Backward Error Measures for Roots of Polynomials

Simon Telen[1], Sascha Timme*[2], and Marc Van Barel†[1]

[1]Department of Computer Science, KU Leuven
[2]Technische Universität Berlin, Chair of Discrete Mathematics/Geometry
simon.telen@kuleuven.be, timme@math.tu-berlin.de, marc.vanbarel@kuleuven.be

May 21, 2020

Abstract

We analyze different measures for the backward error of a set of numerical approximations for the roots of a polynomial. We focus mainly on the element-wise mixed backward error introduced by Mastronardi and Van Dooren, and the tropical backward error introduced by Tisseur and Van Barel. We show that these measures are equivalent under suitable assumptions. We also show relations between these measures and the classical element-wise and norm-wise backward error measures.

keywords: backward error; roots of polynomials; tropical backward error; element-wise mixed backward error; tropical roots.

1 Introduction

In this article we analyze the problem of measuring the backward error for a set of approximations for the roots of a polynomial with complex coefficients. For a general introduction to the notion of backward error analysis, the reader can consult for instance [Hig02, Section 1.5]. Consider a set of approximate solutions $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_d\} \subset \mathbb{C}^* = \mathbb{C} \setminus \{0\}$ of a polynomial equation with nonzero coefficients

$$f = c_0 + c_1 x + \ldots + c_{d-1}x^{d-1} + c_d x^d = c_d(x - x_1)\cdots(x - x_d) = 0, \quad f \in \mathbb{C}[x],$$

with solutions $X = \{x_1, \ldots, x_d\} \subset \mathbb{C}^*$. The *backward error* of $\hat{X}$ is a measure for the 'distance' of f to the polynomial

$$\hat{f} = \hat{c}_0 + \hat{c}_1 x + \ldots + \hat{c}_{d-1}x^{d-1} + c_d x^d = c_d(x - \hat{x}_1)\cdots(x - \hat{x}_d),$$

*This author was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) Graduiertenkolleg *Facets of Complexity* (GRK 2434).

†This author was partially supported by the Research Council KU Leuven, C1-project (Numerical Linear Algebra and Polynomial Computations), and by the Fund for Scientific Research Flanders (Belgium), G.0828.14N (Multivariate polynomial and rational interpolation and approximation), and EOS Project no 30468160.

whose roots are exactly the points in $\hat{X}$. How to measure this distance turns out to be a surprisingly subtle problem. A first and natural measure is the 2-norm distance between the coefficients of f and $\hat{f}$. The *norm-wise backward error* (NBE) of $\hat{X}$ is

$$\mathrm{NBE}(\hat{X}) = \sqrt{\frac{|c_0 - \hat{c}_0|^2 + \cdots + |c_{d-1} - \hat{c}_{d-1}|^2}{|c_0|^2 + \cdots + |c_d|^2}} = \frac{\|c - \hat{c}\|_2}{\|c\|_2},$$

where $c = (c_0, \ldots, c_d) \in \mathbb{C}^d, \hat{c} = (\hat{c}_0, \ldots, \hat{c}_{d-1}, c_d) \in \mathbb{C}^d$. In [AMVW15], an algorithm is proposed that computes a set of approximate solutions $\hat{X}$ satisfying $\mathrm{NBE}(\hat{X}) = O(u)$, where u is the unit round-off. Such an algorithm is called *norm-wise backward stable*. However, it turns out that this type of stability is too 'weak' in a sense we explain by means of an example. Consider the polynomial $f = a(x - 10^6)(x - 10^{-6})$. The set of approximate solutions $\hat{X} = \{10^6 + u10^6, 10^{-6} + u\}$ would satisfy $\mathrm{NBE}(\hat{X}) = O(u)$. Indeed,

$$\hat{c} = a(1 + (u10^6 + u + u^2 10^6), -(10^6 + 10^{-6}) - u(10^6 + 1), 1), \quad c = a(1, -(10^6 + 10^{-6}), 1)$$

and $\|c - \hat{c}\|_2 / \|c\|_2 = O(u)$. This means that we would allow a relative error of size $u10^6$ on the constant coefficient. However, computing the roots of f with $a = 0.2$ using the Julia package `PolynomialRoots` (using the command `roots`) we get

```
julia> abs.((c - chat)./c)
3-element Array{Float64,1}:
 2.7755575615628914e-16
 0.0
 0.0
```

which shows that we can obtain better element-wise accuracy on the coefficient vector. This suggests another, more 'strict', measure for the backward error. The *element-wise backward error* (EBE) of $\hat{X}$ is

$$\mathrm{EBE}(\hat{X}) = \max_{i=0,\ldots,d-1} \left| \frac{c_i - \hat{c}_i}{c_i} \right|.$$

Unfortunately, this measure turns out to be *too* strict. In [MVD15], Mastronardi and Van Dooren show that *no algorithm* for finding the roots of a general quadratic polynomial is element-wise backward stable, meaning that it computes $\hat{X}$ such that $\mathrm{EBE}(\hat{X}) = O(u)$. As an alternative measure, the authors of [MVD15] propose the following definition in the case where $d = 2$. The *element-wise mixed backward error* (EMBE) of $\hat{X}$, denoted $\mathrm{EMBE}(\hat{X})$, is the smallest number $\varepsilon \geq 0$ such that there exists some $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_d\} \subset \mathbb{C}$ and

$$\tilde{f} = \tilde{c}_0 + \tilde{c}_1 x + \cdots + \tilde{c}_{d-1} x^{d-1} + c_d x^d = c_d(x - \tilde{x}_1) \cdots (x - \tilde{x}_d)$$

such that

$$\begin{aligned} |\hat{x}_i - \tilde{x}_i| &\leq \varepsilon |\tilde{x}_i|, \quad i = 1, \ldots, d, \\ |c_i - \tilde{c}_i| &\leq \varepsilon |c_i|, \quad i = 0, \ldots, d-1. \end{aligned}$$

Note that the second set of inequalities is equivalent to $\mathrm{EBE}(\tilde{X}) \leq \varepsilon$. In the same paper, the authors also show the implication $\mathrm{EMBE}(\hat{X}) = O(u) \Rightarrow \mathrm{NBE}(\hat{X}) = O(u)$ in the case

where $d = 2$. The implication $\mathrm{EBE}(\hat{X}) = O(u) \Rightarrow \mathrm{EMBE}(\hat{X}) = O(u)$ is obvious from the definition.
The advantage of EMBE as a measure for the backward error is that it results in a notion of stability, i.e. *element-wise mixed backward stability*, which is stronger than norm-wise backward stability and can provably be obtained for quadratic polynomials [MVD15]. A drawback of this measure is that it is *hard to compute* $\mathrm{EMBE}(\hat{X})$ for a given set of approximate solutions $\hat{X}$ because of the rather abstract definition. In [TVB20], Tisseur and Van Barel define the *min-max element-wise backward error* of $\hat{X}$ as

$$\mathrm{TBE}(\hat{X}) = \max_{i=0,\dots,d-1} \left| \frac{c_i - \hat{c}_i}{r_i c_i} \right|$$

where $r_i \geq 1$ are constants that can be computed *in linear time* from the coefficients of f. The r_i depend only on the tropical roots of f, which is why we will refer to this error measure as the *tropical backward error* (TBE). We will give a definition of the numbers r_i in Section 3. The authors of [TVB20] also provide an algorithm that, under some assumptions on the numerical behavior of a modified QZ-algorithm (see [TVB20, Section 5, Assumption 1]), computes a set of approximate roots $\hat{X}$ satisfying $\mathrm{TBE}(\hat{X}) = O(u)$.

In this paper, we investigate the relations between the TBE and the EMBE. In particular, we show that these error measures are equivalent under suitable assumptions. Here's a simplified version of our first main theorem.

Theorem 1.1. *Assume that the tropical roots of f are of the same order of magnitude as the corresponding classical roots and $|\hat{x}_j|$ are of the same order of magnitude as $|x_j|$. Then we have that $\mathrm{EMBE}(\hat{X}) = O(u)$ implies $\mathrm{TBE}(\hat{X}) = O(u)$.*

The strategy for proving Theorem 1.1 will also allow us to prove that, under the assumptions of the theorem, $\mathrm{EMBE}(\hat{X}) = O(u)$ implies $\mathrm{NBE}(\hat{X}) = O(u)$. This was proved in [MVD15] for the case where $d = 2$. Under some stronger assumptions, we also prove the reverse implication.

Theorem 1.2. *Assume that the tropical roots of f are of the same order of magnitude as the corresponding classical roots and $|\hat{x}_j|$ are of the same order of magnitude as $|x_j|$. Moreover, assume that for each $x_j \in X$, there are two terms $c_{\beta'} x_j^{\beta'}$ and $c_\beta x_j^\beta$ of $f(x_j)$ such that*

$$|c_i x_j^i| \ll |c_{\beta'} x_j^{\beta'}| \quad \text{and} \quad |c_i x_j^i| \ll |c_\beta x_j^\beta|, \quad \text{for all } i \neq \beta', \beta.$$

Then we have that $\mathrm{TBE}(\hat{X}) = O(u)$ implies $\mathrm{EMBE}(\hat{X}) = O(u)$.

We will give numerical evidence that Theorem 1.2 holds without the extra assumption on the polynomial f. In summary, we have the following diagram, where the arrows are implications.

$$\begin{array}{ccc} \mathrm{EBE}(\hat{X}) = O(u) & \longrightarrow & \mathrm{TBE}(\hat{X}) = O(u) \\ \downarrow & \searrow\!\swarrow & \updownarrow \\ \mathrm{NBE}(\hat{X}) = O(u) & \longleftarrow & \mathrm{EMBE}(\hat{X}) = O(u) \end{array} \tag{1}$$

Here, the black arrows are implications which are obvious from the definitions. The blue arrows are implications that we prove under the assumptions of Theorem 1.1. The dashed arrow represents Theorem 1.2, which uses stronger assumptions.

This article is organized as follows. In the next section we prove the equivalence of the tropical and element-wise mixed backward error measures in the case where $d = 2$. This can be seen as an extension of the analysis in [MVD15], and it is instructive for the general case. The proofs for general d are given in Section 3. In Section 4 we show some computational experiments and give numerical evidence that Theorem 1.2 holds under weaker assumptions.

2 Backward Error for Quadratic Polynomials

In this section, we prove that the *element-wise mixed backward error* (EMBE) as introduced by Mastronardi and Van Dooren [MVD15] and the *tropical backward error* (TBE) from [TVB20] are equivalent backward error measures for the roots of a quadratic polynomial

$$f = ax^2 + bx + c = a(x - x_1)(x - x_2).$$

For simplicity, we assume that $a, b, c \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$. For the approximate roots $\hat{X} = \{\hat{x}_1, \hat{x}_2\}$ of f, EMBE$(\hat{X})$ is the smallest number $\varepsilon \geq 0$ such that there exists $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2\}$ with

$$\begin{aligned} \tilde{f} = a(x - \tilde{x}_1)(x - \tilde{x}_2) = ax^2 + \tilde{b}x + \tilde{c}, \\ |\hat{x}_1 - \tilde{x}_1| \leq \varepsilon|\tilde{x}_1|, \quad |\hat{x}_2 - \tilde{x}_2| \leq \varepsilon|\tilde{x}_2|, \\ |b - \tilde{b}| \leq \varepsilon|b|, \quad |c - \tilde{c}| \leq \varepsilon|c|. \end{aligned}$$

In analogy with [TVB20], we define the TBE$(\hat{X})$ to be the smallest number $\varepsilon \geq 0$ such that

$$\begin{aligned} \hat{f} = a(x - \hat{x}_1)(x - \hat{x}_2) = ax^2 + \hat{b}x + \hat{c} \\ |b - \hat{b}| \leq r_b\varepsilon|b|, \quad |c - \hat{c}| \leq \varepsilon|c|, \end{aligned}$$

where $r_b = \max(1, \sqrt{|ac|}/|b|)$. The definition of r_b will be clarified in Section 3. For the approximate roots $\hat{x}_j$ and $\tilde{x}_j$ in these definitions, we will assume that the order of magnitude of $|\tilde{x}_j|$ and $|\hat{x}_j|$ is the same as the order of magnitude of $|x_j|$ (that is, we allow relative errors of size at most 1). Note that by the definition of EMBE, it is sufficient that this is satisfied for $|\hat{x}_j|$.

We will now relate these two error measures. Let $\varepsilon = \text{EMBE}(\hat{X})$. We observe

$$\begin{aligned} \frac{\hat{c} - c}{c} &= \frac{\hat{c} - \tilde{c}}{c} + \frac{\tilde{c} - c}{c} = \frac{\hat{x}_1\hat{x}_2 - \tilde{x}_1\tilde{x}_2}{x_1x_2} + \frac{\tilde{c} - c}{c} \\ &= \frac{(\tilde{x}_1 + (\hat{x}_1 - \tilde{x}_1))(\tilde{x}_2 + (\hat{x}_2 - \tilde{x}_2)) - \tilde{x}_1\tilde{x}_2}{x_1x_2} + \frac{\tilde{c} - c}{c}. \end{aligned}$$

Since $|c - \tilde{c}| \leq \varepsilon|c|$, we have $|1 - \frac{\tilde{x}_1\tilde{x}_2}{x_1x_2}| \leq \varepsilon$ and it follows

$$\left|\frac{\hat{c} - c}{c}\right| \leq (2\varepsilon + \varepsilon^2)\left|\frac{\tilde{x}_1\tilde{x}_2}{x_1x_2}\right| + \varepsilon \lesssim 3\varepsilon\,.$$

For the coefficient $\hat{b}$, we find in an analogous way that

$$\left|\frac{\hat{b}-b}{b}\right| \leq \varepsilon \left(1+\frac{|\tilde{x}_1|+|\tilde{x}_2|}{|x_1+x_2|}\right). \tag{2}$$

If the solutions x_1 and x_2 have different orders of magnitude, there does not occur any cancellation in the denominator of the right hand side of (2) and this implies $\left|\frac{\hat{b}-b}{b}\right| \lesssim 2\varepsilon$. However, if the order of magnitude of both solutions is the same, the factor standing with ε may be significantly larger than 1 due to cancellation. We will now make this precise and relate this to the number r_b. We define $\gamma = x_1/x_2$. By the assumption that $|\tilde{x}_i|$ is of the same order of magnitude as $|x_i|$, (2) can be written as

$$\left|\frac{\hat{b}-b}{b}\right| \leq \varepsilon \left(1+K\frac{|\gamma|+1}{|\gamma+1|}\right) \tag{3}$$

with K a small constant. We assume, without loss of generality, that $0 < |\gamma| \leq 1$. Note that we have for $0 < |\gamma| < 1$ the inequality

$$\frac{|\gamma|+1}{|\gamma+1|} \leq \frac{|\gamma|+1}{||\gamma|-1|} = \frac{|\gamma|+1}{1-|\gamma|}, \tag{4}$$

which follows from $|x-y| \geq ||x|-|y||, \forall x, y \in \mathbb{C}$ applied to $x = \gamma, y = -1$. Assume that

$$\frac{\sqrt{|ac|}}{|b|} = \frac{\sqrt{|\gamma|}}{|\gamma+1|} \leq 1.$$

In this case, we have

$$\frac{|\gamma|+1}{|\gamma+1|} \leq \frac{|\gamma|+1}{\sqrt{|\gamma|}}. \tag{5}$$

Now, assume

$$\frac{\sqrt{|ac|}}{|b|} = \frac{\sqrt{|\gamma|}}{|\gamma+1|} \geq 1.$$

In this case

$$\frac{|\gamma|+1}{|\gamma+1|}\left(\frac{\sqrt{|ac|}}{|b|}\right)^{-1} \leq \frac{|\gamma|+1}{|\gamma+1|} \leq \frac{|\gamma|+1}{1-|\gamma|}. \tag{6}$$

Also,

$$\frac{|\gamma|+1}{|\gamma+1|}\left(\frac{\sqrt{|ac|}}{|b|}\right)^{-1} = \frac{|\gamma|+1}{|\gamma+1|}\left(\frac{\sqrt{|\gamma|}}{|\gamma+1|}\right)^{-1} = \frac{|\gamma|+1}{\sqrt{|\gamma|}}. \tag{7}$$

Using (4)-(7), we find that

$$\frac{|\gamma|+1}{|\gamma+1|}r_b^{-1} \leq \min\left(\frac{|\gamma|+1}{1-|\gamma|}, \frac{|\gamma|+1}{\sqrt{|\gamma|}}\right),$$

which gives

$$\frac{|\gamma|+1}{|\gamma+1|}r_b^{-1} \leq \begin{cases} \frac{\alpha+1}{1-\alpha} = \sqrt{5} & 0 < |\gamma| \leq \alpha \\ \frac{\alpha+1}{\sqrt{\alpha}} = \sqrt{5} & \alpha \leq |\gamma| \leq 1 \end{cases} \Rightarrow \frac{|\gamma|+1}{|\gamma+1|}r_b^{-1} \leq \sqrt{5},$$

where $\alpha = \frac{3}{2} - \frac{\sqrt{5}}{2}$. This is illustrated in Figure 1. It follows immediately from this

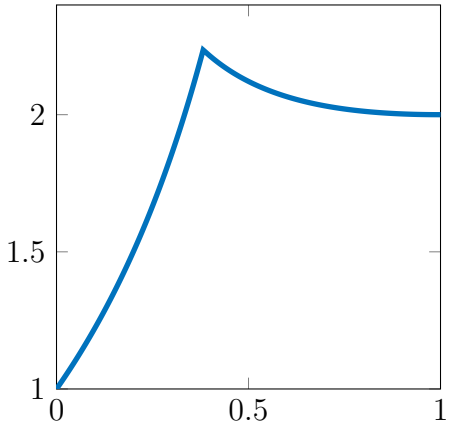

Figure 1: The value of $\min\left(\frac{|\gamma|+1}{1-|\gamma|}, \frac{|\gamma|+1}{\sqrt{|\gamma|}}\right)$ for $0 \leq |\gamma| \leq 1$.

observation and (3) that

$$\left|\frac{\hat{b}-b}{b}\right| \leq \varepsilon\left(\sqrt{5}K + r_b^{-1}\right) r_b \leq \varepsilon(\sqrt{5}K+1)r_b.$$

Figure 2 illustrates the values of r_b and $\frac{|\gamma|+1}{|\gamma+1|}r_b^{-1}$ as a function of γ in the unit disk. This

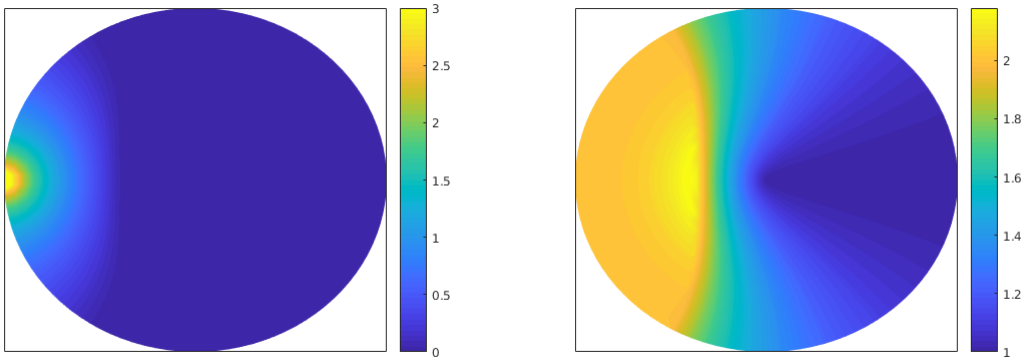

Figure 2: Left: illustration of the value of $\log(r_b)$ as a function of γ for $0 \leq |\gamma| \leq 1$. Right: illustration of the value of $\frac{|\gamma|+1}{|\gamma+1|}r_b^{-1}$ for $0 \leq |\gamma| \leq 1$.

shows that element-wise mixed backward stability implies tropical backward stability. The

converse also holds, as we will now show. Let $\mathrm{TBE}(\hat{X}) = \varepsilon$ for the computed roots $\hat{X} = \{\hat{x}_1, \hat{x}_2\}$. If $r_b = 1$ then we can take $\tilde{x}_1 = \hat{x}_1$ and $\tilde{x}_2 = \hat{x}_2$ such that $\mathrm{EMBE}(\hat{X}) \leq \mathrm{TBE}(\hat{X})$. Suppose now that $r_b = \sqrt{|ac|}/|b| > 1$ and without loss of generality assume $|\hat{x}_1| \leq |\hat{x}_2|$. $\mathrm{TBE}(\hat{X}) = \varepsilon$ implies that there exists $\hat{\delta}_b \in \mathbb{C}$ with $|\hat{\delta}_b| \leq \varepsilon$ such that

$$-\hat{b} = a(\hat{x}_1 + \hat{x}_2) = -b(1 + r_b\hat{\delta}_b)\,.$$

Let $\tilde{x}_1 = \hat{x}_1 + \frac{b}{a} r_b \hat{\delta}_b$ and $\tilde{x}_2 = \hat{x}_2$. Then we have

$$-\tilde{b} = a(\tilde{x}_1 + \tilde{x}_2) = a(\hat{x}_1 + \frac{b}{a} r_b\hat{\delta}_b + \hat{x}_2) = a(\hat{x}_1 + \hat{x}_2) + br_b\hat{\delta}_b = -b(1 + r_b\hat{\delta}_b) + br_b\hat{\delta}_b = -b\,.$$

Note that

$$|\tilde{x}_1 - \hat{x}_1| = \left|\frac{b}{a} r_b\hat{\delta}_b\right| = \sqrt{\left|\frac{c}{a}\right|}|\hat{\delta}_b| = \sqrt{|x_1x_2|}|\hat{\delta}_b| \lesssim |\tilde{x}_1|\varepsilon\,.$$

We also have

$$\tilde{c} - \hat{c} = a\tilde{x}_1\tilde{x}_2 - a\hat{x}_1\hat{x}_2 = a(\hat{x}_1 + \frac{b}{a} r_b\hat{\delta}_b)\hat{x}_2 - a\hat{x}_1\hat{x}_2 = br_b\hat{\delta}_b\hat{x}_2$$

from which we get

$$|\tilde{c} - \hat{c}| = |a|\left|\frac{b}{a} r_b\hat{\delta}_b\hat{x}_2\right| \lesssim \varepsilon|a||\tilde{x}_1\tilde{x}_2| = \varepsilon|\tilde{c}|\,.$$

We conclude that tropical backward stability implies element-wise mixed backward stability.

Remark 1. We assumed $a, b, c \in \mathbb{C}^*$ in this discussion. If $a = 0$, we are solving a linear equation and there is nothing to prove. If $c = 0$, the root $x_1 = 0$ can be deflated and we are again left with a linear equation. If $b = 0$, a similar derivation can be made. We omit the details but give a brief outline. We replace the conditions on $\tilde{b}$ and $\hat{b}$ in the definitions of $\mathrm{EMBE}(\hat{X})$ and $\mathrm{TBE}(\hat{X})$ respectively by $|\tilde{b}| \leq \varepsilon$ and $|\hat{b}| \leq r_b\varepsilon$ where $r_b = \max(1, \sqrt{|ac|}) = \max(1, |ax_1|)$ (note that in this case $|x_1| = |x_2|$). One derives bounds in a similar way for the implication $\mathrm{EMBE}(\hat{X}) = \varepsilon \Rightarrow \mathrm{TBE}(\hat{X}) = O(\varepsilon)$. For the other implication, there is again nothing to prove when $r_b = 1$. When $r_b = |ax_1|$ we observe that we can write $\hat{b} = -r_b\hat{\delta}_b$ with $|\hat{\delta}_b| \leq \varepsilon$ and we set $\tilde{x}_1 = \hat{x}_1 + a^{-1}r_b\hat{\delta}_b, \tilde{x}_2 = \hat{x}_2$.

We arrive at the following statement.

Proposition 2.1. *Let $\hat{X} = \{\hat{x}_1, \hat{x}_2\}$ be a set of approximations for the roots $X = \{x_1, x_2\}$ of a quadratic polynomial $f = ax^2 + bx + c \in \mathbb{C}[x]$, where $a, c \neq 0$. Under the assumption that $|\hat{x}_i|$ has the same order of magnitude as $|x_i|$, $i = 1, 2$, we have that*

$$\mathrm{EMBE}(\hat{X}) = O(u) \quad \text{if and only if} \quad \mathrm{TBE}(\hat{X}) = O(u).$$

3 Backward Error for Polynomials of General Degree

We now generalize the results of the previous section to polynomials of arbitrary degree d. In the following let $f = \sum_{i=0}^{d} c_i x^i \in \mathbb{C}[x]$, $c_0, c_d \neq 0$, be a polynomial of degree d with roots

$X = \{x_1, \ldots, x_d\} \subset \mathbb{C}^*$ and let $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_d\} \subset \mathbb{C}^*$ be the approximate roots. Without loss of generality we assume that the roots are labeled such that $|x_1| \leq |x_2| \leq \ldots \leq |x_d|$.

We first formally generalize the notion of the element-wise mixed backward error due to Mastronardi and Van Dooren [MVD15] from quadratic polynomials to general polynomials of degree d.

Definition 3.1 (Element-wise mixed backward error)**.** The *element-wise mixed backward error* of $\hat{X}$, denoted $\mathrm{EMBE}(\hat{X})$, is the smallest number $\varepsilon \geq 0$ such that there exist points $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_d\} \subset \mathbb{C}^*$ satisfying $|\hat{x}_j - \tilde{x}_j| \leq \varepsilon |\tilde{x}_j|$, $j = 1, \ldots, d$, and

$$\begin{cases} |c_i - \tilde{c}_i| \leq \varepsilon |c_i|, & c_i \neq 0, \\ |c_i - \tilde{c}_i| \leq \varepsilon, & c_i = 0, \end{cases}$$

for all $i = 0, \ldots, d-1$, where the $\tilde{c}_i$ are the coefficients of

$$\tilde{f} = c_d \prod_{j=1}^{d} (x - \tilde{x}_j) = c_d x^d + \sum_{i=0}^{d-1} \tilde{c}_i x^i \,.$$

Before we can state the generalization of the tropical backward error for degree d polynomials we need to introduce some definitions and concepts. We define the tropical polynomial $\mathtt{t}f(\tau)$ associated to $f(x)$ as

$$\mathtt{t}f : \mathbb{R} \cup \{-\infty\} \to \mathbb{R} \cup \{-\infty\}, \quad \tau \mapsto \max_{0 \leq i \leq d} v_i + i\tau \tag{8}$$

where $v_i = \log |c_i|$. The number v_i is the *valuation* of c_i under the valuation map $\log |\cdot|$. Any base for the logarithm can be used in theory. We want to think of the image under the valuation map as the 'order of magnitude' of the modulus of a complex number. In this paper, when we state that $\log |c| \approx 0, c \in \mathbb{C}$, we mean that $|c|$ is of *order 1*. In tropical geometry the map $\log |\cdot|$ is referred to as an Archimedean valuation. For a general introduction to tropical geometry we refer to [Sha11, MS15].

The *Newton polytope* of f is the line segment $[0, d] \subset \mathbb{R}$. The convex hull of the points $\{(i, v_i)\}_{0 \leq i \leq d} \subset \mathbb{R}^2$ is called the *lifted* Newton polytope. We will consider the *upper hull* of the lifted Newton polytope. For a specific example, this is shown as a solid black line in Figure 3. The vertices of this upper hull are the points $(\beta_\ell, v_{\beta_\ell})$, $\ell = 0, \ldots, s$, with

$$0 = \beta_0 < \beta_1 < \ldots < \beta_s = d \,.$$

We call the set $\Delta = \{(\beta_0, \beta_1), (\beta_1, \beta_2), \ldots, (\beta_{s-1}, \beta_s)\}$ the subdivision induced by the coefficients c_i, or the *induced subdivision* for short. We say that a point $\tau \in \mathbb{R} \cup \{-\infty\}$ is a *root* of $\mathtt{t}f$ if the maximum in (8) is attained at least twice. A root of $\mathtt{t}f$ is called a *tropical root* of f. The *multiplicity* of a root τ of $\mathtt{t}f$ is the number $\beta_\ell - \beta_{\ell-1}$ where β_ℓ and $\beta_{\ell-1}$ are the largest, respectively the smallest value of i for which $v_i + i\tau$ is maximal. Counted with multiplicity, $\mathtt{t}f$ has d roots $\tau_1, \ldots, \tau_d$ and we can give a closed formula for them. For $\beta_{\ell-1} < i \leq \beta_\ell$ we have

$$\tau_i = \frac{1}{m_\ell}(v_{\beta_{\ell-1}} - v_{\beta_\ell}) = \log \left| \frac{c_{\beta_{\ell-1}}}{c_{\beta_\ell}} \right|^{\frac{1}{m_\ell}}$$

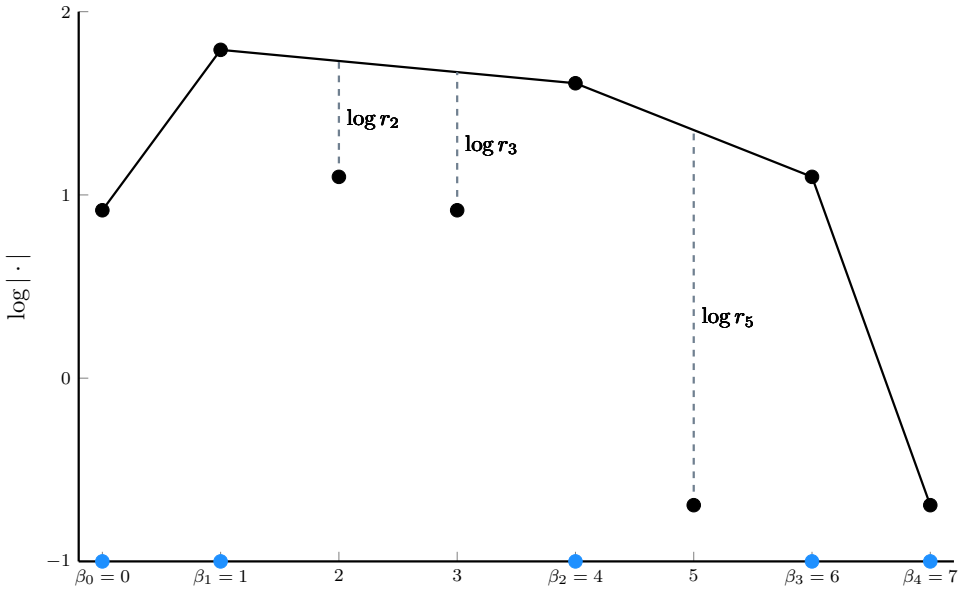

Figure 3: Consider $f(x) = \frac{1}{2}x^7 + 3x^6 + \frac{1}{2}x^5 + 5x^4 + \frac{5}{2}x^3 + 3x^2 + 6x + \frac{5}{2}$. The figure depicts the upper convex hull of the lifted Newton polytope, the geometric derivation of the r_i values and the induced subdivision $\Delta = \{(\beta_0 = 0, \beta_1 = 1), (1, \beta_2 = 4), (4, \beta_3 = 6), (6, \beta_4 = 7)\}$.

where $m_\ell = \beta_\ell - \beta_{\ell-1}$ is the multiplicity. In particular, the definition implies

$$\tau_1 = \tau_{\beta_0} = \cdots = \tau_{\beta_1} < \tau_{\beta_1+1} = \cdots = \tau_{\beta_2} < \cdots < \tau_{\beta_{s-1}+1} = \cdots = \tau_{\beta_s} = \tau_d.$$

Tropical roots of polynomials are used for scaling (matrix) polynomial eigenvalue problems, see for instance [BNS13, GS09, NST15].

Furthermore, for $i = 0, \ldots, d$ we define the constants

$$r_i = \begin{cases} 1, & i = \beta_\ell \text{ for some } \ell \\ \exp(v_{\beta_\ell} + (\beta_\ell - i)\tau_i - v_i), & \beta_{\ell-1} < i < \beta_\ell \text{ for some } \ell \text{ and } v_i \in \mathbb{R} \\ \exp(v_{\beta_\ell} + (\beta_\ell - i)\tau_i), & \beta_{\ell-1} < i < \beta_\ell \text{ for some } \ell \text{ and } v_i = -\infty \end{cases}.$$

Geometrically, if $c_i \neq 0$ then $\log r_i$ is the distance of $v_i = \log|c_i|$ to the upper convex hull of the lifted Newton polytope. Figure 3 illustrates these concepts. Note that τ_i, $\beta_{\ell-1} < i \leq \beta_\ell$, is the negative slope of the line connecting $(\beta_{\ell-1}, v_{\beta_{\ell-1}})$ and $(\beta_\ell, v_{\beta_\ell})$. We note that the complexity of computing the tropical roots of f is *linear* in the degree d, see e.g. [Sha11, Proposition 2.2.1].

Definition 3.2 (Tropical backward error)**.** The *tropical backward error* of $\hat{X}$, denoted $\mathrm{TBE}(\hat{X})$, is the smallest number $\varepsilon \geq 0$ such that for all $i = 0, \ldots, d-1$

$$\begin{cases} |c_i - \hat{c}_i| \leq r_i \varepsilon |c_i|, & c_i \neq 0, \\ |c_i - \hat{c}_i| \leq r_i \varepsilon, & c_i = 0, \end{cases}$$

where the $\hat{c}_i$ are the coefficients of

$$\hat{f} = c_d \prod_{j=1}^{d} (x - \hat{x}_j) = c_d x^d + \sum_{i=0}^{d-1} \hat{c}_i x^i \,.$$

In the following we show that an element-wise mixed backward error of order machine precision also implies a tropical backward error of order machine precision and that the converse holds, under suitable assumptions, as well. As in Section 2 for the quadratic case, we assume in the following that $|x_j|$, $|\hat{x}_j|$ and $|\tilde{x}_j|$ have the same order of magnitude for $j = 1, \ldots, d$.

3.1 Element-wise Mixed Backward Stability implies Tropical Backward Stability

We start by showing that an $\mathrm{EMBE}(\hat{X}) = \varepsilon$ implies $\mathrm{TBE}(\hat{X}) = O(\varepsilon)$. The coefficients of the polynomial f can be considered as functions of the roots $x_1, \ldots, x_d$. Let

$$\sigma_k(x_1, \ldots, x_d) = \sum_{|I|=k} \prod_{i \in I} x_i$$

be the k-th elementary symmetric polynomial. We have the identity

$$f = c_d \prod_{i=1}^{d} (x - x_i) = c_d \sum_{k=0}^{d} (-1)^{d-k} \sigma_{d-k}(x_1, \ldots, x_d) x^k.$$

Lemma 3.1. *If* $\mathrm{EMBE}(\hat{X}) = \varepsilon$*, then the coefficients of*

$$\hat{f} = c_d x^d + \sum_{i=0}^{d-1} \hat{c}_i x^i = c_d (x - \hat{x}_1) \cdots (x - \hat{x}_d)$$

satisfy

$$\left| \frac{\hat{c}_i - c_i}{c_i} \right| \leq \varepsilon \left(1 + (d-i) \frac{\sigma_{d-i}(|\tilde{x}_1|, \ldots, |\tilde{x}_d|)}{|\sigma_{d-i}(x_1, \ldots, x_d)|} + O(\varepsilon) \right), \qquad \text{when } c_i \neq 0$$

and

$$|\hat{c}_i - c_i| \leq \varepsilon (1 + c_d (d-i) \sigma_{d-i}(|\tilde{x}_1|, \ldots, |\tilde{x}_d|) + O(\varepsilon)), \qquad \text{when } c_i = 0.$$

Proof. Suppose $c_i \neq 0$. We have

$$\frac{\hat{c}_i - c_i}{c_i} = \frac{\hat{c}_i - \tilde{c}_i}{c_i} + \frac{\tilde{c}_i - c_i}{c_i}.$$

This implies

$$\begin{aligned}\left|\frac{\hat{c}_i - c_i}{c_i}\right| &\leq \left|\frac{\hat{c}_i - \tilde{c}_i}{c_i}\right| + \left|\frac{\tilde{c}_i - c_i}{c_i}\right| \\ &\leq \frac{|\sigma_{d-i}(\hat{x}_1, \ldots, \hat{x}_d) - \sigma_{d-i}(\tilde{x}_1, \ldots, \tilde{x}_d)|}{|\sigma_{d-i}(x_1, \ldots, x_d)|} + \varepsilon \\ &= \frac{|\sigma_{d-i}(\tilde{x}_1(1+\hat{\delta}_1), \ldots, \tilde{x}_d(1+\hat{\delta}_d)) - \sigma_{d-i}(\tilde{x}_1, \ldots, \tilde{x}_d)|}{|\sigma_{d-i}(x_1, \ldots, x_d)|} + \varepsilon\end{aligned}$$

where $|\hat{\delta}_i| \leq \varepsilon, i = 1, \ldots, d$. Note that the second inequality and the equality both use $\mathrm{EMBE}(\hat{X}) = \varepsilon$. We now observe

$$\sigma_{d-i}(\tilde{x}_1(1+\hat{\delta}_1), \ldots, \tilde{x}_d(1+\hat{\delta}_d)) = \sigma_{d-i}(\tilde{x}_1, \ldots, \tilde{x}_d) + \sum_{j=1}^{d-i} \hat{\delta}_j \sigma_{d-i}(\tilde{x}_1, \ldots, \tilde{x}_d) + \text{h.o.t.},$$

where the 'higher order terms' contain at least two of the $\hat{\delta}_i$. This, together with the triangle inequality, shows

$$\left|\frac{\hat{c}_i - c_i}{c_i}\right| \leq \varepsilon(d-i)\frac{\sigma_{d-i}(|\tilde{x}_1|, \ldots, |\tilde{x}_d|)}{|\sigma_{d-i}(x_1, \ldots, x_d)|} + O(\varepsilon^2) + \varepsilon.$$

The case $c_i = 0$ is completely analogous. □

It is well known that the values $\exp(\tau_i)$ are related to the modulus of the classical roots x_i, see for instance [Sha11]. In what follows, we will make the assumption that the order of magnitude of $\exp(\tau_i)$ is equal to that of $|x_i|$ (and, by our previous assumption, also to $|\hat{x}_i|$ and $|\tilde{x}_i|$). Under this assumption, we have for $\beta_{\ell-1} < i \leq \beta_\ell$ that

$$\sigma_{d-i}(|x_1|, \ldots, |x_d|) = \sum_{|I|=d-i} \prod_{j \in I} |x_j| = D_i \binom{m_\ell}{\beta_\ell - i} \prod_{k=0}^{d-i-1} \exp(\tau_{d-k}),$$

with D_i a not too large constant and

$$\binom{m_\ell}{\beta_\ell - i} = \frac{m_\ell!}{(\beta_\ell - i)!(i - \beta_{\ell-1})!}$$

the binomial coefficient. This can be seen as follows. The important terms in the expansion of $\sigma_{d-i}(|x_1|, \ldots, |x_d|)$ are those containing $d-i$ large roots. By the ordering of the roots by their modulus, these terms have the order of magnitude of $\prod_{k=0}^{d-i-1} \exp(\tau_{d-k})$. We can assume that each of these terms contains the largest $m_{\ell+1} + \cdots + m_s \leq d-i$ roots. For the remaining factors, we can choose $\beta_\ell - i$ roots among $\{x_{\beta_{\ell-1}+1}, \ldots, x_{\beta_\ell}\}$.

By our assumption that $|\tilde{x}_i| \approx |x_i|$, we have

$$\sigma_{d-i}(|\tilde{x}_1|, \ldots, |\tilde{x}_d|) = \sum_{|I|=d-i} \prod_{j \in I} |\tilde{x}_j| = \tilde{D}_i \binom{m_\ell}{\beta_\ell - i} \prod_{k=0}^{d-i-1} \exp(\tau_{d-k}),$$

with $\tilde{D}_i$ a not too large constant. Taking valuations on both sides of this equality, we get

$$\log|\sigma_{d-i}(|\tilde{x}_1|,\ldots,|\tilde{x}_d|)| = w_i + \sum_{k=0}^{d-i-1} \tau_{d-k} \tag{9}$$

where $w_i = \log\left|\tilde{D}_i\binom{m_i}{\beta_\ell - i}\right|$ and $\exp(w_i)$ is a not too large positive number. Equation (9) is the assumption we will use in the next theorem.

Theorem 3.1. *If* $\mathrm{EMBE}(\hat{X}) = \varepsilon$ *and the order of magnitude of* $|x_i|$ *is equal to that of* $|\hat{x}_i|$ *and* $\exp(\tau_i)$*,* $i = 1,\ldots,d$ *and* (9) *holds, then the coefficients of*

$$\hat{f} = c_d x^d + \sum_{i=0}^{d-1} \hat{c}_i x^i = c_d(x-\hat{x}_1)\cdots(x-\hat{x}_d)$$

satisfy

$$\left|\frac{\hat{c}_i - c_i}{c_i}\right| \le \varepsilon\,(1 + (d-i)\exp(w_i)r_i + O(\varepsilon)) \quad \textit{when } c_i \neq 0$$

and

$$|\hat{c}_i - c_i| \le \varepsilon\,(1 + (d-i)\exp(w_i)r_i + O(\varepsilon)) \quad \textit{when } c_i = 0.$$

In particular, $\mathrm{TBE}(\hat{X}) \le \varepsilon \max_i(1 + (d-i)\exp(w_i) + O(\varepsilon))$.

Proof. For $\beta_{\ell-1} \le i \le \beta_\ell$, assume $c_i \neq 0$. Note that

$$\begin{aligned}
\log|r_i| &= v_{\beta_\ell} + (\beta_\ell - i)\tau_{\beta_\ell} - v_i \\
&= v_{\beta_{\ell+1}} + m_{i+1}\tau_{\beta_{\ell+1}} + (\beta_\ell - i)\tau_{\beta_\ell} - v_i \\
&= v_{\beta_{\ell+2}} + m_{i+2}\tau_{\beta_{\ell+2}} + m_{i+1}\tau_{\beta_{\ell+1}} + (\beta_\ell - i)\tau_{\beta_s} - v_i \\
&= \ldots \\
&= v_{\beta_\ell} + \sum_{k=\ell+1}^{\ell} m_k\tau_{\beta_k} + (\beta_\ell - i)\tau_{\beta_\ell} - v_i \\
&= v_d + \sum_{k=0}^{d-i-1} \tau_{d-k} - v_i.
\end{aligned}$$

Now, using (9) and $v_i = v_d + \log|\sigma_{d-i}(x_1,\ldots,x_d)|$ we get

$$\log|r_i| = \log|\sigma_{d-i}(|\tilde{x}_1|,\ldots,|\tilde{x}_d|)| - \log|\sigma_{d-i}(x_1,\ldots,x_d)| - w_i.$$

Therefore

$$\exp(w_i)r_i = \exp(w_i + \log|r_i|) = \frac{\sigma_{d-i}(|\tilde{x}_1|,\ldots,|\tilde{x}_d|)}{|\sigma_{d-i}(x_1,\ldots,x_d)|}$$

and we are done by Lemma 3.1. The proof for $c_i = 0$ is analogous. □

In [MVD15], Mastronardi and Van Dooren show that, for $d = 2$, $\mathrm{EMBE}(\hat{X}) = O(u)$ implies $\mathrm{NBE}(\hat{X}) = O(u)$, where $\mathrm{NBE}(\hat{X})$ is the *norm-wise backward error* as defined in the introduction. Theorem 3.1 allows us to prove this statement for general degrees.

Proposition 3.1. *If* $\mathrm{EMBE}(\hat{X}) = \varepsilon$ *and the assumptions of Theorem 3.1 are satisfied, we have that*

$$\|(c_0, \ldots, c_{d-1}, c_d) - (\hat{c}_0, \ldots, \hat{c}_{d-1}, c_d)\|_2 = O(\varepsilon)\|(c_0, \ldots, c_{d-1}, c_d)\|_2.$$

Proof. Suppose that $c_i \neq 0$. We have

$$|c_i - \hat{c}_i| \leq |c_i - \tilde{c}_i| + |\hat{c}_i - \tilde{c}_i|.$$

Since $\mathrm{EMBE}(\hat{X}) = \varepsilon$, we have that $|(\hat{c}_i - c_i) - (\hat{c}_i - \tilde{c}_i)| \leq \varepsilon |c_i|$. Combining this with Theorem 3.1, we have that

$$|\hat{c}_i - \tilde{c}_i| = O(\varepsilon)|c_i| r_i.$$

Hence, we obtain the bound

$$|c_i - \hat{c}_i| \leq \varepsilon |c_i| + O(\varepsilon)|c_i| r_i.$$

Analogously, when $c_i = 0$ we obtain

$$|c_i - \hat{c}_i| \leq \varepsilon + O(\varepsilon) r_i.$$

It follows that

$$\begin{aligned} \|(c_0 - \hat{c}_0, \ldots, c_{d-1} - \hat{c}_{d-1}, c_d - c_d)\|_2 &\leq \varepsilon \|(c_0, \ldots, c_d)\|_2 + O(\varepsilon)\|(r_0 c_0, \ldots, r_d c_d)\|_2 \\ &\leq \varepsilon \|(c_0, \ldots, c_d)\|_2 + O(\varepsilon)\sqrt{d}\|(c_0, \ldots, c_d)\|_2, \end{aligned}$$

where the last inequality follows from

$$\begin{aligned} \max_{i=1,\ldots,d} |c_i| \leq \|(c_0, \ldots, c_d)\|_2 \leq \sqrt{d} \max_{i=1,\ldots,d} |c_i|, \\ \max_{i=1,\ldots,d} |c_i| \leq \|(r_0 c_0, \ldots, r_d c_d)\|_2 \leq \sqrt{d} \max_{i=1,\ldots,d} |c_i| \end{aligned}$$

because $r_i = 1$ for $i = \mathrm{argmax}_{\ell=1,\ldots,d}|c_\ell|$. □

We note that the proof of Proposition 3.1 can be summarized as

$$\mathrm{EMBE}(\hat{X}) = \varepsilon \overset{\text{Theorem 3.1}}{\Longrightarrow} \mathrm{TBE}(\hat{X}) = O(\varepsilon) \Longrightarrow \mathrm{NBE}(\hat{X}) = O(\varepsilon).$$

3.2 Tropical Backward Stability implies Element-wise Backward Stability?

We now show that a tropical backward error of order ε also implies a mixed element-wise backward error of the same magnitude under some assumptions. For this, consider the perturbed polynomials

$$\hat{f} = f + \hat{\Delta} f = \sum_{i=0}^{d-1} c_i (1 + r_i \delta_i) x^i + c_d x^d$$

and

$$\tilde{f} = f + \tilde{\Delta} f = \sum_{i=0}^{d-1} c_i (1 + \kappa_i \delta_i e^{\sqrt{-1}\theta_i}) x^i + c_d x^d \tag{10}$$

where $\log|\delta_i| \approx \log|\varepsilon| = v_\varepsilon$ and $\kappa_i \in \mathbb{R}, \theta_i \in [0, 2\pi)$ are parameters. We assume that κ_i is not too large, i.e. $\log|\kappa_i| \approx 0$, such that $\tilde{\Delta} f$ is a 'small' perturbation. Observe that for the roots $\hat{x}_j = x_j + \hat{\Delta} x_j$ of $\hat{f}$ we have

$$\begin{aligned} 0 &= (f + \hat{\Delta} f)(x_j + \hat{\Delta} x_j) \\ &= f(x_j + \hat{\Delta} x_j) + \hat{\Delta} f(x_j + \hat{\Delta} x_j) \\ &= f(x_j) + f'(x_j)\hat{\Delta} x_j + \tfrac{f''(x_j)}{2}\hat{\Delta} x_j^2 + \cdots + \tfrac{f^{(d)}(x_j)}{d!}\hat{\Delta} x_j^d \\ &\quad + \hat{\Delta} f(x_j) + \hat{\Delta} f'(x_j)\hat{\Delta} x_j + \tfrac{\hat{\Delta} f''(x_j)}{2}\hat{\Delta} x_j^2 + \cdots + \tfrac{\hat{\Delta} f^{(d)}(x_j)}{d!}\hat{\Delta} x_j^d. \end{aligned} \tag{11}$$

From this we conclude that $\hat{\Delta} x_j$ is a root of the polynomial

$$\hat{E}(x) = \hat{\Delta} f(x_j) + (f'(x_j) + \hat{\Delta} f'(x_j))\frac{x}{1!} + \cdots + (f^{(d)}(x_j) + \hat{\Delta} f^{(d)}(x_j))\frac{x^d}{d!}.$$

Similarly, for the roots $\tilde{x}_1 = x_1 + \tilde{\Delta} x_1, \ldots, \tilde{x}_d = x_d + \tilde{\Delta} x_d$ of $\tilde{f}$ we have that $\tilde{\Delta} x_j$ is a root of the polynomial

$$\tilde{E}(x) = \tilde{\Delta} f(x_j) + (f'(x_j) + \tilde{\Delta} f'(x_j))\frac{x}{1!} + \cdots + (f^{(d)}(x_j) + \tilde{\Delta} f^{(d)}(x_j))\frac{x^d}{d!}.$$

To show that tropical backward stability implies element-wise mixed backward stability we need three assumptions. Each tropical root τ_i should attain the maximum in (8) *exactly* twice, and the other terms should be significantly smaller. Also, the tropical root τ_i should be of the same order of magnitude as $\log|x_i|$.

Lemma 3.2. *If for the tropical root τ_j of f we have*

1. $\{\beta \in \{0, \ldots, d\} \mid v_\beta + \beta\tau_j = \max_i v_i + i\tau_j\} = \{\beta_{\ell-1}, \beta_\ell\}$ *with* $\beta_{\ell-1} < \beta_\ell$,
2. $\log|x_j| \approx \tau_j$,
3. $|c_i x_j^i| \ll |c_{\beta_\ell} x_j^{\beta_\ell}|, i \neq \beta_{\ell-1}, \beta_\ell$,

then for $k \geq 1$

$$\log|f^{(k)}(x_j)| \lesssim v_{\beta_\ell} + (\beta_\ell - k)\tau_j.$$

Here '$\lesssim$' can be replaced by '$\approx$' for $k = 1$.

Proof. We have that $x^k f^{(k)}(x) = \sum_{i=k}^d c_i \frac{i!}{(i-k)!} x^i$. We distinguish three different cases.

1. $(\beta_{\ell-1} - k \geq 0, \beta_\ell - k \geq 0)$. In this case

$$|x_j^k f^{(k)}(x_j)| = K_1 \left| c_{\beta_{\ell-1}} \frac{\beta_{\ell-1}!}{(\beta_{\ell-1} - k)!} x_j^{\beta_{\ell-1}} + c_{\beta_\ell} \frac{\beta_\ell!}{(\beta_\ell - k)!} x_j^{\beta_\ell} \right|,$$

with $\log|K_1| \approx 0$. Since $f(x_j) = 0$ and by assumption $|c_i x_j^i| \ll |c_{\beta_\ell} x_j^{\beta_\ell}|, i \neq \beta_{\ell-1}, \beta_\ell$, we have that

$$|c_{\beta_{\ell-1}} x_j^{\beta_{\ell-1}} + c_{\beta_\ell} x_j^{\beta_\ell}| = \left| \sum_{i \neq \beta_{\ell-1}, \beta_\ell} c_i x_j^i \right| \leq \sum_{i \neq \beta_{\ell-1}, \beta_\ell} |c_i x_j^i| \ll |c_{\beta_\ell} x_j^{\beta_\ell}|.$$

Then

$$\begin{aligned} |x_j^k f^{(k)}(x_j)| &= K_1 \left| \frac{\beta_{\ell-1}!}{(\beta_{\ell-1}-k)!}(c_{\beta_{\ell-1}} x^{\beta_{\ell-1}} + c_{\beta_\ell} x^{\beta_\ell}) + \left(\frac{\beta_\ell!}{(\beta_\ell - k)!} - \frac{\beta_{\ell-1}!}{(\beta_{\ell-1}-k)!} \right) c_{\beta_\ell} x_j^{\beta_\ell} \right| \\ &= K_2 |c_{\beta_\ell} x_j^{\beta_\ell}| \end{aligned}$$

with $\log|K_2| \approx 0$. The lemma now follows from taking valuations.

2. $(\beta_{\ell-1} - k < 0, \beta_\ell - k \geq 0)$. The lemma follows from the observation that in this case

$$|x_j^k f^{(k)}(x_j)| = K_1 \left| c_{\beta_\ell} \frac{\beta_\ell!}{(\beta_\ell - k)!} x_j^{\beta_\ell} \right| = K_2 |c_{\beta_\ell} x_j^{\beta_\ell}|,$$

with $\log|K_1| \approx 0, \log|K_2| \approx 0$.

3. $(\beta_{\ell-1} - k < 0, \beta_\ell - k < 0)$. In this case

$$|x_j^k f^{(k)}(x_j)| = \left| \sum_{i=k}^{d} c_i \frac{i!}{(i-k)!} x_j^i \right| \leq \sum_{i=k}^{d} \left| c_i \frac{i!}{(i-k)!} x_j^i \right| \ll |c_{\beta_\ell} x_j^{\beta_\ell}|.$$

Note that if $k = 1$, the third case is not possible because $\beta_\ell > \beta_{\ell-1} \geq 0$. □

Lemma 3.3. *Under the assumptions of Lemma 3.2, we have that*

$$\log|\hat{\Delta} f(x_j)| \lesssim v_{\beta_\ell} + v_\varepsilon + \beta_\ell \tau_j, \quad \log|\tilde{\Delta} f(x_j)| \lesssim v_{\beta_\ell} + v_\varepsilon + \beta_\ell \tau_j,$$

$$\log|\hat{\Delta} f'(x_j)| \lesssim v_{\beta_\ell} + v_\varepsilon + (\beta_\ell - 1)\tau_j, \quad \log|\tilde{\Delta} f'(x_j)| \lesssim v_{\beta_\ell} + v_\varepsilon + (\beta_\ell - 1)\tau_j,$$

$$\log|f^{(k)}(x_j) + \hat{\Delta} f^{(k)}(x_j)| \lesssim v_{\beta_\ell} + (\beta_\ell - k)\tau_j, \quad \log|f^{(k)}(x_j) + \tilde{\Delta} f^{(k)}(x_j)| \lesssim v_{\beta_\ell} + (\beta_\ell - k)\tau_j.$$

In the last line, for $k = 1$ we can replace '$\lesssim$' by '$\approx$'.

Proof. We have

$$|\hat{\Delta} f(x_j)| = \left| \sum_{i=0}^{d-1} c_i \delta_i r_i x_j^i \right| = K_1 \left| \sum_{\beta_{\ell-1} \leq i \leq \beta_\ell}^{d-1} c_i \delta_i r_i x_j^i \right| \leq K_1 (\beta_\ell - \beta_{\ell-1}) |c_{\beta_\ell} \delta_{\beta_\ell} x_j^{\beta_\ell}|,$$

with $\log|K_1(\beta_\ell - \beta_{\ell-1})| \approx 0$, which proves the first statement. The second statement is proven by a completely analogous argument. The third statement follows from

$$|x_j \hat{\Delta} f'(x_j)| = \left| \sum_{i=1}^{d-1} c_i i \delta_i r_i x_j^i \right| \leq K_1 (\beta_\ell - \beta_{\ell-1}) |c_{\beta_\ell} \delta_{\beta_\ell} x_j^{\beta_\ell}|,$$

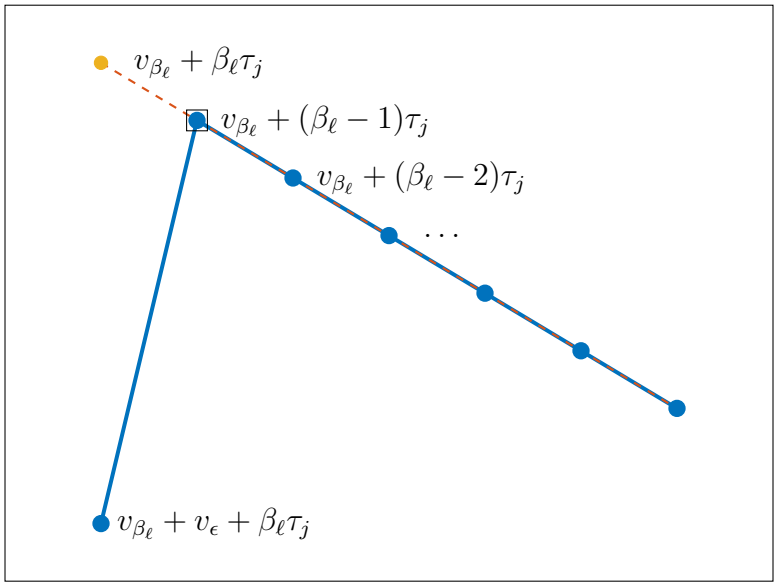

Figure 4: The blue line shows an upper bound for the lifted Newton polytopes of $\hat{E}(x)$ and $\tilde{E}(x)$. The actual lifted polytopes will meet the blue line (approximately) in the point $(1, v_{\beta_\ell} + (\beta_\ell - 1)\tau_j)$ (indicated with a small box).

with $\log|K_1(\beta_\ell - \beta_{\ell-1})| \approx 0$. The fourth statement is analogous. The fifth statement follows from

$$\log|f^{(k)}(x_j) + \hat{\Delta} f^{(k)}(x_j)| \approx \log|f^{(k)}(x_j)|$$

and Lemma 3.2. The sixth statement follows again from an analogous argument. □

It follows from Lemma 3.3 that we can bound the lifted Newton polytopes of the polynomials $\hat{E}, \tilde{E}$ from above. An example is shown in Figure 4. The expansion (11) is used to approximate $\hat{\Delta} x_j$ as

$$\hat{\Delta} x_j \approx \frac{-\hat{\Delta} f(x_j)}{f'(x_j) + \hat{\Delta} f'(x_j)}. \tag{12}$$

It is clear that this is an approximation for the smallest root of $\hat{E}(z)$, which corresponds to the smallest tropical root $\tau_{\hat{E}}$ of $\hat{E}$ which is bounded by (see Figure 4)

$$\tau_{\hat{E}} \leq (v_{\beta_\ell} + v_\varepsilon + \beta_\ell \tau_j) - (v_{\beta_\ell} + (\beta_\ell - 1)\tau_j) = \tau_j + v_\varepsilon. \tag{13}$$

Analogously, we have for the smallest tropical root $\tau_{\tilde{E}}$ of $\tilde{E}$ that $\tau_{\tilde{E}} \leq \tau_j + v_\varepsilon$. We will also make our usual assumption that the tropical roots give an indication for the order of magnitude of the classical roots, i.e.

$$\log|\hat{\Delta} x_j| \lesssim \tau_j + v_\varepsilon, \quad \log|\tilde{\Delta} x_j| \lesssim \tau_j + v_\varepsilon. \tag{14}$$

We conclude that the assumptions of Lemma 3.2 imply that $\hat{X}$ and $\tilde{X}$ have a relative *forward* error of size $O(\varepsilon)$. This implies that $\mathrm{EMBE}(\hat{X}) = O(\varepsilon)$ (take $\tilde{x}_j = x_j$), which gives the following result.

Theorem 3.2. *Under the assumptions of Lemma 3.2, if* $\mathrm{TBE}(\hat{X}) = \varepsilon$ *and the order of magnitude of* $|x_j|$ *is equal to that of* $|\hat{x}_j|$ *and* $\exp(\tau_j)$*, then* $\mathrm{EMBE}(\hat{X}) = O(\varepsilon)$.

In fact, the assumptions of Lemma 3.2 imply that the roots X of f are well-conditioned. Consider the first order approximation

$$\frac{|\Delta x_j|}{|x_j|} \approx \frac{\max_{i=0,\dots,d} |c_i x_j^i|}{|f'(x_j)||x_j|} \frac{|\Delta f(x_j)|}{\max_{i=0,\dots,d} |c_i x_j^i|},$$

for a perturbation Δf on f causing a perturbation Δx_j on x_j. Here $\max_{i=0,\dots,d} |c_i x_j^i|$ is used to measure the residual $|\Delta f(x_j)|$ with a relative criterion. The condition of a root x_j can be measured by

$$\frac{\max_{i=0,\dots,d} |c_i x_j^i|}{|f'(x_j)||x_j|} = \frac{c_{\beta_\ell} \exp(\tau_j)^{\beta_\ell}}{|f'(x_j)||x_j|}.$$

Using Lemma 3.2, we find that this number is of order 1.

In what follows, we will give a constructive proof for Theorem 3.2. That is, in the proof of Theorem 3.3 (which implies Theorem 3.2) we will give values for κ_i, θ_i in (10) which realize a small EMBE. It uses the following lemma.

Lemma 3.4. *Under the assumptions of Lemma 3.2, we have that*

$$\log \left| \hat{\Delta} x_j - \frac{-\hat{\Delta} f(x_j)}{f'(x_j)} \right| \lesssim \tau_j + 2v_\varepsilon, \quad \log \left| \tilde{\Delta} x_j - \frac{-\tilde{\Delta} f(x_j)}{f'(x_j)} \right| \lesssim \tau_j + 2v_\varepsilon.$$

Proof. It follows from $\hat{E}(\hat{\Delta} x_j) = 0$ that

$$f'(x_j)\hat{\Delta} x_j = -\left(\hat{\Delta} f(x_j) + \hat{\Delta} x_j \hat{\Delta} f'(x_j) + \sum_{k=2}^{d} \hat{\Delta} x_j^k \frac{f^{(k)}(x_j) + \hat{\Delta} f^{(k)}(x_j)}{d!} \right).$$

The valuation of the first neglected term in the approximation (12) is

$$\log \left| \frac{\hat{\Delta} x_j \hat{\Delta} f'(x_j)}{f'(x_j)} \right| \lesssim \tau_j + v_\varepsilon + v_{\beta_\ell} + v_\varepsilon + (\beta_\ell - 1)\tau_j - (v_{\beta_\ell} + (\beta_\ell - 1)\tau_j) = 2v_\varepsilon + \tau_j,$$

where we used Lemma 3.2, Lemma 3.3 and (14). For the term corresponding to $k = 2$, we have

$$\log \left| \frac{\hat{\Delta} x_j^2 f^{(2)}(x_j)}{2 f'(x_j)} \right| \lesssim 2(\tau_j + v_\varepsilon) + v_{\beta_\ell} + (\beta_\ell - 2)\tau_j - (v_{\beta_\ell} + (\beta_\ell - 1)\tau_j) = 2v_\varepsilon + \tau_j.$$

For the terms corresponding to higher values of k, we get in the same way a valuation of $\tau_j + k v_\varepsilon < \tau_j + 2v_\varepsilon$. The reasoning for $\tilde{\Delta} x_j$ is completely analogous. □

Theorem 3.3. *Under the assumptions of Lemma 3.2, there are choices of the parameters* κ_i, θ_i *with* $\log |\kappa_i| \approx 0$ *such that* $\log |\hat{x}_j - \tilde{x}_j| \approx v_\varepsilon + \tau_j$.

Proof. By Lemma 3.4 we have $\hat{x}_j = x_j - \frac{\hat{\Delta}f(x_j)}{f'(x_j)} + O(\varepsilon^2 \exp(\tau_j))$ and $\tilde{x}_j = x_j - \frac{\tilde{\Delta}f(x_j)}{f'(x_j)} + O(\varepsilon^2 \exp(\tau_j))$. Hence it suffices to show that the valuation of

$$|\hat{x}_j - \tilde{x}_j| \approx \left|\frac{1}{f'(x_j)}\right| \left|\sum_{i=0}^{d-1} c_i \delta_i (r_i - \kappa_i e^{\sqrt{-1}\theta_i}) x_j^i\right|$$

is bounded by $v_\varepsilon + \tau_j$. We have

$$\begin{aligned}\left|\frac{1}{f'(x_j)}\right| \left|\sum_{i=0}^{d-1} c_i \delta_i (r_i - \kappa_i e^{\sqrt{-1}\theta_i}) x_j^i\right| &\leq \left|\frac{1}{f'(x_j)}\right| \sum_{i=0}^{d-1} |c_i||\delta_i||r_i - \kappa_i e^{\sqrt{-1}\theta_i}||x_j^i| \\ &\leq \varepsilon \left|\frac{1}{f'(x_j)}\right| \sum_{i=0}^{d-1} |c_i||r_i - \kappa_i e^{\sqrt{-1}\theta_i}||x_j^i|.\end{aligned}$$

We now specify the parameters κ_i, θ_i. For $r_i = O(1)$, we choose κ_i, θ_i such that $r_i = \kappa_i e^{\sqrt{-1}\theta_i}$. Note that $\log|\kappa_i| \approx 0$. For the other i, we set $\kappa_i = \theta_i = 0$. We get

$$|\hat{x}_j - \tilde{x}_j| \leq \varepsilon \left|\frac{1}{f'(x_j)}\right| \sum_{r_i \gg 1} |c_i||r_i||x_j^i|.$$

Since by assumption $\log|x_j| \approx \tau_j$, the dominant terms in the sum are those with $\beta_{\ell-1} < i < \beta_\ell$, where $\tau_{\beta_{\ell-1}+1} = \cdots = \tau_{\beta_\ell} = \tau_j$. Therefore

$$|\hat{x}_j - \tilde{x}_j| \leq K\varepsilon \left|\frac{1}{f'(x_j)}\right| \sum_{\beta_{\ell-1} < i < \beta_\ell} |c_i||r_i||x_j^i|$$

with $\log|K| \approx 0$. The valuation of one of the terms in the sum is

$$\log|K| + v_\varepsilon - \log|f'(x_j)| + v_i + v_{\beta_\ell} + (\beta_\ell - i)\tau_j - v_i + i\tau_j$$

which is equal to $\log|K| - \log|f'(x_j)| + v_\varepsilon + v_{\beta_\ell} + \beta_\ell \tau_j$. Note that this is independent of i, and hence we get

$$\log|\hat{x}_j - \tilde{x}_j| \approx -\log|f'(x_j)| + v_\varepsilon + v_{\beta_\ell} + \beta_\ell \tau_j.$$

Using Lemma 3.2 we get

$$\log\left|\frac{\hat{x}_j - \tilde{x}_j}{\tilde{x}_j}\right| \approx v_\varepsilon.$$

□

4 Computational Experiments

In Subsection 3.2 we proved that a tropical backward error of order ε also implies a mixed element-wise backward error of the same magnitude under some assumptions. Unfortunately, we were not able to prove this result in general. However, based on several numerical experiments that we performed, we are convinced that a small TBE implies a small

EMBS also in general. To support this conjecture, the following numerical experiment was performed.

Numerical Experiment 1 Take 1000 polynomials of degree d with coefficients whose modulus is chosen as 10^e with e uniformly randomly chosen between $-k$ and k and whose argument is uniformly randomly chosen between 0 and 2π. These polynomials will not always satisfy the necessary assumptions for Theorem 3.2. The zeros of these polynomials are approximated by applying an eigenvalue method from the Julia package `Polynomials` resulting in the computed zeros $\hat{X}$. For these approximate roots the TBE is computed. To compute an upper bound for the EMBE, the roots $\hat{x}_j$ are separately refined using Newton's method in extended precision based on the original polynomial. The correspondence between the TBE and EMBE is shown in Figure 5 and clearly indicates that a small TBE implies a small EMBE.

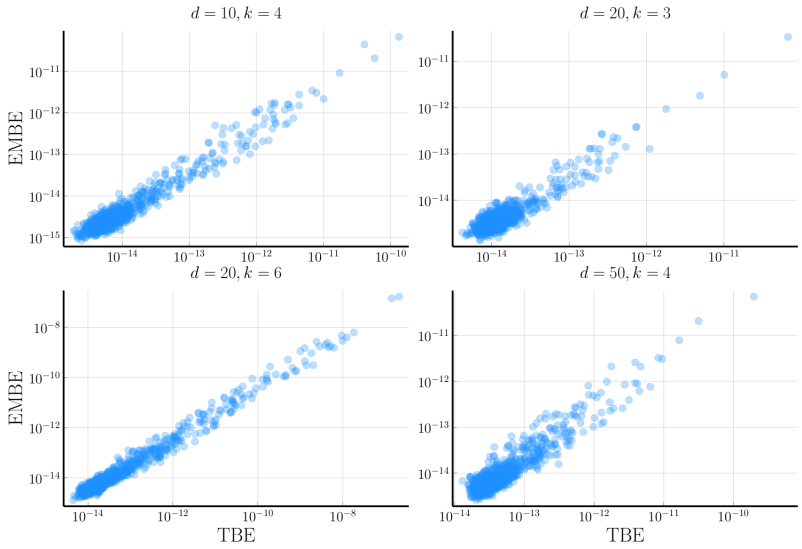

Figure 5: Results of Numerical Experiment 1.

In several of our statements, we assumed that the tropical roots of f are of the same order of magnitude as the corresponding classical roots. To check if this is a reasonable assumption in practice, we performed the following numerical experiment.

Numerical Experiment 2 Ten thousand polynomials of degree d are taken with coefficients whose modulus is chosen as 10^e with e uniformly randomly chosen between $-k$ and k and whose argument is uniformly randomly chosen between 0 and 2π. For each of these polynomials the tropical roots are compared to the roots computed in high precision. Figure 6 gives a histogram of the measured ratios $|\tau_i/x_i|$. The results show that for the vast majority of roots the magnitude differs by at most 10 percent.

5 Conclusion

We have shown the relations (1) between different measures for the backward error of an approximate set of roots $\hat{X}$ of a polynomial. Under some assumptions the tropical

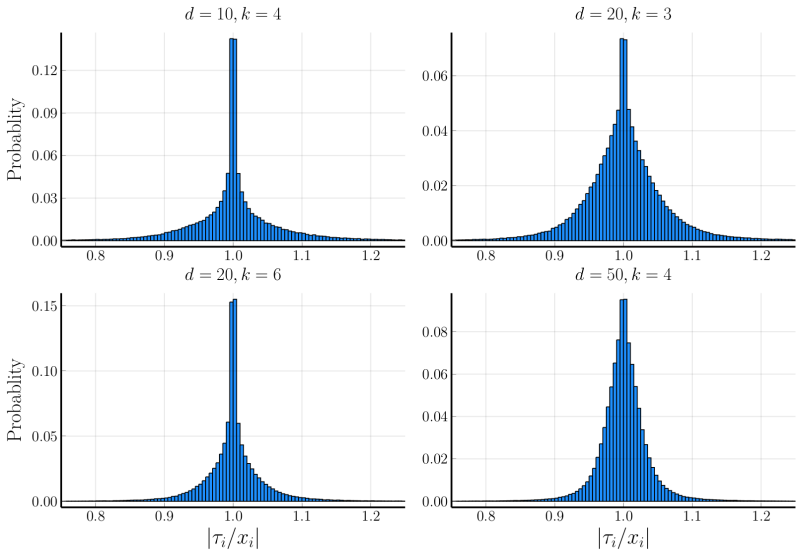

Figure 6: Results of Numerical Experiment 2.

backward error measure of [TVB20], which is easy to compute, is shown to be equivalent to the element-wise mixed backward error measure defined in [MVD15] for $d = 2$. We have given numerical evidence that the equivalence holds more generally.

References

[AMVW15] Jared L Aurentz, Thomas Mach, Raf Vandebril, and David S Watkins. Fast and backward stable computation of roots of polynomials. *SIAM Journal on Matrix Analysis and Applications*, 36(3):942–973, 2015.

[BNS13] Dario A Bini, Vanni Noferini, and Meisam Sharify. Locating the eigenvalues of matrix polynomials. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1708–1727, 2013.

[GS09] Stéphane Gaubert and Meisam Sharify. Tropical scaling of polynomial matrices. In *Positive systems*, volume 389 of *Lecture Notes in Control and Information Sciences*, pages 291–303. Springer, 2009.

[Hig02] Nicholas J Higham. *Accuracy and stability of numerical algorithms*, volume 80. Siam, 2002.

[MS15] Diane Maclagan and Bernd Sturmfels. *Introduction to tropical geometry*, volume 161. American Mathematical Soc., 2015.

[MVD15] Nicola Mastronardi and Paul Van Dooren. Revisiting the stability of computing the roots of a quadratic polynomial. *Electronic Transactions on Numerical Analysis*, 44:73–82, 2015.

[NST15] Vanni Noferini, Meisam Sharify, and Françoise Tisseur. Tropical roots as approximations to eigenvalues of matrix polynomials. *SIAM Journal on Matrix Analysis and Applications*, 36(1):138–157, 2015.

[Sha11] Meisam Sharify. *Scaling Algorithms and Tropical Methods in Numerical Matrix Analysis: Application to the Optimal Assignment Problem and to the Accurate Computation of Eigenvalues*. PhD thesis, 2011.

[TVB20] Françoise Tisseur and Marc Van Barel. Min-max elementwise backward error for roots of polynomials and a corresponding backward stable root finder. *arXiv:2001.05281*, 2020.