Efficient Algorithms for Computing the Nearest Polynomial with a Real Root and Related Problems*

Markus A. Hitz
Department of Mathematics
and Computer Science
North Georgia College & State University
Dahlonega, GA 30597
E-mail: mahitz@ngcsu.edu

Erich Kaltofen
Mathematics Department
North Carolina State University
Raleigh, NC 27695-8205
E-mail: kaltofen@eos.ncsu.edu
URL: www.math.ncsu.edu/~kaltofen

Lakshman Y. N.
Computing Sciences Research
Bell Labs
Murray Hill, NJ 07974
E-mail: ynl@research.bell-labs.com

1 Introduction

We present three new algorithms in the general area of input-sensitivity analysis: a problem formulation, possibly with floating point coefficients, lacks an expected property because the inputs are slightly perturbed. A task is to efficiently compute the nearest problem that has the desired property. Nearness to the desired property can lead to problems for numerical algorithms: for example, an almost singular linear system cannot be solved by classical numerical techniques. In such case one can approach the problem of locating the nearest problem with the desired property by symbolic computation techniques, for instance, by exact arithmetic.

Our three properties are:

1. A univariate polynomial with real coefficients has a real root
2. A square matrix with real entries has a real eigenvalue
3. A bivariate polynomial with complex coefficients has a linear complex factor

Obviously, if we take input data with any of the above properties and change coefficients/entries in the slightest, as would be the case, for example, when the input data is the result of the numerical computation or a physical measurement, the perturbed problem can lose the property. At task is to recover a nearby input data with the wanted property in an efficient manner. Although that data may still only be an approximation of the actual problem, it now can be processed under the assumption of the known property. For example, a multiple real root can be removed.

Nearest polynomial with a real root The problem of finding the nearest polynomial with a given root has been studied before (see [1, 15, 12, 13, 7]). The problem in this paper is very special: nearness is measured coefficient-wise, i.e., in infinity norm. And the root locus is parametric, namely, the real axis. All previous solutions seem to have required that for parametric root locations the distance expression is at least differentiable: they and we have proven results using the Euclidean distance. Infinity norm leads us [7] to linear programming problems, whose parametric versions we do not know how to solve efficiently. We can solve our specific problem efficiently, that is in polynomial-time in the degree and input length, because an explicit expression for the distance to the nearest polynomial with a real root can be derived from a result in [17], or alternatively by eigenvalue analysis of companion matrices in [15, Section 4.2]. The expression involves absolute values, but by a stroke of luck can be minimized over the entire real axis.

Nearest matrix with a real eigenvalue The coefficient space of polynomials is not restricted to coefficients before powers of the variable. Often a polynomial is represented as the characteristic polynomial of a matrix. One must perturb the entries of the matrix to achieve a property for the characteristic polynomial. The measure of distance perhaps must be chosen differently, as entry-wise minimization can lead to NP-hard problems [16]. Stability problems, i.e., where the locus of the eigenvalues is to remain in an open subset of complex plane, are discussed for complex matrices and the induced 2-norm in [19]. The computation of the nearest matrix with a multiple eigenvalue is discussed in [14]. We give a polynomial-time solution to the problem of finding, given a real matrix, the nearest real matrix with a real eigenvalue. We measure nearness in terms of the matrix norms induced by the infinity- and one-norms on the n-dimensional vectors. Our methods are similar to the previous problem, but

*This material is based on work supported in part by the National Science Foundation under Grant No. CCR-9712267 (Erich Kaltofen).

in this case the explicit formulas for the distances are all well-known.

Nearest multivariate complex polynomial with a linear factor The problem of computing in polynomial time the nearest multivariate polynomial that factors over the complex numbers is posed in [11]. It remains unsolved. Here we take a partial step, in that we compute the nearest polynomial that has a linear factor. Our methods generalize to any factor of fixed degree. Distance is measured in the Euclidean norm of the coefficient vector, and the technique is taken from [12, 7]. Having a linear factor can be formulated as a least squares problem. Because of explicit determined solution, the coefficients of the linear factor can be taken as parameters, and the least square solution can be optimized for those parameters.

The algorithms to each of the three problems are given in a separate section below.

2 The nearest polynomial with a real root

2.1 Preliminaries

Tchebycheff gives a variant of least squares optimization, which finds the nearest constant vector in terms of entry-wise (infinity) norm that makes a linear system consistent:

$$\min_{\hat{x}} \Big(\max_{1\le i\le m} \big| b_i - \sum_{j=1}^{n} a_{i,j}\hat{x}_j \big| \Big)$$

By introducing a new variable δ we can derive the minimum by solving the linear program due to Tchebycheff.

$$\begin{array}{ll} \text{minimize: } \delta & \\ \text{linear constraints:} & \delta \ge \quad b_i - \sum_{j=1}^{n} a_{i,j}\hat{x}_j \quad (1 \le i \le m) \\ & \delta \ge -b_i + \sum_{j=1}^{n} a_{i,j}\hat{x}_j \quad (1 \le i \le m) \end{array}$$

Stiefel gives algorithms for solving the Tchebycheff problem based on work by Vallée-Poussin [17] and on the simplex method [18]. We will use the explicit formula for the optimum solution, which Stiefel [17] has found under special circumstances.

Theorem 1 *Let* $\mathbf{A}$ *be a matrix*

$$\mathbf{A} = \begin{bmatrix} a_{0,0} & \cdots & a_{0,n-1} \\ \vdots & & \vdots \\ a_{n,0} & \cdots & a_{n,n-1} \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}$$

of rank n *such that no row of* $\mathbf{A}$ *is the zero vector, and let* $\mathbf{b} = [b_0, \ldots, b_n] \in \mathbb{R}^{n+1}$ *such that* $\mathbf{Ax} \neq \mathbf{b}$ *for all* $\mathbf{x} = [x_0, \ldots, x_{n-1}] \in \mathbb{R}^n$. *Then*

$$\delta = \min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_\infty = \left| \frac{\sum_{i=0}^{n} \lambda_i b_i}{\sum_{i=0}^{n} |\lambda_i|} \right| ,$$

where $\Lambda = [\lambda_0, \ldots, \lambda_n]^{tr} \neq 0$ *is a linear dependency among the rows of* $\mathbf{A}$, *i.e.,* $\Lambda^{tr}\mathbf{A} = 0$.

Proof: Let $\mathbf{A}_i = [a_{i,0}, \ldots, a_{i,n-1}]$ denote the i^{th} row vector of $\mathbf{A}$. Each equation $\mathbf{A}_i\mathbf{x} - b_i = 0$ defines a hyperplane in $\mathbb{R}^n$, while $\mathbf{A}_i$ is a normal vector for that plane. Because of the rank condition, any n out of $n+1$ hyperplanes intersect in a single point, whereas all $n+1$ hyperplanes form a *simplex* in $\mathbb{R}^n$. This is a convex polytop with $n+1$ vertices (a triangle for $n = 2$, a tetrahedron for $n = 3$, and so on). We want to characterize the inner points of this simplex. Because the rows $\mathbf{A}_i$ are linearly dependent, there must exist a vector $\Lambda = [\lambda_0, \ldots, \lambda_n]^{tr}$ such that $\Lambda^{tr}A = 0$, and $\lambda_i \neq 0$ for $0 \le i \le n$. If we remove the k^{th} equation $\mathbf{A}_k\mathbf{x} - b_k = 0$ then the remaining n equations have a unique solution $\mathbf{x}^{(k)} \in \mathbb{R}^n$. The point $\mathbf{x}^{(k)}$ is one of the vertices of the simplex. Every inner point has to be on the same side of the k^{th} hyperplane as $\mathbf{x}^{(k)}$. The sign of the function $d_k(\mathbf{x}) = \mathbf{A}_k\mathbf{x} - b_k$ indicates whether the point $\mathbf{x} \in \mathbb{R}^n$ lies to the "left" or "right" of the hyperplane with respect to the direction of the normal vector $\mathbf{A}_k$. For the extreme point $\mathbf{x}^{(k)}$ we have a residue $r_k = d_k(\mathbf{x}^{(k)}) \neq 0$. Since $\mathbf{A}_l\mathbf{x}^{(k)} - b_l = 0$ for $l \neq k$ we get

$$\lambda_k r_k = \Lambda^{tr}(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}) = (\Lambda^{tr}\mathbf{A})\mathbf{x}^{(k)} - \Lambda^{tr}\mathbf{b} = - \Lambda^{tr}\mathbf{b} . \quad (1)$$

The product $-\Lambda^{tr}\mathbf{b}$ does not depend on k, therefore the sign of $d_k(\mathbf{x})$ of an inner point $\mathbf{x}$ has to satisfy:

$$\begin{array}{l} \text{sgn } d_k(\mathbf{x}) = \text{ sgn } \lambda_k \\ \qquad\qquad \text{or} \qquad\qquad \text{for } \textbf{every } k,\ 0 \le k \le n . \\ \text{sgn } d_k(\mathbf{x}) = -\text{sgn } \lambda_k \end{array} \quad (2)$$

Note that the "or" is exclusive, and—depending on the sign of $-\Lambda^{tr}\mathbf{b}$—either one of the two cases applies to all $d_k(\mathbf{x})$.

Evidently, the minimum $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_\infty$ is attained where one of the vertices of a hypercube centered at $\mathbf{b}$ touches the range of $\mathbf{A}$ (a hyperplane in $\mathbb{R}^{n+1}$), i.e., when all residues $d_i(\mathbf{x})$ have the same absolute value δ. Figure 1 illustrates the situation for the two-dimensional case: any deviation from the optimal point increases the absolute value of some $d_i(\mathbf{x})$, and in turn $\max_{0\le i\le n} |d_i(\mathbf{x})|$.

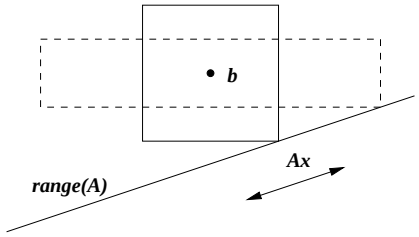

Figure 1: Projecting $\mathbf{b}$ onto the range of $\mathbf{A}$ in the infinity norm.

Therefore, we can express the residues at the optimal point $\mathbf{x}$, which by geometric reasons must be an inner point [17, Satz 3], as $d_i(\mathbf{x}) = d \cdot \text{sgn } \lambda_i$, where $d = \pm\delta$. Finally, from (1) we get:

$$\sum_{i=0}^{n} \lambda_i (d \cdot \text{sgn } \lambda_i) = \sum_{i=0}^{n} \lambda_i d_i(\mathbf{x}) = \Lambda^{tr}(\mathbf{Ax} - \mathbf{b}) = -\Lambda^{tr}\mathbf{b} = -\sum_{i=0}^{n} \lambda_i b_i ,$$

hence

$$d\sum_{i=0}^{n}\lambda_i \text{sgn } \lambda_i = -\sum_{i=0}^{n}\lambda_i b_i\,,$$

$$\text{thus} \quad \delta = \left|\sum_{i=0}^{n}\lambda_i b_i \Big/ \sum_{i=0}^{n}|\lambda_i|\right|. \quad (3)$$

Once we know the sign of every λ_i we can compute δ as well as the minimizing vector $\mathbf{x}$ by solving the linear system $\tilde{\mathbf{A}}\tilde{\mathbf{x}} - \mathbf{b} = 0$, where

$$\tilde{\mathbf{A}} = \begin{bmatrix} a_{0,0} & \cdots & a_{0,n-1} & \sigma_0 \\ \vdots & & \vdots & \vdots \\ a_{n,0} & \cdots & a_{n,n-1} & \sigma_n \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}\,,$$

$$\Sigma = [\sigma_0,\ldots,\sigma_n]^{tr} = [\text{sgn } \lambda_0,\ldots,\text{sgn } \lambda_n]^{tr}\,, \quad \text{and}$$

$$\tilde{\mathbf{x}} = [x_0,\ldots,x_{n-1},d]^{tr}\,.$$

Here, $d = \pm\delta$ plays the role of a slack variable, and it will eventually contain the norm expression (3) up to its sign. Geometrically, the "x-part" $\mathbf{x} = [x_0,\ldots,x_{n-1}]$ of the solution $\tilde{\mathbf{x}}$ defines the intersection point $\mathbf{Ax}$ of the range of $\mathbf{A}$ with that diagonal of the hypercube around $\mathbf{b}$ which has Σ as directional vector in $\mathbb{R}^{n+1}$. Due to the rank conditions, we have an immediate proof for the existence and uniqueness of $\mathbf{x}$ and δ. □

If we would replace the hypercube around the point $\mathbf{b}$ by a "diamond" in $\mathbb{R}^{n+1}$, we would get the minimum in the 1-norm in lieu of the infinity norm. From the proof given above, one can easily derive the following corollary, that we include without proof here:

Corollary 1 *Let* $\mathbf{A},\mathbf{x}$*, and* $\mathbf{b}$ *be defined as in theorem 1. Then* $\min\|\mathbf{Ax}-\mathbf{b}\|_1$ *is attained at one of the vertices* $\mathbf{x}^{(k)}$ *of the simplex mentioned in the proof of theorem 1:*

$$\min_{\mathbf{x}\in\mathbb{R}^n}\|\mathbf{Ax}-\mathbf{b}\|_1 = \min_{0\le k\le n}|\mathbf{A}_k\mathbf{x}^{(k)} - b_k|\,.$$

In general, exactly one vertex of the hyper-diamond will touch the range of $\mathbf{A}$. Therefore at the 1-norm minimum, we will have a residue r_k in one particular coordinate x_k only. If an edge or a facet of the diamond or the hypercube, in the case of the 1-norm or infinity norm respectively, is in any way colinear to the range of $\mathbf{A}$, there will exist infinitely many solutions to the minimization problem.

The sign rules (2) will prove to be essential for deriving our algorithm in the next section. They also suggest that an extension of theorem 1 to the case of complex numbers seems difficult to be accomplished along this line of reasoning.

2.2 The general case

In this section, we assume that

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

is a polynomial with real coefficients that does not have any real roots. Evidently the degree n of f has to be even, i.e., $n = 2m$ for some $m \in \mathbb{Z}$. Furthermore, the roots of f are m pairs of complex conjugates. We want to perturb the coefficients of f minimally in the infinity norm, such that the perturbed polynomial $\tilde{f}$ (of equal or lesser degree than f) has at least one real root. As shown in [7] and [6], we can state the problem as a linear minimization problem:

$$\delta = \min_{\substack{\tilde{f}\in\mathbb{R}[x]\\ \deg\tilde{f}\le\deg f}}\|\tilde{f}-f\|_\infty = \min_{\mathbf{u}\in\mathbb{R}^n}\|\mathbf{Pu}-\mathbf{b}\|_\infty\,, \quad (4)$$

where $\|\tilde{f}-f\|$ is the norm of the coefficient vector, and

$$\mathbf{b} = [a_0,\ldots,a_n]^{tr}\,,$$

$$\mathbf{u} = [c_0,\ldots,c_{n-1}]^{tr}\,,$$

$$\mathbf{P} = \begin{bmatrix} -\alpha & & & \\ 1 & -\alpha & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & -\alpha \\ & & & 1 \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}\,.$$

The vector $\mathbf{u}$ contains the coefficients of the co-factor of the linear factor $x-\alpha$, while the matrix $\mathbf{P}$ represents multiplication by $x-\alpha$. The indeterminate $\alpha \in \mathbb{R}$ is the real root of $\tilde{f}$. Its actual value will be determined by *parametric minimization* in a second step.

In the following we want to use the result of theorem 1 to express the symbolic minimum as a function in α. In order to satisfy the prerequisites for the theorem, we have to exclude $\alpha = 0$. However, we will see after the fact that the result also covers this special case. Now, it can easily be verified that the vector $\Lambda = [1,\alpha,\alpha^2,\ldots,\alpha^n]$ represents a linear dependency among the rows of $\mathbf{P}$. Therefore, we can express the symbolic minimum as the function (see also [7], [6], and [15]):

$$\delta(\alpha) = \left|\frac{\sum_{i=0}^{n}\lambda_i a_i}{\sum_{i=0}^{n}|\lambda_i|}\right| = \left|\frac{f(\alpha)}{\sum_{i=0}^{n}|\alpha^i|}\right|.$$

For $\alpha = 0$, the expression becomes: $\delta(0) = |f(0)/1| = |a_0|$, i.e., the minimum perturbation is obtained by dropping the constant coefficient a_0, while $a_1,\ldots,a_n$ each may be perturbed arbitrarily, as long as the absolute value of each perturbation does not exceed $|a_0|$. Therefore, we have infinitely many solutions to the minimization problem for $\alpha = 0$, although the value of δ is consistent with theorem 1.

We still have to find the critical values of $\delta(\alpha)$ in order to determine the overall minimum, and finally to compute α. In general, we would have exponentially many possible values for the sign vector Σ, representing all possible 2^{n+1} vertices of the hypercube mentioned in the proof of theorem 1. However for our problem, we only have to check *two* cases, namely $\alpha < 0$ and $\alpha > 0$, because the λ_i only depend on the single parameter α. We will denote the two norm functions by $d^-(\alpha)$ and $d^+(\alpha)$ for $\alpha < 0$ and $\alpha > 0$, respectively. In order to determine the root α that yields the global minimum, we have to compute the derivative of d^- and d^+ *symbolically*. Any real zero of one of the two derivatives is a candidate for the optimal choice of α. However, we only have to run the root-finder on the appropriate domain, i.e., $\alpha < 0$ or $\alpha > 0$. Finally, we have to evaluate our norm expressions at those values, and select the one(s) that minimize $\delta(\alpha)$. Once we know the minimizing α, we can compute the actual perturbations by solving the linear system mentioned above.

The following example illustrates the individual steps of our method:

Example 1 *$f(x) = x^2 + 1$ has the complex roots $+\mathbf{i}$ and $-\mathbf{i}$. We look at the norm expressions:*

$$d^+(\alpha) = \frac{\alpha^2+1}{\alpha^2+\alpha+1}, \quad \text{and} \quad d^-(\alpha) = \frac{\alpha^2+1}{\alpha^2-\alpha+1}.$$

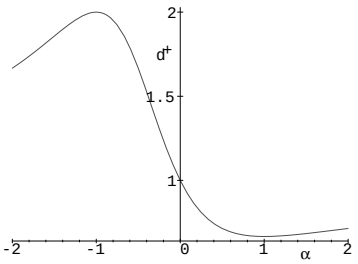

Figure 2: $d^+(\alpha)$ for $f(x) = x^2 + 1$

For the given polynomial, the graphs for $d^+(\alpha)$ and $d^-(\alpha)$ are symmetric about the y-axis. The plot for $d^+(\alpha)$ in figure 2 shows a minimum at $\alpha = 1$, therefore, we compute the perturbations by solving the linear system

$$\begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ d \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = 0$$

for a_0, a_1, and d, and finally determine $\tilde{f}(x) = \frac{1}{3}(x^2 - 2x + 1)$. The perturbations have absolute value $\delta = 2/3$, and $\tilde{f}$ has a double root at $x = 1$. We could derive another solution for $\tilde{f}$, yielding the same value for δ, by using $\alpha = -1$.

For even n, the denominator $\sum_{i=0}^{n} |\alpha|^i$ of $\delta(\alpha)$ does not have any real roots, because the roots of $\alpha^{n+1} - 1 = (\alpha-1)\sum_{i=0}^{n} \alpha^i$, as well as of $\alpha^{n+1}+1 = (\alpha-1)\sum_{i=0}^{n}(-1)^i\alpha^i$, are located on the unit circle in the complex plane. Therefore, we do not have to take poles of either $d^+(\alpha)$ or $d^-(\alpha)$ into account. We only have to compute the zeros of the numerator of their derivatives to find the critical points. Furthermore, we can use $f(\alpha)$ in place of $|f(\alpha)|$ within our norm expressions, as $f(\alpha)$ is either positive or negative for all $\alpha \in \mathbb{R}$.

When perturbing the coefficients of a generic polynomial to move one or more roots to a given locus, we may encounter the following special cases (see [6]):

1. The minimal perturbation is attained for a polynomial of lesser degree than the degree of the given polynomial.
2. The resulting root can become zero. This special case leads to infinitely many solutions if we use the infinity norm or the 1-norm.
3. The derivative of the norm expression vanishes for the entire domain under consideration.

We will prove that none of these can happen in our context of real polynomials of even degree $n > 1$. The following theorem is instrumental; it shows that setting the leading coefficient of the polynomial to zero is always a sub-optimal perturbation with respect to infinity norm.

Theorem 2 *Let $f(x) = a_n x^n +$ lower order terms $\in \mathbb{R}[x]$ be a polynomial of degree n with no real root. Then there exists an $\epsilon > 0$ and a polynomial $\tilde{f} \in \mathbb{R}[x]$ of degree n with a real root such that*

$$\|f - \tilde{f}\|_\infty \leq |a_n| - \epsilon.$$

PROOF: Since $f(x)$ has no real root, n is even. We set $f(x) = a_n x^n + g(x)$ where $g(x)$ is the reductum polynomial of degree less than n. First suppose that $g(x)$ has odd degree. Then $g(x)$ has a real root of odd multiplicity m that is isolated in a disc of radius δ around this root in the complex plane. We prove that there is an $\epsilon > 0$ such that $\tilde{f}_\epsilon(x) = \epsilon x^n + g(x)$ only has m roots in this disc. First, the boundary of the disc is a circle C of radius δ which since it isolates the root cannot contain a root of $g(x)$. We will choose ϵ such that $|g(z)| > \epsilon|z^n|$ for all $z \in C$. Then by Rouché's theorem of complex function theory, $g(z)$ and $g(z) + \epsilon z^n$ have the same number of roots inside C. Therefore, $\tilde{f}_\epsilon$ has m complex roots within C, one of which, since m is odd and $\tilde{f}_\epsilon$ has real coefficients, must be real.

We finally extend the proof when $g(x)$ has even degree, which must then be less than $n-1$. One simply uses $\tilde{g}(x) = a_n/2\, x^{n-1} + g(x)$ in place of $g(x)$ and enforces $\epsilon < a_n/2$ in addition to the above restrictions. □

Theorem 2 immediately excludes the first special case. It can also be applied to the second case:

Corollary 2 *Given $f \in \mathbb{R}[x]$ as in theorem 2, there exists an $\epsilon > 0$ and a polynomial $\tilde{f} \in \mathbb{R}[x]$ with a real root such that*

$$\|f - \tilde{f}\|_\infty \leq |a_0| - \epsilon.$$

PROOF: Either apply theorem 2 to the *reverse* polynomial $f_r(x) := a_0 x^n + \cdots + a_{n-1}x + a_n$, or conduct the proof of theorem 2 for $\tilde{f}_\epsilon(x) = xg(x) + \epsilon$. □

Finally, the derivative of $d^+(\alpha)$ vanishes if $f(x) = c\sum_{k=0}^{n} x^k$, where $c \in \mathbb{R}$, i.e., if $f(x)$ is a multiple of the denominator of $d^+(x)$. Likewise, the derivative of $d^-(\alpha)$ vanishes if $f(x)$ is a multiple of the denominator of $d^-(x)$. Both $d^+(\alpha)$ and $d^-(\alpha)$ can vanish at the same time for polynomials of degree zero only. The following is an example of a quadratic polynomial with $d^-(\alpha)$ vanishing:

Example 2 *$f(x) = 2x^2 - 2x + 2$ has the complex roots $\frac{1}{2}(1 \pm \mathbf{i}\sqrt{3})$.*

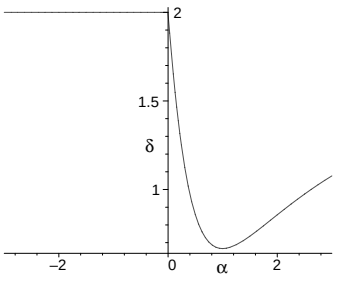

Figure 3: $\delta(\alpha)$ for $f(x) = 2x^2 - 2x + 2$

The plot for

$$\delta(\alpha) = \frac{2\alpha^2 - 2\alpha + 2}{|\alpha|^2 + |\alpha| + 1}$$

in figure 3 shows that the minimum at $\alpha = 1$ gives us a unique solution for the perturbed polynomial $\tilde{f}$. Although the derivative of $d^-(\alpha)$ vanishes, this cannot simultaneously happen to $d^+(\alpha)$. The minimal perturbation is $\delta(1) = \frac{2}{3}$, and $\tilde{f}(x) = \frac{4}{3}(x^2 - 2x + 1)$.

We summarize the discussion of this section in the description of the algorithm:

Algorithm U:

Input: $f \in \mathbb{R}[x]$, $\deg f = n > 1$ even, and $f(x) \neq 0$ for all $x \in \mathbb{R}$.

Output: $\tilde{f} \in \mathbb{R}[x]$, and $\alpha \in \mathbb{R}$, such that $\tilde{f}(\alpha) = 0$ and $\delta = \min_{\tilde{f} \in \mathbb{R}[x]} \|\tilde{f} - f\|_\infty$.

U1: Let $d^+(\alpha) := f(\alpha)/(\sum_{i=0}^{n} \alpha^i)$ and $d^-(\alpha) := f(\alpha)/(\sum_{i=0}^{n} (-1)^i \alpha^i)$.

U1.1: Determine the derivative of $d^+(\alpha)$ and $d^-(\alpha)$ symbolically.

U1.2: Determine all *real* roots $\alpha_k^+ > 0$ of the numerator of the derivative of $d^+(\alpha)$, and evaluate $d_k^+ := d^+(\alpha_k^+)$.

U1.3: Determine all *real* roots $\alpha_k^- < 0$ of the numerator of the derivative of $d^-(\alpha)$, and evaluate $d_k^- := d^-(\alpha_k^-)$.

U2: From the values d_k^+ and d_k^- computed in step U1.2 and U1.3, select $\delta = \min_{p,m}\{|d_p^+|, |d_m^-|\}$ and set $\alpha = \alpha_p^+$ or $\alpha = \alpha_m^-$ accordingly.

U3: Solve the linear system shown at the end of the proof of theorem 1; return $\tilde{f}$ and α.

2.3 Preserving monicity

In [7], we showed how a given monic polynomial f can be minimally perturbed such that $\tilde{f}$ is also monic. The norm expression derived from theorem 1 becomes

$$\delta(\alpha) = \left| \frac{f(\alpha)}{\sum_{i=0}^{n-1} |\alpha^i|} \right| .$$

Because n is even, $\sum_{i=0}^{n-1} |\alpha^i|$ contains the factor $|\alpha| + 1$, and $d^+(\alpha)$ has a single pole at $\alpha = -1$, while $d^-(\alpha)$ has a pole at $\alpha = 1$. Their derivatives also have a singularity at either $\alpha = -1$ or $\alpha = 1$. The algorithm has to account for these special cases.

2.4 Other "perfidious" polynomials

One of the applications of our method is sensitivity analysis of root locations of polynomials subject to coefficient perturbations. It was shown in [8] and [2] that root finding becomes ill-conditioned whenever the root is "close" to a multiple root. In our case, we convert one or more pairs of complex conjugates to double real roots. Therefore, we are able to compute the distance to the "set of ill-conditioned" problems in the infinity norm.

As an example, we look at $w(x) = \prod_{k=1}^{10}(x - k - \mathbf{i})(x - k + \mathbf{i})$, a polynomial of degree $n = 20$. It is a sibling of the Wilkinson-polynomial $f(x) = \prod_{k=1}^{20}(x - k)$ [20] which exhibits extreme sensitivity to small perturbations of its coefficients, despite of its simple structure.

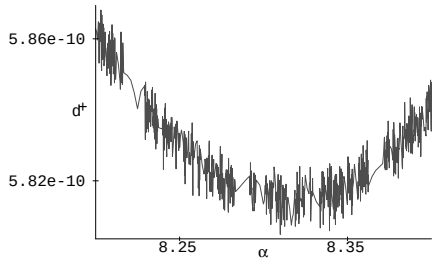

Figure 4: $d^+(\alpha)$ for $w(x)$

Figure 4 is a plot of $d^+(\alpha) = w(\alpha)/\sum_{k=0}^{20} \alpha^k$ near its local minimum between $\alpha = 8.25$ and $\alpha = 8.35$. It was produced using the default setting for the numerical precision in Maple. The oscillations are an indicator for the increased sensitivity in this area. The plot shows that the minimal infinity norm perturbation that forces (at least) one pair of complex roots to the real axis is very small: less than $5.82 \cdot 10^{-10}$.

One could argue that the picture would look differently if we would use the factored form of $w(\alpha)$, or that increasing the numerical precision in Maple would "smooth" out the plot. However, if the coefficients of the polynomial are derived from experimental data then we are always given the un-factored form. In fact, one of the common tasks will be to factor the polynomial. On the other hand, if we apply numerical methods to root finding we usually cannot adjust the precision, we are stuck with standard floating point arithmetic (e.g. IEEE 754).

3 Nearest matrix with a real eigenvalue

Van Dooren [19] made us aware of the work on matrix perturbations for purpose of moving an eigenvalue onto a given curve, for instance, the unit circle (Schur stability) or the imaginary axis (Hurwitz stability). Van Dooren is mostly interested in matrix-2-norm. The methods of section 2 apply to this problem when distance between matrices is measured in matrix-∞-norm and matrix-1-norm. For a matrix $\mathbf{B}$ we have

$$\|\mathbf{B}\|_{\infty,\infty} = \max_i \sum_j |b_{i,j}|, \quad \|\mathbf{B}\|_{1,1} = \max_j \sum_i |b_{i,j}|, \quad (5)$$

where in general

$$\|\mathbf{B}\|_{p,q} = \max_{\mathbf{x} \neq 0} \|\mathbf{B}\mathbf{x}\|_p / \|\mathbf{x}\|_q$$

is the induced matrix norm with $\|\cdot\|_l$ denoting the vector-l-norm. For the entry-wise norm, which was the subject of section 2, no such results are to be expected, as the problem of finding the nearest singular matrix, i.e., one with eigenvalue 0, is NP-hard [16].

The distance to the nearest matrix with a given eigenvalue is a direct consequence of a theorem on the distance

to the nearest singular matrix attributed to Gastinel in [9, p. 775].*

Theorem 3 *Let* $\mathbf{A}$ *be a complex matrix and* μ *be a complex number.*

$$\delta_{\mathbf{A}}(\mu) = \min_{\tilde{\mathbf{A}}:\ \mu \text{ is an eigenvalue of } \tilde{\mathbf{A}}} \|\mathbf{A} - \tilde{\mathbf{A}}\| = \frac{1}{\|(\mu\mathbf{I} - \mathbf{A})^{-1}\|}$$

where $\|\cdot\| = \|\cdot\|_{p,p}$ *is an induced matrix norm.*

In addition, one can efficiently compute the actual perturbation matrix [9].

Suppose now that we wish to compute, given a real $n \times n$ matrix $\mathbf{A}$ with no real eigenvalue, a matrix $\tilde{\mathbf{A}}$ with a real eigenvalue such that $\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\infty,\infty}$ is minimized. By Theorem 3 we first compute, for the variable parameter μ, the matrix with polynomial entries

$$\mathbf{B}(\mu) = [b_{i,j}(\mu)]_{1 \le i,j \le n} = \text{adjoint}(\mu\mathbf{I} - \mathbf{A})$$

and the characteristic polynomial

$$f(\mu) = \det(\mu\mathbf{I} - \mathbf{A}).$$

From definition (5) and Theorem 3 we conclude that one must minimize

$$\frac{1}{\|(\mu\mathbf{I} - \mathbf{A})^{-1}\|} = \frac{|f(\mu)|}{\max_i \sum_j |b_{i,j}(\mu)|}$$

over the real parameter μ. We proceed by computing $n^{O(1)}$ many real intervals such that in each interval the sign of the polynomial $b_{i,j}(\mu)$ is fixed. In each interval, we intersect for all i the polynomials $\sum_j |b_{i,j}(\mu)|$ to determine subintervals with the following property: for each subinterval there is an i_0 with $\sum_j |b_{i_0,j}(\mu)| \ge \sum_j |b_{i,j}(\mu)|$ for all i and for all μ within the subinterval. Finally, for each subinterval we minimize $|f(\mu)|/\sum_j |b_{i_0,j}(\mu)|$, this by standard techniques from calculus. Note that $f(\mu)$ has no real root, hence has one and the same sign on the subinterval. The minimum distance to the nearest matrix with a real eigenvalue is the smallest of all minima of the subintervals. Our method is polynomial-time in n and the size of the entries in $\mathbf{A}$; the method is a special case of quantifier elimination over the reals with a fixed number of variables. The algorithm for matrix-1-norm is the same.

4 Approximate factorization of bivariate polynomials

In this section, we formulate the problems of computing approximate factorization of bivariate polynomials with real or complex coefficients as parametric minimization problems. The formulations generalize to the case of multivariate polynomials easily.

*In [3] a corresponding theorem is established for Euclidean norm viewing the matrices as vectors of dimension n^2. A referee points to the alternative view for matrix-2-norm of $\delta_{\mathbf{A}}(\mu)$ being the smallest singular value of $\mu\mathbf{I} - \mathbf{A}$, which is similar to the approach taken in [3].

We state the problem for the case of polynomials with real coefficients, with the reminder that the complex coefficients case is analogous (see remark below). Let $f \in \mathbb{R}[x, y]$, of total degree m. Canonically, we write it as

$$f = \sum_{\substack{i,j=0 \\ i+j \le n}} f_{i,j} x^i y^j.$$

For normalization purposes, we regard it as monic in x, i.e., $f_{n,0} = 1$.

Approximate Factorization: Given f as above, we wish to compute a polynomial

$$\tilde{f} = \sum_{\substack{i,j=0 \\ i+j \le n}} \tilde{f}_{i,j} x^i y^j$$

$\mathbb{R}[x, y]$ of total degree n and monic in x such that $\tilde{f}$ factors over $\mathbb{R}$ in a "pre-determined" manner minimizing

$$\|f - \tilde{f}\|_2 = \sum_{\substack{i,j=0 \\ i+j \le n}} (f_{i,j} - \tilde{f}_{i,j})^2.$$

By "pre-determined", we mean that $\tilde{f}$ has two factors g, h of degrees $k, n-k$ (and monic in x) respectively for a fixed $k > 0$. More generally, one can look for factors with specific sparsity patterns and such. However, to keep the exposition simple, we formulate the minimization problem for the above version, for $k = 1$, indicating possible variations along the way. Let

$$\begin{aligned} g &= x + g_{0,1} y + g_{0,0}, \\ h &= \sum_{i,j=0; i+j<n} h_{i,j} x^i y^j, \end{aligned}$$

with $h_{n-1,0} = 1$. As we want

$$\tilde{f} = gh,$$

we should have

$$\tilde{f}_{i,j} = h_{i-1,j} + g_{0,1} h_{i,j-1} + g_{0,0} h_{i,j},$$

for $0 \le i, j; i + j \le n$. If f has a linear factor, we would have

$$\begin{aligned} f_{i,j} &= \tilde{f}_{i,j} \\ &= h_{i-1,j} + g_{0,1} h_{i,j-1} + g_{0,0} h_{i,j} \end{aligned}$$

for $0 \le i, j; i + j \le n, i \ne n$, a total of $n(n+3)/2$ linear equations in the $(n-1)(n+2)/2$ unknowns $h_{i,j}$. We re-write this system in matrix form as

$$\mathbf{Mh} = \mathbf{f}$$

where $\mathbf{h}$ is the $(n-1)(n+2)/2$-dimensional column vector

$$[h_{0,0}\ h_{1,0}\ h_{0,1} \ldots h_{n-2,1}]^{tr},$$

$\mathbf{f}$ is the $n(n+3)/2$-dimensional column vector

$$[f_{0,0}\ f_{1,0}\ f_{0,1} \ldots f_{n-1,1}]^{tr},$$

and $\mathbf{M}$ is the $n(n+3)/2 \times (n-1)(n+2)/2$ coefficient matrix corresponding to the system of linear equations above (each

row has 3 non-zero entries, $1, g_{0,1}, g_{0,0}$). When f does not have an exact linear factor, we want $\tilde{f}$ that minimizes the norm of the residual

$$\|\tilde{f} - f\|_2 = \min_{g,\,\mathbf{h}} \|\mathbf{M}\mathbf{h} - \mathbf{f}\|_2.$$

This is a least squares problem with the twist that the entries of the matrix $\mathbf{M}$ are parametric. Symbolically, $\mathbf{M}$ has full rank and we can write down the solution to the least squares problem as

$$\mathbf{h} = (\mathbf{M}^{tr}\mathbf{M})^{-1}\mathbf{M}^{tr}\mathbf{f}$$

where $\mathbf{M}^{tr}$ denotes matrix transposition. The residual is

$$\|\mathbf{f} - \mathbf{M}(\mathbf{M}^{tr}\mathbf{M})^{-1}\mathbf{M}^{tr}\mathbf{f}\|_2$$

which is a function of the parameters $g_{0,1}, g_{0,0}$. The polynomial closest (in 2-norm) to f that has a linear factor is that $\tilde{f}$ for which the above function of $g_{0,1}, g_{0,0}$ is minimized. To obtain such an $\tilde{f}$, we need to

F1: find the global minimum of

$$\|\mathbf{f} - \mathbf{M}(\mathbf{M}^{tr}\mathbf{M})^{-1}\mathbf{M}^{tr}\mathbf{f}\|_2.$$

See the remark in [13, page 658] on computing the minimum of a bivariate polynomial and the papers cited there.

F2: compute

$$\mathbf{h} = (\mathbf{M}^{tr}\mathbf{M})^{-1}\mathbf{M}^{tr}\mathbf{f}$$

at the global minimum.

Remark 1: The algorithm can be generalized to the case of f with complex coefficients and complex perturbations by separating the real and imaginary parts of each equation in the system $\mathbf{M}\mathbf{h} = \mathbf{f}$ above. One now has twice the number of equations, unknowns and parameters as before.

Remark 2: Generalization to the case of factors of degree higher than 1 and polynomials in several variables is obvious. The optimization problem gets much harder.

5 Concluding discussion

The algorithms presented in this paper are of polynomial-time complexity in the input size with the coefficients represented in any of the customary exact ways. There are several questions whose answers are not entirely explored.

1. For the problem of finding the nearest polynomial with a root on a curve, such as the real axis, how important is the choice of distance norm? In this paper we consider the theoretically more challenging infinity norm, but it is not clear when in a given situation one should switch from the Euclidean norm to infinity norm, especially since the Euclidean norm algorithms [1, 7] are much less costly. We suppose the same question arises in least squares problems, where the infinity norm approximation requires the solution of a linear programming problem.

2. The practicality of all of our algorithms is open. Implementations of our algorithms with fixed precision floating point arithmetic may be fairly unstable, as the goal polynomials have multiple roots. For the problem of approximately factoring a bi-variate polynomial over the complex numbers, numerical instability has been observed for the polynomial-time solution [10, 5, 4]. In fact, our algorithms open an exact approach to imprecisely presented inputs that would be numerically unmanageable. The price is a higher computational complexity.

3. The parametric optimization approach of [12, 13], which our algorithms have utilized, appears to reach its limit when the GCDs or factors of the perturbed inputs are to have a high degree. These problems are solvable via the quantifier elimination algorithm, and perhaps the most pressing problem is to first put them into polynomial-time.

Acknowledgements: Rob Corless directed us to the results about the nearest polynomial with a given root in [1, 15]. Gilles Villard brought [14] to our attention. We also thank the reviewers for their cogent comments.

References

Note: many of Erich Kaltofen's publications are accessible through links in the online BibTeX bibliography database at `www.math.ncsu.edu/~kaltofen/bibliography/`.

[1] Corless, R. M., Gianni, P. M., Trager, B. M., and Watt, S. M. The singular value decomposition for polynomial systems. In *Proceedings of the 1995 International Symposium on Symbolic and Algebraic Computation, ISSAC'95* (1995), pp. 195–207.

[2] Demmel, J. W. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Mathematik 51* (1987), 251–289.

[3] Eckart, C., and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika 1*, 3 (Sept. 1936), 211–218.

[4] Galligo, A., and Watt, S. A numerical absolute primality test for bivariate polynomials. In *ISSAC 97 Proc. 1997 Internat. Symp. Symbolic Algebraic Comput.* (New York, N. Y., 1997), W. Küchlin, Ed., ACM Press, pp. 217–224.

[5] Hitz, M. Porting computer algebra algorithms to numerical computing – the difficult case. *SIGSAM Bulletin 30*, 1 (Mar. 1996), 44–45. ECCAD'96 abstract.

[6] Hitz, M. A. *Efficient Algorithms for Computing the Nearest Polynomial with Constrained Roots.* PhD thesis, Rensselaer Polytechnic Institute, Troy, N.Y., 1998.

[7] Hitz, M. A., and Kaltofen, E. Efficient algorithms for computing the nearest polynomial with constrained roots. In *Proceedings of the 1998 International Symposium on Symbolic and Algebraic Computation, ISSAC'98* (1998), pp. 236–243.

[8] Hough, D. G. *Explaining and ameliorating the ill condition of zeros of polynomials.* PhD thesis, University of California, Berkeley, C.A., 1977.

[9] Kahan, W. Numerical linear algebra. *Canadian Math. Bull. 9* (1966), 757–801.

[10] Kaltofen, E. Fast parallel absolute irreducibility testing. *J. Symbolic Comput. 1*, 1 (1985), 57–67. Misprint corrections: *J. Symbolic Comput.* vol. 9, p. 320 (1989).

[11] Kaltofen, E. Polynomial factorization 1987-1991. In *Proc. LATIN '92* (Heidelberg New York, 1992), I. Simon, Ed., vol. 583 of *Lect. Notes Comput. Sci.*, Springer, pp. 294–313.

[12] Karmarkar, N., and Lakshman Y. N. Approximate polynomial greatest common divisors and nearest singular polynomials. In *Proceedings of the 1996 International Symposium on Symbolic and Algebraic Computation, ISSAC'96* (1996), pp. 35–39.

[13] Karmarkar, N. K., and Lakshman Y. N. On approximate GCDs of univariate polynomials. *J. Symbolic Comput. 26*, 6 (1998), 653–666. Special issue on Symbolic Numeric Algebra for Polynomials S. M. Watt and H. J. Stetter, editors.

[14] Lippert, R. A., and Edelman, A. The computation and sensitivity of double eigenvalues. Manuscript; see link at `http://www-math.mit.edu/~edelman/`, Jan. 1998.

[15] Manocha, D., and Demmel, J. Algorithms for intersecting parametric and algebraic curves II: Multiple intersections. *Computer Vision, Graphics and Image Processing: Graphical Models and Image Processing 57*, 2 (1995), 81–100.

[16] Poljak, S., and Rohn, J. Checking robust nonsingularity is NP-hard. *Math. Control Signals Systems 6* (1993), 1–9.

[17] Stiefel, E. Über diskrete und lineare Tschebyscheff-Approximationen. *Numerische Mathematik 1* (1959), 1–28.

[18] Stiefel, E. Note on Jordan elimination, linear programming, and Tschebyscheff approximation. *Numerische Mathematik 2* (1960), 1–17.

[19] Van Dooren, P., and Vermaut, V. On stability radii of generalized eigenvalue problems. In *Proc. European Conference on Control* (1997).

[20] Wilkinson, J. H. The perfidious polynomial. In *Studies in Numerical Analysis*, G. H. Golub, Ed., vol. 24 of *Studies in Mathematics*. M.A.A., 1984, pp. 1–28.