bar{y} \ , \end{cases}$$

which means that the function g is found by linear interpolation at $\bar{x}$ and $\bar{y}$. This set of linear equations is easily solved for α and β, so that the next iterate, obtained from $-\alpha/\beta$, is given by

$$\bar{x} - \frac{f(\bar{x})(\bar{x} - \bar{y})}{f(\bar{x}) - f(\bar{y})} \ ,$$

which turns out to be one step of the secant method, namely, a method that can be viewed as obtained from Newton's method by approximating the derivative by a finite difference. Its iteration formula is given by

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - y_k)}{f(x_k) - f(y_k)} \ .$$

The third method, Halley's method, the "method of osculating hyperbolae", can also be derived using the same principle we just illustrated (see, e.g., [16] and [17]). In this case, $g(x) = \alpha + \beta/(x - \gamma)$, and the approximation requires that f and g coincide up to second derivatives at a given point, resulting in the iteration formula

$$x_{k+1} = x_k - \frac{2f(x_k)f'(x_k)}{2(f'(x_k))^2 - f(x_k)f''(x_k)} \ .$$

Several more existing methods can be derived in this unifying way, but its main advantage lies in its flexibility. Consider, for example, the equation $f(x) = 0$, where f is of the form $f(x) = f_1(x) + f_2(x)$. Standard methods are typically defined by an iteration formula rigidly applied to all of f, based on an approximation function that does not necessarily bear any relation to f. Instead, a more efficient

method can often be obtained by approximating the functions f_1 and f_2 each in a different and more appropriate way.

However, it is not always possible or easy to find a convenient approximation. In such cases, one might attempt to adapt the problem to a certain method, by which is meant formulating an equivalent problem for which a certain method exhibits desired convergence properties. Such an equivalent problem might be obtained by a transformation of variables $x = \varphi(z)$ for a suitable function φ so that the equivalent problem becomes $f(\varphi(z)) = 0$. Another possibility is the use of a nonzero multiplier $\mu(x)$ that transforms the problem into $\mu(x)f(x) = 0$, which has the same solutions as $f(x) = 0$. Sometimes, a combination of techniques is called for. In the following section, we take a detailed look at concrete applications of these techniques.

3. Case studies

We present four examples, taken from the literature, to illustrate the techniques of the previous section for both the construction and analysis of methods for solving nonlinear equations. In the first example, a method is developed for solving an equation using adaptive approximation, where an approximation function is chosen to resemble the function that is being approximated.

Example 1 (Secular Equation - adaptive approximation). This example considers the solution of a so-called *secular equation* [5], which lies at the heart of the fast and widely used divide and conquer method from [4] to compute the eigenvalues of a symmetric matrix. After some simplification, this secular equation takes the form $f(x) = 0$, where

$$f(x) := 1 + \sum_{j=1}^{n} \frac{b_j}{d_j - x}\,, \tag{1}$$

with $b_j > 0$ for all j and $d_1 < d_2 < \cdots < d_{n-1} < d_n$. This function has n poles and n roots, one on each interval (d_j, d_{j+1}) for $j = 1, \ldots, n-1$, and one on $(d_n, +\infty)$. The value of n can be very large and the goal is to compute all of the roots quickly and accurately.

For the sake of this example, we concentrate on the ith root, with $1 \le i \le n-1$, after the origin is shifted to d_i, so that from here on we consider the computation of the unique root of f on the interval $(d_i, d_{i+1}) = (0, d_{i+1})$, shown in Figure 1 as the lower curve in thicker line.

Figure 1 illustrates a typical (although mild) situation, where a root's location is close to one of the singularities, as often happens in practice. A method such as Newton's method would generate iterates outside the interval from most starting points in $(0, d_{i+1})$, and trying to find a starting point to guarantee convergence might require a larger effort than the actual computation of the root. It is, in fact, not surprising that Newton's method would not be appropriate here as it is

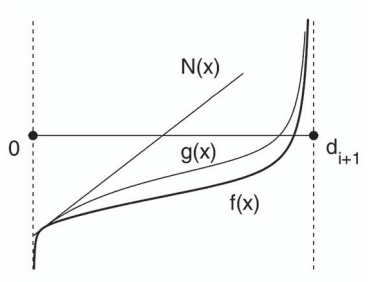

Fig.1. The functions $f(x)$, $g(x)$, and $N(x)$ on $(0, d_{i+1})$.

based on a linear approximation, while the function f is rational. We have made a comparison to Newton's method here simply because it is well-known, and similar reasoning applies to other standard methods.

Method construction. We can expect to obtain a more efficient method by using an approximation that more closely resembles the function f and distinguishes between different parts of f. There are several candidates for such an approximation, which were extensively studied in [3], [9], [11] and [12]. Here we will consider the approximation used in [3]. To do this, we first rewrite the function f as $f = 1 + f_1 + f_2$, with

$$f_1(x) := \sum_{j=1}^{i} \frac{b_j}{d_j - x} \quad \text{and} \quad f_2(x) := \sum_{j=i+1}^{n} \frac{b_j}{d_j - x} \ ,$$

and construct an iterative numerical method by separately approximating f_1 and f_2 by g_1 and g_2, respectively, so that, at a given point, g_1 agrees with f_1 in function and first derivative values, while g_2 does the same for f_2. The approximation g to f is then defined as $g := 1 + g_1 + g_2$. One possible choice, as in [3], is to use the rational functions

$$g_1(x) := \frac{\alpha}{\beta - x} \quad \text{and} \quad g_2(x) := \gamma + \frac{\delta}{d_{i+1} - x} \ .$$

In other words, at a given point $\bar{x}$, the following conditions must be satisfied:

$$\text{(2)} \qquad \begin{cases} f_1(\bar{x}) = g_1(\bar{x}) \\ f_1'(\bar{x}) = g_1'(\bar{x}) \end{cases} \quad \text{and} \quad \begin{cases} f_2(\bar{x}) = g_2(z) \\ f_2'(\bar{x}) = g_2'(\bar{x}) \ . \end{cases}$$

The conditions in (2), which require the same computational effort as Newton's method (function and derivative values), determine the parameters α, β, γ, and δ, which, in turn, define the functions g_1 and g_2. The approximation g is well defined on the interval $(0, d_{i+1})$ since the approximation conditions yield

$$\beta = \bar{x} + \frac{f_1(\bar{x})}{f_1'(\bar{x})} = \frac{\bar{x} f_1'(\bar{x}) + f_1(\bar{x})}{f_1'(\bar{x})} = \frac{1}{f_1'(\bar{x})} \sum_{j=1}^{i} \frac{b_j d_j}{(d_j - \bar{x})^2} < 0 \ .$$

Moreover, $\alpha = (\beta - \bar{x})^2 f_1'(\bar{x}) > 0$ and $\delta = (d_{i+1} - \bar{x})^2 f_2'(\bar{x}) > 0$, so that $g'(x) > 0$ on the interval, implying that g is strictly increasing (to $+\infty$). It must therefore have a unique root on $(0, d_{i+1})$ if $\bar{x}$ is chosen such that $f(\bar{x}) = g(\bar{x}) < 0$. The next iterate of a method based on this approximation is the root of g in the interval $(0, d_{i+1})$, which is obtained as the appropriate solution of a simple quadratic, since

$$1+\frac{\alpha}{\beta - x}+\gamma+\frac{\delta}{d_{i+1} - x} = 0 \Longrightarrow (1+\gamma)(\beta-x)(d_{i+1}-x)+\alpha(d_{i+1}-x)+\delta(\beta-x) = 0 \; .$$

The approximation at a point $\bar{x}$, deliberately chosen to be far from the root, is shown in Figure 1 in a thinner line just above the function f. For comparison, we have also shown the tangent $N(x)$ at the same point $\bar{x}$, which clearly demonstrates the advantage of a rational approximation.

Convergence properties. To address the global convergence of the method just described, we now first show that g dominates f on $(0, d_{i+1})$. Since $g_1(x) = \alpha/(\beta - x)$ approximates $f_1(x)$ to first order at $\bar{x} \in (0, d_{i+1})$, $(\beta - x)/\alpha$ approximates $1/f_1(x)$ to first order at $\bar{x}$, i.e., $(\beta - x)/\alpha$ is the linear approximation to $1/f_1$ at $\bar{x}$. A straightforward calculation yields

$$\left(\frac{1}{f_1}\right)'' = \frac{2(f_1')^2 - f_1 f_1''}{f_1^3} = \frac{-2(-f_1')^2 + (-f_1)(-f_1'')}{(-f_1)^3} \; . \tag{3}$$

The function $-f_1$ is positive on the interval $(0, d_{i+1})$ and it satisfies the conditions of Lemma 2.3 in [11] for $\rho = -1$, a parameter in that lemma, which in this case states that $-2(-f_1')^2 + (-f_1)(-f_1'') \geq 0$. It then follows from (3) that $1/f_1$ is a convex function on $(0, d_{i+1})$, so that it dominates its linear approximation at any point in that interval. As a result,

$$\frac{\beta - x}{\alpha} \leq \frac{1}{f_1(x)} \Longrightarrow \frac{\alpha}{\beta - x} \geq f_1(x) \; ,$$

i.e., $g_1(x) \geq f_1(x)$. We also have that $g_2(x) = \gamma + \delta/(d_{i+1} - x)$ approximates f_2 to first order at $\bar{x}$, which means that $\gamma(d_{i+1} - x) + \delta$ is the linear approximation of $(d_{i+1} - x) f_2(x)$. Some algebra yields

$$(d_{i+1} - x) f_2(x) = \left(\sum_{j=i+1}^{n} b_j \right) - \sum_{j=i+1}^{n} \frac{b_j (d_j - d_{i+1})}{d_j - x} \; ,$$

which is a concave function because $b_j(d_j - d_{i+1}) \geq 0$ when $j \geq i + 1$. This means that it is dominated by its linear approximation, leading to

$$\gamma(d_{i+1} - x) + \delta \geq (d_{i+1} - x) f_2(x) \Longrightarrow g_2(x) = \gamma + \frac{\delta}{d_{i+1} - x} \geq f_2(x) \; .$$

As a result, we obtain that $g(x) \geq f(x)$ on $(0, d_{i+1})$. Consequently, if $\bar{x}$ lies to the left of the root, then $f(\bar{x}) = g(\bar{x}) < 0$, implying that the unique root of g, which necessarily lies to the right of $\bar{x}$, also lies to the left of the root of f, ensuring

monotonic convergence. Moreover, it was shown in [11] that the convergence order is quadratic.

We conclude by deriving a starting point $x_0 \in (0, d_{i+1})$ satisfying $f(x_0) < 0$. Such a point can be found by observing that

$$1 + \sum_{j=1}^{i-1} \frac{b_j}{d_j - d_{i+1}} - \frac{b_i}{x} + \frac{b_{i+1}}{d_{i+1} - x} + \sum_{j=i+2}^{n} \frac{b_j}{d_j - d_{i+1}} \geq f(x) \ ,$$

so that initial point x_0 can be found as the root in $(0, d_{i+1})$ of the strictly increasing function

$$\left(1 + \sum_{\substack{j=1 \\ j \neq i, i+1}}^{n} \frac{b_j}{d_j - d_{i+1}}\right) - \frac{b_i}{x} + \frac{b_{i+1}}{d_{i+1} - x} \ ,$$

which is obtained as the appropriate root of a quadratic.

An additional consideration in the case of a parallel computation of the n roots of f is that the efficiency of such an implementation is determined by the root that requires the most time, so that uniform performance of the method is also important. In practice, the more an approximation resembles the function, the better this requirement will be fulfilled.

In the following example, the equation from Example 1 is solved in a different way, namely, by first carrying out a transformation of variables before applying adaptive approximation with a function having similar properties as the one being approximated.

Example 2 (Secular Equation – transformation and adaptive approximation). In Example 1 a method was derived to compute the root of f, defined in (1), on the interval $(0, d_{i+1})$ by using rational approximations. Here, the approach from [11, Section 3.3] is used, which consists of transforming the variable to obtain an equivalent, but more convenient, equation.

Method Construction. A relatively natural idea is to try and mitigate the effect of the singularities at the endpoints of the interval. A transformation that then suggests itself [11], is to set $x = 1/z$, which eliminates the singularity at the origin by sending it to infinity, so that the original interval is mapped to $(1/d_{i+1}, +\infty)$, and the problem becomes the computation of the (necessarily) unique root of $f(1/z) = 0$ on this interval. For convenience, we define $F(z) = f(1/z)$. So far, this is an idea that seems reasonable, but we still need to show that the resulting transformed equation $F(z) = 0$ exhibits properties that make it easier to solve than $f(x) = 0$. Straightforward algebra shows that

$$F(z) = 1 + \sum_{j=1,\, j\neq i}^{n} \frac{b_j}{d_j} - b_i z + \sum_{j=1,\, j\neq i}^{n} \frac{b_j/d_j^2}{z - 1/d_j} \ .$$

All the negative singularities of F lie in the interval $(1/d_{i-1}, 0)$ and all its positive singularities lie in the interval $(0, 1/d_{i+1})$. Its first and second derivatives show that

that F is strictly decreasing and convex on the interval $(1/d_{i+1}, +\infty)$, with

$$\lim_{z\to(1/d_{i+1})^+} F(z) = +\infty \qquad \text{and} \qquad \lim_{z\to+\infty} F(z) = -\infty \ .$$

Because F is convex, it dominates its linear approximation, so that Newton's method converges monotonically from any starting point between $1/d_{i+1}$ and the root. Figure 2 shows the function $F(z)$ (top curve in thicker line), along with its tangent $N(z)$ at a particular point. The root of the tangent is the next Newton iterate from that point.

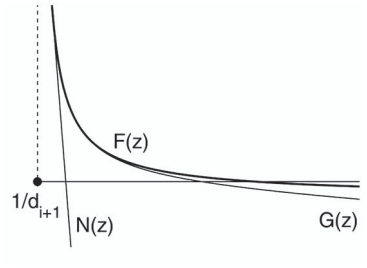

Fig. 2. The functions $F(z)$, $G(z)$, and $N(z)$ on $(1/d_{i+1}, +\infty)$.

The transformation of variables $x = 1/z$ has transformed the original equation into one for which Newton's method is guaranteed to converge, which was not the case for the original equation, as we saw in Example 1. However, the equation still involves a rational function that should preferably be approximated by a rational, rather than a linear, function. To construct such an approximation, we observe that the behavior of F on $(1/d_{i+1}, +\infty)$ is significantly affected by its singularity at $1/d_{i+1}$, which indicates that, if possible, the singularity should be included in the approximation. Writing F as

$$F(z) = F_1(z) + \frac{b_{i+1}/d_{i+1}^2}{z - 1/d_{i+1}} \ ,$$

we approximate it at a point $\bar{x}$ by

$$G(z) = \alpha + \beta z + \frac{b_{i+1}/d_{i+1}^2}{z - 1/d_{i+1}} \ ,$$

where $\alpha + \beta z$ is the linear approximation of F_1 at $\bar{x}$. Since $\beta = F_1'(\bar{x}) < 0$ for any $\bar{x} \in (1/d_{i+1}, +\infty)$, G is strictly decreasing, while it becomes unbounded as $z \to (1/d_{i+1})^+$, implying that it has a unique root on $(1/d_{i+1}, +\infty)$. The next iterate of a method based on this approximation is therefore the root of G, which is found as the appropriate root of the quadratic $(z - 1/d_{i+1})(\alpha + \beta z) + b_{i+1}/d_{i+1}^2$.

Convergence properties. Since, like F, F_1 is convex, $\alpha + \beta z$ will be dominated by F_1, and the entire approximation will therefore be dominated by F. As a result, the convergence of a method based on this approximation will be, like Newton's

method, monotonic from any starting point between $1/d_{i+1}$ and the root of F. Unlike Newton's method, it is globally convergent from any point in $(1/d_{i+1}, +\infty)$: if the initial iterate lies to the right of the root of F, the second iterate will lie between $1/d_{i+1}$ and the root. Moreover, because the approximation resembles F more closely than a line, we expect such a method to perform better than Newton's method, while requiring the same computational effort, namely, the computation of $F(\bar{z})$ and $F'(\bar{z})$. Figure 2 clearly shows this to be the case, where the nonlinear approximation (at the same point that was used for the linear approximation) is the curve just below $F(z)$, drawn in thinner line. Its root is a much better approximation to the root of F than the one obtained from Newton's method, which is the root of the linear approximation (tangent line). The quadratic convergence of such a method immediately follows from the quadratic convergence of Newton's method.

Similar arguments as in Example 1 show that a initial point to the right of the root of F is obtained as the root of

$$1 + \sum_{j=1,\, j\neq i}^{n} \frac{b_j}{d_j} + \sum_{j=1,\, j\neq i,i+1}^{n} \frac{b_j/d_j^2}{1/d_{i+1} - 1/d_j} - b_i z + \frac{b_{i+1}/d_{i+1}^2}{z - 1/d_{i+1}} \,,$$

which dominates F, while a point to the left of the root is obtained as the root of

$$1 + \sum_{j=1,j\neq i}^{n} \frac{b_j}{d_j} - b_i z + \frac{b_{i+1}/d_{i+1}^2}{z - 1/d_{i+1}} \,,$$

which is dominated by F. Both are computed as the appropriate root of a quadratic. An initial point can then be chosen as the average of these two points. In practice, the performance of the method in this example is similar to that of the method in Example 1.

We conclude with the observation that the method we obtained can easily be improved by including more terms of F in the approximation function G. For example, we could define

$$G(z) = \alpha + \beta z + \frac{b_{i+1}/d_{i+1}^2}{z - 1/d_{i+1}} + \frac{b_{i+2}/d_{i+2}^2}{z - 1/d_{i+2}} + \frac{b_{i+3}/d_{i+3}^2}{z - 1/d_{i+3}} + \frac{b_{i+4}/d_{i+4}^2}{z - 1/d_{i+4}} \,,$$

and compute its root with a method like the one just obtained. Such an approach could be advantageous for large values of n, as is often the case in practice.

In the next example a nonzero multiplier is used to transform the equation into one for which Newton's method exhibits global convergence.

Example 3 (Knapsack – multiplier). The equation we consider in this example is obtained in the course of solving the *nonlinear continuous knapsack problem* from [13]. In the original and simplest form of the (discrete and linear) knapsack problem, a knapsack of limited volume is filled with items that have different volumes and values, with the goal of maximizing the total value of the

items in the knapsack. It has been generalized to nonlinear continuous problems with many applications bearing no relation to knapsacks (see [6]), as in our case here, where the problem originated in the scheduling of the servicing of chemical processing units.

Before we state the equation to be solved, we define $h(x) = 1-(1+1/x)e^{-1/x}$ on $(0,+\infty)$ and its inverse function $\varphi(x) = h^{-1}(x)$, defined on $(0,1)$. We note that φ is well-defined since $h'(x) = -x^{-3}e^{-1/x} < 0$ for $x > 0$. Figure 3 shows the functions h and φ. The function φ will play a crucial rule, and to gain a better understanding of its properties, we make the following observations. Set $y = \varphi(x)$ to obtain

$$y = \varphi(x) \implies h(y) = x \implies h'(y)y' = 1 \implies y' = -y^3 e^{1/y} ,$$

from which it follows that $y'' = -y(3y-1)e^{1/y}y' = y^4(3y-1)e^{2/y}$. These calculations show that φ is a strictly decreasing function with a single inflection point at $h(1/3)$. Moreover,

$$\lim_{x\to 0^+} \varphi(x) = +\infty \text{ and } \lim_{x\to 1^-} \varphi'(x) = -\infty .$$

The function φ does not have an explicit functional expression: to compute $y = \varphi(x)$, one needs to compute the solution y of the nonlinear equation $h(y) = x$, which will be briefly addressed further on.

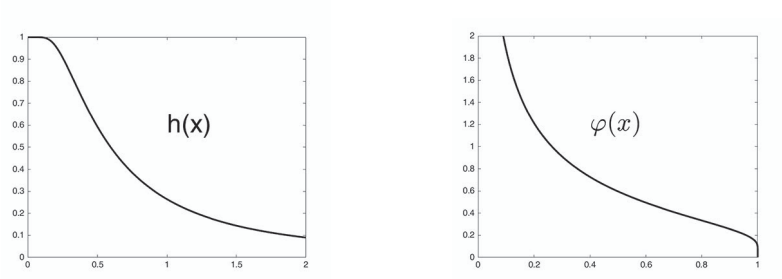

Fig. 3. The functions $h(x)$ and $\varphi(x)$.

We are now ready to state the problem to be solved, which is to compute the root of the function f, originating from the dual formulation of the problem, defined by

$$f(x) := \sum_{j=1}^{n} \alpha_j \varphi(\beta_j x) - K ,$$

where $\alpha_j, \beta_j, K > 0$, on the interval $(0,\gamma)$, with $\gamma := \min_j\{1/\beta_j\}$. Because it is a positive linear combination of scaled versions of φ, the properties of f are similar to those of φ : it is a strictly decreasing function, which becomes unbounded as $x \to 0^+$, and has an unbounded derivative when $x \to \gamma^-$. It is the curve in thicker line in both graphs of Figure 4. In what follows, we assume that $K > \lim_{x\to\gamma^-} f(x)$, implying that f has a unique root on $(0,\gamma)$.

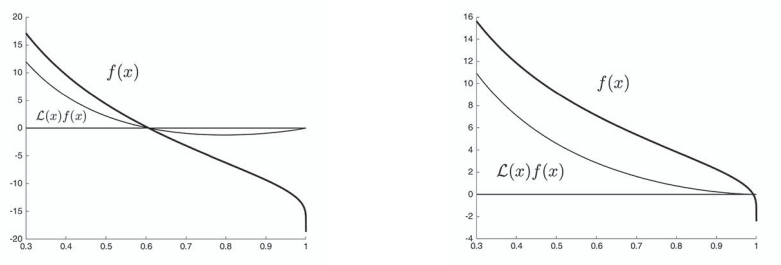

Fig. 4. The functions $f(x)$ and $\mathcal{L}(x)f(x)$.

Convexifying multiplier. The function f exhibits several difficulties: it does not have an explicit functional expression, and it does not have a simple shape, as it is neither convex nor concave over $(0, \gamma)$. The graph on the left in Figure 4 shows a mild situation, whereas, on the right, the root lies very close to γ, as inevitably happens in practice. In the latter case, a standard method (including Newton's method) can easily produce iterates that fall outside the interval, unless a starting point is available that already lies very close to the root. There does not seem to be an obvious appropriate approximation function that one could use, nor is a transformation of variables of much help. The best course of action may therefore be to try and use a multiplier to obtain a function with the same root, but with properties that make it easy to apply a standard method. In other words, we will adapt the problem to a method, rather than constructing a specific method for the problem. An easy choice for such a standard method would be Newton's method, which is simple, asymptotically fast, and guarantees convergence for convex and concave functions. The problem then becomes to find a nonzero multiplier $\mathcal{L}(x)$ such that the function $\mathcal{L}(x)f(x)$ is either convex or concave. In [13] it is shown that $\mathcal{L}f$ is convex on $(0, \gamma)$ if $\mathcal{L}$ satisfies a specific differential inequality. The simplest such multiplier is $\mathcal{L}(x) = 1 - \gamma^{-1}x$. We will not reproduce its derivation as it is rather technical and irrelevant for our purposes, but we will verify that $(1 - \gamma^{-1}x)f(x)$ is indeed convex on $(0, \gamma)$. To do this, we set $y = \varphi(\beta_j x)$, where $1 \leq j \leq n$, so that $h(y) = 1 - (1 + 1/y)e^{-1/y} = \beta_j x$, from which it follows that

$$e^{1/y} = \frac{y+1}{y(1-\beta_j x)}\,. \tag{4}$$

We now use the expressions for φ' and φ'' that were previously derived, the expression for $e^{1/y}$ from (4), and the fact that $\gamma^{-1} = \max_j\{\beta_j\}$, to obtain that

(5)
$$\begin{aligned}\big((1-\gamma^{-1}x)\varphi(\beta_j x)\big)'' &= -2\gamma^{-1}\beta_j\varphi'(\beta_j x) + \beta_j^2(1-\gamma^{-1}x)\varphi''(\beta_j x)\\ &= 2\gamma^{-1}\beta_j y^3 e^{1/y} + \beta_j^2(1-\gamma^{-1}x)y^4(3y-1)e^{2/y}\\ &= \beta_j y^3 e^{1/y}\big(2\gamma^{-1} + \beta_j(1-\gamma^{-1}x)y(3y-1)e^{1/y}\big)\\ &= \beta_j y^3 e^{1/y}\Big(2\gamma^{-1} + \frac{\beta_j(1-\gamma^{-1}x)}{1-\beta_j x}(3y-1)(y+1)\Big)\end{aligned}$$

$$= \beta_j y^3 e^{1/y}\Big(\frac{\beta_j(1-\gamma^{-1}x)}{1-\beta_j x}(3y^2+2y) + 2\gamma^{-1} - \Big(\frac{1-\gamma^{-1}x}{1-\beta_j x}\Big)\beta_j\Big)$$
$$\geq \beta_j y^3 e^{1/y}\Big(\frac{\beta_j(1-\gamma^{-1}x)}{1-\beta_j x}(3y^2+2y) + \gamma^{-1}\Big) > 0\ .$$

Since

$$(1-\gamma^{-1}x)f(x) = \sum_{j=1}^{n}\alpha_j(1-\gamma^{-1}x)\varphi(\beta_j x) - K(1-\gamma^{-1}x)$$

and $(1-\gamma^{-1}x)'' = 0$, we conclude with the help of (5) that $\mathcal{L}f$ is convex on $(0,\gamma)$. Figure 4 shows the function $(1-\gamma^{-1}x)f(x)$ in thinner line.

Convergence properties. The convexity of $\mathcal{L}f$ implies that if Newton's method is started from any point to the left of its root, then the iterates will converge monotonically to that root. Moreover, it is not hard to show that the root is simple, so that Newton's method converges quadratically.

A starting point to the left of the root of f can be found using the fact that, for $x > 0$, $\gamma^{-1}x \geq \beta_j x$, so that $\varphi(\gamma^{-1}x) \leq \varphi(\beta_j x)$, which in turn implies that

$$\sum_{j=1}^{n}\alpha_j\varphi(\beta_j x) - K \geq \sum_{j=1}^{n}\alpha_j\varphi(\gamma^{-1}x) - K\ . \tag{6}$$

The function in the right-hand side of (6) is strictly decreasing, becomes unbounded at the origin, and is negative as $x \to 1^-$. It must therefore have a unique root in $(0,1)$ and, since (6) shows that it is dominated by f, that root must lie to the left of the root of f. A starting point x_0 is therefore given by

$$x_0 = \gamma\, h\Big(\frac{K}{\sum_{j=1}^{n}\alpha_j}\Big)\ .$$

Although it is not the focus of this example, we conclude by briefly mentioning the function value computation of φ. From $y = \varphi(x)$, we obtain $h(y) = 1 - 1(1+1/y)e^{-1/y} = x$, so that y is the solution of a nonlinear equation on $(0,1)$. Setting $y = 1/z$ transforms this equation into $1-(1+z)e^{-z} - x = 0$, for which it was shown in [13] that Halley's method converges from any point in $(0,1)$.

In the following example, a transformation of variables is used to facilitate adaptive approximation. In this case, the approximation function is chosen to satisfy constraints on the iterates, rather than to resemble the function it approximates.

Example 4 (Pellet – transformation and adaptive approximation). In this example, we consider Pellet's theorem [10, Th. (2,8)] for a polynomial $p(z) = \sum_{j=0}^{n} a_j z^j$. It states that if, for some ℓ with $1 \leq \ell \leq n-1$, $a_\ell \neq 0$, the real polynomial $q(z) := |a_n|z^n + \cdots + |a_{\ell+1}|z^{\ell+1} - |a_\ell|z^\ell + |a_{\ell-1}|z^{\ell-1} + \cdots + |a_0|$ has two distinct positive roots, namely, the *Pellet ℓ-radii* ρ_1 and ρ_2, with $\rho_1 < \rho_2$, then p has exactly ℓ zeros in the closed disk $|z| \leq \rho_1$, and no zeros in the open annulus

$\rho_1 < |z| < \rho_2$. In other words, the theorem can sometimes detect gaps between the moduli of zeros. It is a direct consequence of Rouché's theorem [8, Theorem 1.6]. The theorem has been generalized to matrix polynomials [2], [14], leading to a real polynomial of the same kind as q. The graph of q has a form that is very similar to the lower curve on the left in Figure 5 in thicker line. It is of overriding importance for the equation $q(x) = 0$ to be solved by a method whose iterates converge to the solutions from the inside of the interval $[\rho_1, \rho_2]$, because, in such a case, one has the option to stop iterating at any moment and still have correct bounds. If, on the other hand, the iterates converge from outside the interval, none of the iterates provide correct bounds until they have fully converged.

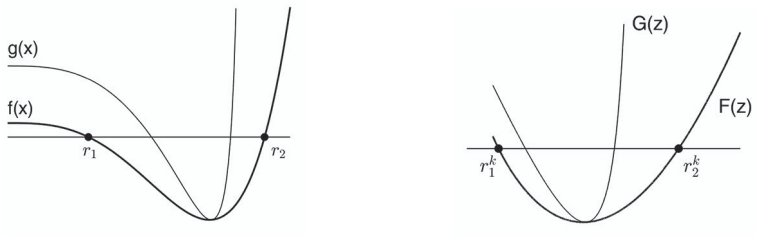

Fig. 5. The functions $f(x)$ and $F(z)$ with their respective approximations $g(x)$ and $G(z)$.

An efficient method to compute the Pellet radii from inside the interval was derived in [15], and consists of two phases. In the first phase the polynomial q is approximated by a trinomial with roots inside the interval, an application of using an adaptive approximation that resembles the function it approximates, while in the second phase, this trinomial is solved using adaptive approximation, where the approximation function is constructed to ensure that the iterates are properly constrained. Here we concentrate on the second phase, as it provides a good and uncomplicated example of adaptive approximation designed to satisfy specific requirements on the iterates.

The trinomial equation we need to solve is given by $f(x) := ax^n - bx^k + c = 0$, where $a, b, c > 0$, $n \geq 3$, and $1 \leq k \leq n-1$, under the assumption that f has two positive roots r_1 and r_2 with $r_1 < r_2$. The goal is to find an approximation to f that dominates it so that the approximation has roots in the interval $[r_1, r_2]$. To facilitate this, we use the transformation of variables $z = x^k$, which transforms f into $F(z) = f(z^{1/k}) = az^{n/k} - bz + c$. The function F is convex, so that using a linear approximation is not appropriate as it would be dominated by F. On the other hand, $z^{-n/k}$ is also convex, so that it dominates its linear approximation. Consequently, the reciprocal of this linear approximation then approximates $z^{n/k}$ to first order, and it dominates $z^{n/k}$, which is precisely what we need, as will soon become clear. This is the general idea that we now consider in more detail.

Method construction. The first order approximation of $z^{n/k}$ at $z = \bar{z}$ by $R(z) = \alpha/(\beta - z)$ is obtained by setting $R(\bar{z}) = \bar{z}^{n/k}$ and $R'(\bar{z}) = (n/k)\bar{z}^{n/k-1}$. A straightforward calculation shows that $\alpha = (k/n)\bar{z}^{1+n/k} > 0$ and $\beta = (1 + k/n)\bar{z} > \bar{z} > 0$.

Since $R(z)$ approximates $z^{n/k}$ to first order, $1/R(z)$, which is linear, approximates $z^{-n/k}$ to first order. Since $z^{-n/k}$ is convex, this means that

$$\frac{1}{R(z)} = \frac{\beta - z}{\alpha} \leq z^{-n/k} \Longrightarrow R(z) = \frac{\alpha}{\beta - z} \geq z^{n/k} .$$

As a result, we have obtained, with $G(z) = aR(z) - bz + c$, that $G(z) \geq F(z)$. If $F(\bar{z}) = G(\bar{z}) < 0$, then the approximation necessarily has roots in the interval $[r_1^k, r_2^k]$, since $G(0) > 0$ and $G(z) \to +\infty$ as $z \to \beta$. There are two such roots as they are the solution of a quadratic equation, and they become the next iterates. Figure 5 shows the trinomial $f(x)$ and its approximation $g(x) = G(x^k)$, as well as $F(z)$ and its approximation $G(z)$.

Convergence properties. A method based on the approximation G, starting from a point with negative function value, iterates with the smaller or larger root of G, to converge monotonically to r_1^k or r_2^k, respectively, from within the interval. It is a direct consequence of the domination of F by G. The order of convergence is quadratic [15].

A starting point z_0 is most conveniently computed as the minimum argument of F, obtained from

$$F'(z) = \frac{n}{k} a z^{n/k-1} - b = 0 \Longrightarrow z_0 = \left(\frac{kb}{na}\right)^{\frac{k}{n-k}} .$$

REFERENCES

[1] Acton, Forman S. *Numerical Methods That Work*, Updated and revised reprint of the 1970 edition. Mathematical Association of America, Washington, DC, 1990.

[2] Bini, D.A., Noferini, V., and Sharify, M. *Locating the eigenvalues of matrix polynomials*, SIAM J. Matrix Anal. Appl., **34** (2013), 1708–1727.

[3] Bunch, J.R., Nielsen, C.P., Sorensen, D.C. *Rank-one modification of the symmetric eigenproblem*, Numer. Math., **31** (1978), 31–48.

[4] Cuppen, J.J.M. *A divide and conquer method for the Symmetric Tridiagonal Eigenproblem*, Numer. Math., **36** (1981), 177-195.

[5] Golub, G.H. *Some modified matrix eigenvalue problems*, SIAM Rev., **15** (1973), 318–334.

[6] Hochbaum, D. *A nonlinear knapsack problem*, Oper. Res. Lett., **17** (1995), 103–110.

[7] Isaacson, E. and Keller, H.B. *Analysis of Numerical Methods*, Dover Publications, Inc., New York, 1994.

[8] Lang, S. *Complex Analysis*, Fourth edition, Graduate Texts in Mathematics, Springer-Verlag, New York, 1999.

[9] Li, R-C. *Solving secular equations stably and efficiently*, Tech. Report UCB//CSD-94-851, Computer Science Division, University of California, Berkeley, CA. Also: LAPACK Working Notes 89, 1994.

[10] Marden, M. *Geometry of Polynomials*, Second edition. Mathematical Surveys, No. 3, American Mathematical Society, Providence, R.I., 1966.

[11] Melman, A. *A unifying convergence analysis of numerical methods for secular equations*, Math. Comp., **66** (1997), 333–344.

[12] Melman, A. *Analysis of third-order methods for secular equations*, Math. Comp., **67** (1998), 271–286.

[13] Melman, A. and Rabinowitz, G. *An efficient method for a class of continuous nonlinear knapsack problems*, SIAM Rev., **42** (2000), 440–448.

[14] Melman, A. *Generalization and variations of Pellet's theorem for matrix polynomials*, Linear Algebra Appl. **439** (2013), 1550–1567.

[15] Melman, A. *Implementation of Pellet's theorem*, Numer. Algorithms, **65** (2014), 293–304.

[16] Salehov, G.S. *On the convergence of the process of tangent hyperbolas*, Dokl. Akad. Nauk SSSR, **82** (1952), 525–528 (in Russian).

[17] Scavo, T.R. and Thoo, B. *On the geometry of Halley's method*, Amer. Math. Monthly, **102** (1995), 417–426.

[18] Stoer, J. and Bulirsch, R., *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[19] Traub, J.F. *Iterative Methods for the Solution of Equations*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

Department of Applied Mathematics, Santa Clara University, Santa Clara, CA 95053

E-mail: amelman@scu.edu