

## A UNIFYING CONVERGENCE ANALYSIS OF SECOND-ORDER METHODS FOR SECULAR EQUATIONS

A. MELMAN

ABSTRACT. Existing numerical methods of second-order are considered for a so-called secular equation. We give a brief description of the most important of these methods and show that all of them can be interpreted as improvements of Newton's method for an equivalent problem for which Newton's method exhibits convergence from any point in a given interval. This interpretation unifies the convergence analysis of these methods, provides convergence proofs where they were lacking and furnishes ways to construct improved methods. In addition, we show that some of these methods are, in fact, equivalent. A second secular equation is also briefly considered.

### 1. INTRODUCTION

Modifying eigenvalue problems plays a role in several applications. Among these are, to name just a few, updating the singular value decomposition of matrices (e.g., [4]) and divide and conquer methods, based on the important paper by J.J.M. Cuppen ([6]), for the singular value decomposition of matrices ([1]) and for eigenproblems for symmetric matrices (e.g., [6, 7, 13, 16, 17, 18]). In these applications, a so-called secular equation appears ([14]). Related secular equations are found when solving constrained least squares type problems (e.g., [5, 9, 11, 12, 15, 19, 23, 24, 25]). A similar equation appears in invariant subspace computations ([10]), or when using the “escalator method” for computing the eigenvalues of a matrix ([8, pp. 183–192]). In this paper we consider some general approximation results and apply them to analyze numerical methods for the solution of two types of secular equations, the first of which is given by

$$(1) \quad 1 + \sigma \sum_{j=1}^n \frac{\zeta_j^2}{d_j - \lambda} = 0.$$

All quantities are assumed real. If the  $\zeta_j$ 's are all nonzero and the  $d_j$ 's are distinct, then this equation in  $\lambda$  has  $n$  solutions separated by the  $n$  values  $d_j$ . We assume without loss of generality that  $\sigma > 0$ . Should this not be the case, then  $d_j$  can be replaced by  $-d_{n-j+1}$  and  $\sigma$  by  $-\sigma$ . Such an equation is obtained, e.g., when computing the eigenvalues of the matrix  $D + \sigma z z^T$ , where  $z$  is a real vector with components  $\zeta_j$  and  $D$  a real diagonal matrix with diagonal entries  $d_j$ . Since equation (1) has poles, Newton's method for its solution is not appropriate and other

---

Received by the editor February 12, 1995 and, in revised form, November 13, 1995.

1991 *Mathematics Subject Classification*. Primary 65F15, 65H05.

*Key words and phrases*. Symmetric eigenvalues, secular equation, nonlinear approximation, global convergence.

methods, based on nonlinear first-order approximations have been proposed. The most widely used method seems to be the one in [3]. Recently, different and more efficient methods were developed in [21] and in [22]. These methods all exhibit a quadratic order of convergence.

We use a few basic results to show that all these methods can be interpreted as improvements of Newton's method for an equivalent problem for which Newton's method converges from any point in a given interval. This not only unifies the convergence analysis of these methods, it also provides convergence proofs where none existed before, shortens existing ones and furnishes ways of improving some of these methods. In particular, we show how an approximation, deemed inadequate in [21], can nevertheless be used to yield a better method. In addition, we show that some of the methods appearing in [21] are equivalent to methods in [22]. We do not go into implementation details, such as when certain methods should be preferred over others, or initial points and stopping rules, as this was rather extensively covered in [21]. We also do not consider higher-order methods such as, e.g., "Gragg's zero finder" in [2].

The second type of secular equation we consider is given by

$$(2) \quad \sum_{j=1}^n \frac{\zeta_j^2}{(\lambda - d_j)^2} - s^2 = 0,$$

where the unknown is, again,  $\lambda$ . All quantities are real and  $d_1 < d_2 < \dots < d_n$ . This equation appears, e.g., in [12] and [24]. We have largely used the same notation as in [12]. This equation needs to be solved for the smallest root, which lies to the left of  $d_1$ . A problem of this kind is encountered, e.g., when computing the smallest  $\lambda$  such that  $b^T(A - \lambda I)^{-2}b = \alpha^2$  is satisfied, with  $\alpha$  a real number,  $b$  a real vector and  $A$  a real symmetric matrix with known eigenvalues. We again use our basic results to prove the convergence of a method, proposed in [12, 24, 25]. Other types of secular equations can be treated similarly.

In §2 preliminary results are presented, which are used in §3 to obtain our main results concerning equation (1). In §4 we consider equation (2). All quantities in this paper are real.

## 2. PRELIMINARY RESULTS

**Lemma 2.1.** *Let  $f(t)$  be a strictly positive and twice continuously differentiable real function, defined on some interval  $I \subset \mathbb{R}$ . With  $\rho$  a nonzero integer, consider the real function of  $t$ :  $a(t+b)^\rho$ , where the parameters  $a$  and  $b$  are such that it interpolates  $f$  up to first order at a point  $\bar{t} \in I$  with  $f'(\bar{t}) \neq 0$ . Then  $a(t+b)^\rho = (L(t))^\rho$ , where  $L(t)$  denotes the linear Taylor approximation to  $f^{1/\rho}(t)$  at  $t = \bar{t}$ . Moreover, if*

$$\frac{1-\rho}{\rho} f'^2(t) + f(t)f''(t)$$

*is positive (negative) for all  $t \in I$ , then for all  $t$  such that  $a(t+b)^\rho \geq 0$ , the interpolant lies below (above) the function  $f$ .*

*Proof.* We obtain from the interpolation requirements that  $a(\bar{t}+b)^\rho = f(\bar{t})$ . Keeping in mind that  $f(\bar{t}) > 0$ , we therefore have that  $a^{1/\rho}$  is well defined. Now, since  $a(t+b)^\rho$  interpolates  $f(t)$  up to first order,  $\epsilon a^{1/\rho}(t+b)$  interpolates  $f^{1/\rho}(t)$  up to first order also, where, depending on the situation,  $\epsilon = \pm 1$  ( $\epsilon$  appears only for  $\rho$  even). This means that  $a(t+b)^\rho = (\epsilon a^{1/\rho}(t+b))^\rho = (L(t))^\rho$ , where  $L(t)$  denotes

the linear Taylor approximation to  $f^{1/\rho}(t)$  at  $t = \bar{t}$ . This proves the first part of the lemma. We now compute  $\left(f^{\frac{1}{\rho}}\right)''$  :

$$\left(f^{\frac{1}{\rho}}\right)'' = \left(\frac{1}{\rho}f^{\frac{1}{\rho}-1}f'\right)' = \frac{1}{\rho}f^{\frac{1}{\rho}-2}\left(\frac{1-\rho}{\rho}f'^2 + ff''\right) .$$

This means, for  $\rho > 0$ , that if for all  $t \in I$  :

$$\left(\frac{1-\rho}{\rho}f'^2(t) + f(t)f''(t)\right) \geq 0 ,$$

then  $f^{\frac{1}{\rho}}(t)$  is a convex function. The linear approximation at a point  $\bar{t}$  consequently lies below it, i.e.,  $L(t) \leq f^{\frac{1}{\rho}}(t)$ , and therefore, as long as  $L(t) \geq 0$ , we have  $(L(t))^\rho \leq f(t)$ . The opposite is true for a concave function. We proceed analogously for  $\rho < 0$ , where now  $L(t) \geq f^{\frac{1}{\rho}}(t)$  implies  $(L(t))^\rho \leq f(t)$ . This proves the lemma.  $\square$

*Remark.* When  $\rho$  is odd, Lemma 2.1 is easily adapted for negative functions as well.

*Notation.* When there can be no misunderstanding, we will denote the linear Taylor approximation to a function  $f$  at a point  $t = \bar{t}$  by  $L_f$ .

**Lemma 2.2.** *In this lemma, all interpolations are assumed to be with respect to the same point  $t = \bar{t}$ .*

(i) *If  $w(t)$  and  $f(t)$  are real and continuously differentiable functions of  $t$  with  $w(t)$  nonzero, and  $a$  and  $b$  are such that  $a + \frac{b}{w(\bar{t})}$  interpolates  $f(t)$  up to first order, then  $b + aw(t)$  interpolates  $w(t)f(t)$  up to first order.*

(ii) *The following holds for the linear Taylor approximations of continuously differentiable real functions  $f$  and  $g$ :*

$$L_{(f+g)} = L_f + L_g, \quad L_{fg} = L_fL_g = L_gL_f = L_{L_fL_g} .$$

(iii) *If  $L_{fg} = 1$ , then  $L_f \equiv L_{1/L_g}$ .*

*Proof.* The proof of parts (i) and (ii) is straightforward. For part (iii), we have  $L_{fL_g} = 1$  and the interpolation was carried out at the point  $t = \bar{t}$ , which means that  $f(\bar{t}) = (L_g(\bar{t}))^{-1}$ . It also means that  $(fL_g)'(\bar{t}) = 0$ . This gives

$$f'(\bar{t}) = -\frac{f(\bar{t})L'_g(\bar{t})}{L_g(\bar{t})} = \left(\frac{1}{L_g}\right)'(\bar{t}) .$$

This concludes the proof.  $\square$

**Lemma 2.3.** *The function*

$$f(t) = \sum_{j=1}^n \alpha_j(t + \beta_j)^\rho ,$$

*with  $\rho$  a nonzero integer and the  $\alpha_j$ 's nonnegative, satisfies*

$$\frac{1-\rho}{\rho}f'^2(t) + f(t)f''(t) \geq 0 ,$$

*for all  $t$  such that  $\forall j : t + \beta_j \geq 0$ .*

*Proof.* Let us first compute  $f'$  and  $f''$  :

$$f'(t) = \rho \sum_{j=1}^n \alpha_j (t + \beta_j)^{\rho-1}, \quad f''(t) = \rho(\rho-1) \sum_{j=1}^n \alpha_j (t + \beta_j)^{\rho-2}.$$

We then have

$$\begin{aligned} \frac{\rho-1}{\rho} f'^2(t) &= \rho(\rho-1) \left( \sum_{j=1}^n \alpha_j (t + \beta_j)^{\rho-1} \right)^2 \\ &= \rho(\rho-1) \left( \sum_{j=1}^n \sqrt{\alpha_j} (t + \beta_j)^{\frac{\rho}{2}} \sqrt{\alpha_j} (t + \beta_j)^{\frac{\rho}{2}-1} \right)^2. \end{aligned}$$

We note that  $\frac{1}{\rho}(\rho-1)$  and  $\rho(\rho-1)$  are nonnegative when  $\rho$  is a nonzero integer. Applying the Cauchy-Schwarz inequality yields

$$\frac{\rho-1}{\rho} f'^2(t) \leq \rho(\rho-1) \sum_{j=1}^n \left( \sqrt{\alpha_j} (t + \beta_j)^{\frac{\rho}{2}} \right)^2 \sum_{j=1}^n \left( \sqrt{\alpha_j} (t + \beta_j)^{\frac{\rho}{2}-1} \right)^2 = f(t) f''(t).$$

This completes the proof.  $\square$

### 3. SECULAR EQUATION 1

We now consider the secular equation from [3], which is equation (1) in our introduction. To compute the  $i$ th root ( $1 \leq i < n$ ), the transformation of variables  $\lambda = d_i + \sigma t$  is carried out, which, with  $\delta_j = \frac{d_j - d_i}{\sigma}$ , yields the following rootfinding problem on  $(0, \delta_{i+1})$  :

$$f(t) \triangleq 1 + \sum_{j=1}^i \frac{\zeta_j^2}{\delta_j - t} + \sum_{j=i+1}^n \frac{\zeta_j^2}{\delta_j - t} = 0,$$

where

$$\delta_1 < \cdots < \delta_i = 0 < \delta_{i+1} < \cdots < \delta_n.$$

This transformation is not essential to our results. On the interval  $(0, \delta_{i+1})$ ,  $f$  is monotonically increasing and has simple poles at  $t = 0$  and  $t = \delta_{i+1}$ . In what follows,  $\delta_{i+1}$  will be denoted by  $\delta$ . Note that  $i$  is fixed and  $1 \leq i < n$ . We do not consider the case  $i = n$ , as it can be treated analogously and is simpler (there is only one pole). We define, as in [3],

$$\psi(t) = \sum_{j=1}^i \frac{\zeta_j^2}{\delta_j - t}, \quad \phi(t) = \sum_{j=i+1}^n \frac{\zeta_j^2}{\delta_j - t}.$$

The following easily verified properties hold for all  $t \in (0, \delta)$  :

- (1)  $\psi(t) < 0, \quad \psi'(t) > 0, \quad \psi''(t) < 0$  ;
- (2)  $\phi(t) > 0, \quad \phi'(t) > 0, \quad \phi''(t) > 0$  ;
- (3)  $((\delta - t)\psi(t))'' = (\delta - t)\psi''(t) - 2\psi'(t) < 0$  ;
- (4)  $((\delta - t)\phi(t))'' = \left( \sum_{j=i+1}^n \frac{(\delta - t)\zeta_j^2}{\delta_j - t} \right)'' = \left( \sum_{j=i+1}^n \left( \zeta_j^2 + \frac{(\delta - \delta_j)\zeta_j^2}{\delta_j - t} \right) \right)'' < 0$  ;
- (5)  $(t\phi(t))'' = t\phi''(t) + 2\phi'(t) > 0$  ;
- (6)  $(t\psi(t))'' = \left( \sum_{j=1}^i \frac{t\zeta_j^2}{\delta_j - t} \right)'' = \left( \sum_{j=1}^i \left( -\zeta_j^2 + \frac{\delta_j\zeta_j^2}{\delta_j - t} \right) \right)'' > 0$  ;
- (7)  $(\delta - t)f(t)$  and  $tf(t)$  are concave and convex functions, respectively .

*Notation.* Throughout the remainder of this section we will consider  $i$  fixed, and the solution on  $(0, \delta)$  of the resulting problem  $f(t) = 0$  will be denoted by  $t^*$ .

**3.1. The BNS1 and BNS2 methods.** The method by Bunch, Nielsen and Sorensen in [3], which will henceforth be called the “BNS1 method”, is based on local nonlinear approximations, where  $\psi(t)$  is interpolated up to first order by  $p(q - t)^{-1}$  and  $\phi(t)$  by  $r + s(\delta - t)^{-1}$ . The constants  $p, q, r$  and  $s$  are determined by the interpolation requirements at some point  $\bar{t} \in (0, t^*]$ , where  $f$  has a negative function value. The new iterate is then obtained by solving

$$1 + \frac{p}{q - t} + r + \frac{s}{\delta - t} = 0 .$$

It is straightforward to show that  $p, s > 0$  and  $q < 0$  (see also [3]). This means that the approximating function is strictly increasing on  $(0, \delta)$ . Since its value is negative at  $\bar{t}$  and tends to  $+\infty$  at  $\delta^-$ , it has exactly one root in  $(0, \delta)$ . The convergence as well as the quadratic order of convergence of this method were proved in [3].

An analogous method can be devised (see [21]) by interpolating  $\phi$  with a rational function of the type that was used to interpolate  $\psi$  and vice versa, i.e., the interpolant becomes

$$1 + \bar{r} + \frac{\bar{s}}{t} + \frac{\bar{p}}{\bar{q} - t} .$$

The method based on this approximation will be called the “BNS2 method”. Convergence results can be proved analogously to the the proof in [3] for the BNS1 method.

We now use the results from §2 to present an alternative, and quite a bit shorter, convergence proof. This proof’s distinctive feature is that it shows that both methods can be interpreted as an improved Newton method for an equivalent problem, for which Newton’s method converges from any point in a suitably chosen interval.

**Theorem 3.1.** *The BNS1 and BNS2 methods monotonically converge to the root  $t^*$  of  $f(t)$  on  $(0, \delta)$  at least as fast as Newton’s method applied to computing the (same) root on  $(0, \delta)$  of  $(\delta - t)f(t)$  and  $tf(t)$ , respectively, and starting from an initial point in  $(0, t^*]$  and  $[t^*, \delta)$ , respectively.*

*Proof.* The theorem will be proved for the BNS1 method. The proof for the BNS2 method is very similar and we will only outline it. We start by deriving a few

inequalities on  $(0, \delta)$ . From part (i) in Lemma 2.2 we know that if  $r + s(\delta - t)^{-1}$  interpolates  $\phi(t)$  up to first order at  $t = \bar{t}$ , then  $s + r(\delta - t)$  interpolates  $(\delta - t)\phi(t)$  up to first order, i.e.,

$$r + s(\delta - t)^{-1} = \frac{s + r(\delta - t)}{\delta - t} = \frac{L_{\Delta\phi}(t)}{\delta - t},$$

where  $\Delta(t) = \delta - t$ . From part (ii) in Lemma 2.2 we have that  $L_{\Delta\psi} \equiv L_{\Delta L_\psi}$ . We also have that  $((\delta - t)L_\psi(t))' = -2L'_\psi(t) = -2\psi'(\bar{t}) < 0$ , i.e., the function  $(\delta - t)L_\psi$  is concave. Therefore,

$$(3) \quad L_{\Delta\psi}(t) = L_{\Delta L_\psi}(t) \geq (\delta - t)L_\psi(t).$$

Applying part (iii) in Lemma 2.2 with  $f = \psi$  and  $g = 1/\psi$  yields  $L_\psi \equiv L_{(L_{1/\psi})^{-1}}$ . We also observe that Lemma 2.1 holds for  $-\psi$  with  $\rho = -1$ . Since  $p(t - q)^{-1}$  is the approximation to  $-\psi(t)$ , we have  $(L_{1/\psi}(t))^{-1} = p(q - t)^{-1}$ . As we mentioned before,  $p > 0$  and  $q < 0$ , which means that  $(L_{1/\psi}(t))^{-1}$  is concave on  $(0, \delta)$ . Therefore,

$$(4) \quad L_\psi(t) = L_{(L_{1/\psi})^{-1}}(t) \geq (L_{1/\psi}(t))^{-1}.$$

We now use (3) and (4) to obtain the following for  $t \in (0, \delta)$ :

$$(5) \quad \begin{aligned} L_{\Delta f}(t) &= \delta - t + L_{\Delta\psi}(t) + L_{\Delta\phi}(t) \geq \delta - t + (\delta - t)L_\psi(t) + L_{\Delta\phi}(t) \\ &\geq \delta - t + \frac{\delta - t}{L_{1/\psi}(t)} + L_{\Delta\phi}(t) = (\delta - t) \left( 1 + \frac{1}{L_{1/\psi}(t)} + \frac{L_{\Delta\phi}(t)}{\delta - t} \right). \end{aligned}$$

With  $\rho = -1$ , Lemma 2.1, together with Lemma 2.3, yields  $-(L_{1/\psi}(t))^{-1} \leq -\psi(t)$ . Also,  $(\delta - t)\phi(t)$  is a concave function, and therefore  $L_{\Delta\phi}(t) \geq (\delta - t)\phi(t)$ . As a consequence, inequality (5) implies for  $t \in (0, \delta)$

$$(6) \quad \frac{L_{\Delta f}(t)}{\delta - t} \geq 1 + \frac{1}{L_{1/\psi}(t)} + \frac{L_{\Delta\phi}(t)}{\delta - t} \geq f(t),$$

or

$$\frac{L_{\Delta f}(t)}{\delta - t} \geq 1 + \frac{p}{q - t} + r + \frac{s}{\delta - t} \geq f(t).$$

We observe that  $L_{\Delta f}(t)$  and  $\frac{L_{\Delta f}(t)}{\delta - t}$  have the same root and that  $f(t) < 0$  for  $t \in (0, t^*)$ , implying  $(\Delta f)'(t) = -f(t) + (\delta - t)f'(t) > 0$ . We also know from the definition of the method that the interpolating function has exactly one root in  $(0, \delta)$ . Therefore, the meaning of (6) is that this method is at least as fast as Newton's method for the equivalent (i.e., having the same root) problem  $(\delta - t)f(t) = 0$  on  $(0, \delta)$ . Since this is a concave and increasing function for  $t \in (0, t^*]$ , monotonic convergence and second-order convergence are immediate from any starting point in  $(0, t^*]$  (see, e.g., [20]). For the BNS2 method the appropriate functions are convex instead of concave (see properties (5) and (6)). This yields for  $t \in (0, \delta)$

$$(7) \quad \frac{L_{tf}(t)}{t} \leq 1 + \frac{L_{t\psi}(t)}{t} + \frac{1}{L_{1/\phi}(t)} \leq f(t),$$

or

$$\frac{L_{tf}(t)}{t} \leq 1 + \bar{r} + \frac{\bar{s}}{t} + \frac{\bar{p}}{\bar{q} - t} \leq f(t).$$

For  $t \in (t^*, \delta)$ :  $f(t) > 0$ , and therefore  $(tf)'(t) = f(t) + tf'(t) > 0$ . The same conclusions can be drawn as for the BNS1 method. This concludes the proof.  $\square$

*Remark.* We implicitly use the fact that one iteration of Newton’s method, in the situation that it is applied here, will yield a better approximation to the root the closer the starting point lies to that root. The reason for this is the convexity or concavity (whichever applies) of the functions involved. The same is true for the BNS methods and the fixed weight methods (which are still to come).

The BNS1 and BNS2 methods are called in [21] “approaching from the left” and “approaching from the right”, respectively. In [21], a method based on the BNS methods, “the middle way”, is also considered. It is based on interpolation of  $\psi(t)$  and  $\phi(t)$  by  $a + bt^{-1}$  and  $c + d(\delta - t)^{-1}$ , respectively. Using similar arguments as in the previous proof, it is not hard to show that

$$1 + \frac{L_{t\psi}(t)}{t} + \frac{1}{L_{1/\phi}(t)} \leq 1 + a + \frac{b}{t} + c + \frac{d}{\delta - t} \leq 1 + \frac{1}{L_{1/\psi}(t)} + \frac{L_{\Delta\phi}(t)}{\delta - t} .$$

We note that for this method convergence cannot be guaranteed unless the starting point lies close enough to the root. The second-order convergence then follows from the last inequality and the second-order convergence of the BNS methods.

**3.2. Fixed weight and hybrid methods.** The methods in this subsection are taken from [21], where they are stated without convergence proofs. We first briefly describe these methods before proving their convergence.

**(1) The fixed weight 1 method.** The function  $f(t)$  is interpolated up to first order at a point  $\bar{t}$  by an expression of the form

$$r + \frac{s}{\delta - t} - \frac{\zeta_i^2}{t} .$$

The next iterate is then found, as usual, by computing the root of the interpolant. It is straightforward to show that the interpolant is strictly increasing on  $(0, \delta)$  from  $-\infty$  to  $+\infty$ , regardless of which side of the root  $\bar{t}$  lies on, and therefore that it has a unique root in  $(0, \delta)$ . As we shall see, the iterates approach the root of  $f(t)$  from the left-hand side. We abbreviate this method as the “FW1 method”.

**(2) The fixed weight 2 method.** Here, the interpolant is given by a function of the form

$$\bar{r} + \frac{\bar{s}}{t} + \frac{\zeta_{i+1}^2}{\delta - t} .$$

This function has a unique root on  $(0, \delta)$  and now the iterates approach the root of  $f(t)$  from the right-hand side. We abbreviate this method as the “FW2 method”.

**(3) Hybrid methods.** The interpolating function in this case is arrived at by including more than two poles. For example, when three poles are included, the interpolant could take the form

$$\hat{r} + \frac{\hat{s}}{\delta - t} + \frac{\zeta_{i-1}^2}{\delta_{i-1} - t} - \frac{\zeta_i^2}{t} .$$

Again, the interpolant has a unique root on  $(0, \delta)$ . An alternative interpolant can be constructed in the same way as in the FW2 method.

We now proceed to prove the convergence of these methods. Let us define

$$\begin{aligned} \tilde{\psi}(t) &= \psi(t) + \frac{\zeta_i^2}{t} = \sum_{j=1}^{i-1} \frac{\zeta_j^2}{\delta_j - t}, & \hat{\phi}(t) &= \phi(t) - \frac{\zeta_{i+1}^2}{\delta - t} = \sum_{j=i+2}^n \frac{\zeta_j^2}{\delta_j - t}, \\ \tilde{f}(t) &= f(t) + \frac{\zeta_i^2}{t} = 1 + \tilde{\psi}(t) + \phi(t), & \hat{f}(t) &= f(t) - \frac{\zeta_{i+1}^2}{\delta - t} = 1 + \psi(t) + \hat{\phi}(t). \end{aligned}$$

The following theorem then shows the convergence of the fixed weight and hybrid methods.

**Theorem 3.2.** *The fixed weight and hybrid methods converge from any point in  $(0, \delta)$  to the solution  $t^*$  of  $f(t) = 0$  and their order of convergence is at least quadratic.*

*Proof.* We start with the FW1 method. Analogously to the proof of Theorem 3.1, we have the following inequalities, valid on  $(0, \delta)$  :

$$\begin{aligned} L_{\Delta f}(t) &= L_{\Delta \tilde{f}}(t) + L_{-\Delta \zeta_i^2/t}(t) \geq L_{\Delta \tilde{f}}(t) + (\delta - t) \frac{\zeta_i^2}{-t} \\ &\geq (\delta - t) \tilde{f}(t) + (\delta - t) \frac{\zeta_i^2}{-t} = (\delta - t) f(t). \end{aligned}$$

We have therefore obtained

$$(8) \quad \frac{L_{\Delta f}(t)}{\delta - t} \geq \frac{L_{\Delta \tilde{f}}(t)}{\delta - t} - \frac{\zeta_i^2}{t} \geq f(t),$$

or

$$\frac{L_{\Delta f}(t)}{\delta - t} \geq r + \frac{s}{\delta - t} - \frac{\zeta_i^2}{t} \geq f(t).$$

As in the proof of Theorem 3.1, this proves the convergence of the method starting from any point in  $(0, t^*]$ , along with the second-order convergence. Now, let us have a look at what happens when the starting point belongs to  $(t^*, \delta)$ . We know that the zero of the interpolant must lie in  $(0, \delta)$  and because (8) holds, this zero will lie to the left of the root. From here on, the iterates remain in  $(0, t^*]$ , and convergence is monotonic.

The proof for the FW2 method goes along the same lines. In this case one obtains

$$\frac{L_{t f}(t)}{t} \leq \frac{L_{t \hat{f}}(t)}{t} + \frac{\zeta_{i+1}^2}{\delta - t} \leq f(t),$$

or

$$\frac{L_{t f}(t)}{t} \leq \bar{r} + \frac{\bar{s}}{t} + \frac{\zeta_{i+1}^2}{\delta - t} \leq f(t).$$

Convergence from any point in  $(0, \delta)$  and the quadratic order of convergence follow as before. Note that the iterates now approach the root from the right-hand side. The convergence proof for hybrid methods is analogous, and for the interpolant mentioned as an example in the definition of the method, e.g., one obtains

$$\frac{L_{\Delta f}(t)}{\delta - t} \geq \frac{L_{\Delta \tilde{f}}(t)}{\delta - t} - \frac{\zeta_i^2}{t} \geq \frac{L_{\Delta h}(t)}{\delta - t} + \frac{\zeta_{i-1}^2}{\delta_{i-1} - t} - \frac{\zeta_i^2}{t} \geq f(t),$$

where

$$h(t) = f(t) - \frac{\zeta_{i-1}^2}{\delta_{i-1} - t} + \frac{\zeta_i^2}{t},$$

or

$$\frac{L_{\Delta f}(t)}{\delta - t} \geq r + \frac{s}{\delta - t} - \frac{\zeta_i^2}{t} \geq \hat{r} + \frac{\hat{s}}{\delta - t} + \frac{\zeta_{i-1}^2}{\delta_{i-1} - t} - \frac{\zeta_i^2}{t} \geq f(t).$$

This shows that this particular hybrid method is an improvement over the FW1 method, which concludes the proof.  $\square$

The following theorem shows that using a rational approximation as in Lemma 2.1 improves the fixed weight methods, in spite of the fact that in [21] such an approximation was considered to be inadequate. Analogously, a similar improvement can be obtained for the hybrid methods.

**Theorem 3.3.** *Interpolating  $\tilde{f}$  ( $\hat{f}$ ) as in the BNS1 (BNS2) method improves the FW1 (FW2) method.*

*Proof.* We prove the theorem for the FW1 method and, as it is analogous, only outline the proof for the FW2 method. We have from (8)

$$(9) \quad \frac{L_{\Delta f}(t)}{\delta - t} \geq \frac{L_{\Delta \tilde{f}}(t)}{\delta - t} - \frac{\zeta_i^2}{t} .$$

Now, because  $\tilde{\psi}$  has similar properties as  $\psi$ , inequality (6) still holds with  $\Delta f$  replaced by  $\Delta \tilde{f}$ ,  $\psi$  replaced by  $\tilde{\psi}$  and  $f$  replaced by  $f + \zeta_i^2/t$ . This gives

$$(10) \quad \frac{L_{\Delta \tilde{f}}(t)}{\delta - t} \geq 1 + \frac{1}{L_{1/\tilde{\psi}}(t)} + \frac{L_{\Delta \phi}(t)}{\delta - t} \geq f(t) + \frac{\zeta_i^2}{t} .$$

Inequalities (9) and (10) therefore yield

$$\frac{L_{\Delta f}(t)}{\delta - t} \geq \frac{L_{\Delta \tilde{f}}(t)}{\delta - t} - \frac{\zeta_i^2}{t} \geq 1 + \frac{1}{L_{1/\tilde{\psi}}(t)} + \frac{L_{\Delta \phi}(t)}{\delta - t} - \frac{\zeta_i^2}{t} \geq f(t) .$$

This completes the proof for the FW1 method. For the FW2 method, one obtains

$$\frac{L_{t f}(t)}{t} \leq \frac{L_{t \hat{f}}(t)}{t} + \frac{\zeta_{i+1}^2}{\delta - t} \leq 1 + \frac{L_{t \psi}(t)}{t} + \frac{1}{L_{1/\hat{\phi}}(t)} + \frac{\zeta_{i+1}^2}{\delta - t} \leq f(t) .$$

This concludes the proof. □

An iterative method, based on the improvement in Theorem 3.3, requires the solution of a cubic equation. This can be accomplished with, e.g., Newton’s method. This means that in the hybrid method, where an iterative method has to be used in any case to find the root of the interpolant, the BNS interpolants should be favored over the ones in the fixed weight methods.

**3.3. Transformation methods.** In [22], a transformation of variables of the form  $t = 1/w(\gamma)$  is used to obtain a convex function for which Newton’s method converges from any point in a given interval. Such a transformation transforms  $f(t)$  into  $F(\gamma) = f(1/w(\gamma))$ . Here we consider a particular transformation,  $w(\gamma) = \gamma$ . For this transformation one obtains after some algebra

$$F(\gamma) = 1 + \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\zeta_j^2}{\delta_j} - \zeta_i^2 \gamma + \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\left(\frac{\zeta_j}{\delta_j}\right)^2}{\gamma - \frac{1}{\delta_j}} .$$

The approximation to  $F$  at a point  $\gamma = \bar{\gamma}$ , suggested in [22], is obtained by retaining the term having  $1/\delta_{i+1} \equiv 1/\delta$  as a pole and by interpolating the remaining terms at  $\gamma = \bar{\gamma}$  with a linear function (the linear Taylor approximation). It takes the form

$$(11) \quad a + b\gamma + \frac{\left(\frac{\zeta_{i+1}}{\delta_{i+1}}\right)^2}{\gamma - \frac{1}{\delta}} .$$

The monotonic convergence and quadratic order of convergence on  $(1/\delta, +\infty)$  of this method were proved in [22]. The following theorem shows that this method is equivalent to the FW2 method. Thus, [22] provides yet another convergence proof.

**Theorem 3.4.** *The transformation method in [22] with  $w(\gamma) = \gamma$  is equivalent to the FW2 method.*

*Proof.* All interpolations are implicitly understood to be with respect to the same point  $\gamma = \bar{\gamma}$ . The function in (11) can be rewritten as

$$\left(a - \frac{\zeta_{i+1}^2}{\delta}\right) + b\gamma + \frac{\zeta_{i+1}^2}{\delta - \frac{1}{\gamma}}.$$

With  $r = a - \zeta_{i+1}^2/\delta$  and  $s = b$ , the interpolant takes the form

$$(12) \quad r + s\gamma + \frac{\zeta_{i+1}^2}{\delta - \frac{1}{\gamma}}.$$

Since  $r$  and  $s$  are such that  $r + s\gamma$  interpolates the function

$$-\zeta_i^2\gamma + \sum_{\substack{j=1 \\ j \neq i, i+1}}^n \frac{\zeta_j^2}{\delta_j - \frac{1}{\gamma}}$$

up to first order,  $r + st^{-1}$  must also interpolate up to first order the function

$$-\frac{\zeta_i^2}{t} + \sum_{\substack{j=1 \\ j \neq i, i+1}}^n \frac{\zeta_j^2}{\delta_j - t}.$$

Substituting this back into (12) and setting  $t = 1/\gamma$  in the last term yields exactly the interpolant used in the FW2 method.  $\square$

The idea of the hybrid method is also suggested in [22]. However, here the form of the approximation to the transformed function facilitates the computation of its root if more poles are included, as it is always convex and decreasing, so that Newton's method exhibits guaranteed convergence from any feasible point to the left of the (transformed) root. We stress that this is not true for the hybrid methods in [21].

The FW1 method is similarly equivalent to interpolating  $F(\gamma)$  with  $\bar{r} + \bar{s}(\gamma - 1/\delta)^{-1} - \zeta_i^2\gamma$  and the convergence of this method can be proved analogously.

#### 4. SECULAR EQUATION 2

We now briefly consider the following secular equation :

$$(13) \quad \sum_{j=1}^n \frac{\zeta_j^2}{(t - d_j)^2} - s^2 = 0,$$

with  $s$  a given constant and  $d_1 < d_2 < \dots < d_n$ . This is equation (2) from the introduction, where, for consistency, we have used  $t$  instead of  $\lambda$  to denote the variable. This equation needs to be solved for the smallest root, which lies to the left of  $d_1$ . Let us define

$$h(t) = \sum_{j=1}^n \frac{\zeta_j^2}{(t - d_j)^2}.$$

The method proposed for this problem in [12, 23, 25] interpolates  $h(t)$  up to first order with the rational function  $a(t - b)^{-2}$ . The convergence properties of the iterative method based on this interpolant, are given in the following theorem.

**Theorem 4.1.** *The iterative method for equation (13), based on the successive roots of the first-order interpolant  $a(t - b)^{-2} - s^2$ , converges to  $t^* < d_1$  from any initial point in  $[t^*, d_1)$ .*

*Proof.* From Lemma 2.1, with  $\rho = -2$ , the interpolant  $a(t-b)^{-2}$  can be written as  $(L_{h^{-1/2}}(t))^{-2}$ . Since this function interpolates  $h(t)$  up to first order,  $L_{h^{-1/2}}(t)$  interpolates  $h^{-1/2}(t)$  up to first order also. Now, since  $(L_{h^{-1/2}}(t))^{-2} - s^2 = 0$  is equivalent to  $L_{h^{-1/2}}(t) - |s|^{-1} = 0$ , we obtain the exact same iterates by applying Newton's method to the equivalent problem  $h^{-1/2}(t) - |s|^{-1} = 0$ . Since  $(h^{-1/2})'' = -(1/2)h^{-5/2}(-(3/2)h'^2 + hh'')$ , we have from Lemma 2.3 with  $\rho = -2$  that  $(h^{-1/2})'' < 0$  and therefore that  $h^{-1/2}$  is a concave function. It is also easily verified that  $h^{-1/2}(t)$  is decreasing for  $t \in [t^*, \delta_1)$ . Therefore, Newton's method converges for  $h^{-1/2}(t) - |s|^{-1} = 0$  from any point in  $[t^*, \delta_1)$  (see, e.g., [20]), and therefore our method converges likewise on the same interval with a quadratic order of convergence.  $\square$

A similar proof appears in [24] and [25], but we have included this proof nevertheless as a further illustration of our techniques.

## REFERENCES

1. Arbenz, P., Golub, G.H. (1988): On the spectral decomposition of Hermitian matrices modified by low rank perturbations with applications. *SIAM J. Matrix Anal. Appl.* **9**, pp. 40–58. MR **89c**:15028
2. Borges, C.F., Gragg, W.B. (1993): A parallel divide and conquer algorithm for the generalized real symmetric definite tridiagonal eigenproblem. In *Numerical Linear Algebra and Scientific Computing*, L. Reichel, A. Ruttan and R.S. Varga, eds., pp. 10–28. de Gruyter, Berlin. MR **94k**:65051
3. Bunch, J.R., Nielsen, C.P., Sorensen, D.C. (1978): Rank-one modification of the symmetric eigenproblem. *Numer. Math.* **31**, pp. 31–48. MR **80g**:65038
4. Bunch, J.R., Nielsen, C.P. (1978): Updating the singular value decomposition. *Numer. Math.* **31**, pp. 111–129. MR **80m**:65025
5. Chan, T.F., Olkin, J.A., Cooley, D.W. (1992): Solving quadratically constrained least squares using black box solvers. *BIT* **32**, pp. 481–495. MR **93e**:65091
6. Cuppen, J.J.M. (1981): A divide and conquer method for the symmetric tridiagonal eigenvalue problem. *Numer. Math.* **36**, pp. 177–195. MR **82d**:65038
7. Dongarra, J.J., Sorensen, D.C. (1987): A fully parallel algorithm for the symmetric eigenvalue problem. *SIAM J. Sci. Stat. Comput.* **8**, pp. s139–s154. MR **88f**:65054
8. Faddeeva, V.N. (1959): *Computational Methods of Linear Algebra*. Dover Publications, New York. MR **20**:6777
9. Forsythe, G.E., Golub, G.H. (1965): On the stationary values of a second-degree polynomial on the unit sphere, *J. Soc. Indust. Appl. Math.* **13**, pp. 1050–1068. MR **33**:3453
10. Fuhrmann, D.R. (1988): An algorithm for subspace computation with applications in signal processing. *SIAM J. Matrix Anal. Appl.* **9**, pp. 213–220. MR **89f**:65040
11. Gander, W. (1981): Least squares with a quadratic constraint. *Numer. Math.* **36**, pp. 291–307. MR **82c**:65026
12. Gander, W., Golub, G.H., von Matt, U. (1989): A constrained eigenvalue problem. *Linear Algebra Appl.* **114–115**, pp. 815–839. MR **90e**:15008
13. Gill, D., Tadmor, E. (1990): An  $O(N^2)$  method for computing the eigensystem of  $N \times N$  symmetric tridiagonal matrices by the divide and conquer approach. *SIAM J. Sci. Stat. Comput.* **11**, pp. 161–173. MR **91d**:65058
14. Golub, G.H. (1973): Some modified matrix eigenvalue problems. *SIAM Rev.* **15**, pp. 318–334. MR **48**:7569
15. Golub, G.H., von Matt, U. (1991): Quadratically constrained least squares and quadratic problems. *Numer. Math.* **59**, pp. 561–580. MR **92f**:65049
16. Gragg, W.B., Reichel, L. (1990): A divide and conquer method for unitary and orthogonal eigenproblems. *Numer. Math.* **57**, pp. 695–718. MR **91h**:65052
17. Gu M., Eisenstat S.C. (1994): A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. *SIAM J. Matrix Anal. Appl.* **15**, pp. 1266–1276. MR **96c**:65057

18. Handy, S.L., Barlow, J.L. (1994): Numerical solution of the eigenproblem for banded, symmetric Toeplitz matrices. *SIAM J. Matrix Anal. Appl.* **15**, pp. 205-214. MR **94j**:65053
19. Hanson, R., Phillips, J. (1975): An adaptive numerical method for solving linear Fredholm integral equations of the first kind. *Numer. Math.* **24**, pp. 291-307. MR **53**:7081
20. Henrici, P. (1964): *Elements of numerical analysis*. John Wiley & Sons, Inc., New York. MR **29**:4173
21. Li, R.C. (1994): Solving secular equations stably and efficiently. Technical Report UCB//CSD-94-851, Computer Science Division, Department of EECS, University of California at Berkeley. (LAPACK working note No. 93).
22. Melman, A. (1995): Numerical solution of a secular equation. *Numer. Math.* **69**, pp. 483-493. MR **95j**:65050
23. Reinsch, C.H. (1967): Smoothing by spline functions. *Numer. Math.* **10**, pp. 177-183. MR **45**:4598
24. Reinsch, C.H. (1971): Smoothing by spline functions II. *Numer. Math.* **16**, pp. 451-454. MR **45**:4598
25. von Matt, U. (1993): *Large constrained quadratic problems*. Verlag der Fachvereine, Zürich.

DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT, BEN-GURION UNIVERSITY,  
BEER-SHEVA 84105, ISRAEL

*E-mail address:* `melman@bgumail.bgu.ac.il`