Penalty Function Based Critical Point Approach to Compute Real Witness Solution Points of Polynomial Systems

Wenyuan Wu[1,2], Changbo Chen[1,2(✉)], and Greg Reid[3]

[1] Chongqing Key Laboratory of Automated Reasoning and Cognition, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China
{wuwenyuan,chenchangbo}@cigit.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Applied Mathematics Department, Western University, London, Canada
reid@uwo.ca

Abstract. We present a critical point method based on a penalty function for finding certain solution (witness) points on real solutions components of general real polynomial systems. Unlike other existing numerical methods, the new method does not require the input polynomial system to have pure dimension or satisfy certain regularity conditions.

This method has two stages. In the first stage it finds approximate solution points of the input system such that there is at least one real point on each connected solution component. In the second stage it refines the points by a homotopy continuation or traditional Newton iteration. The singularities of the original system are removed by embedding it in a higher dimensional space.

In this paper we also analyze the convergence rate and give an error analysis of the method. Experimental results are also given and shown to be in close agreement with the theory.

1 Introduction

Computational real algebraic geometry is the study of the global structure of real solution sets of polynomial systems, including positive dimensional solution components (see [2] for a background text on algorithms for exact real algebraic geometry). This paper is a contribution to the development of numerical algorithms for computational real algebraic geometry directed at numerically describing such global structure. In contrast, conventional numerical methods seek local solutions which are points, and generally do not give information on positive dimensional solution components.

Numerical algebraic geometry [16,29] was pioneered by Sommese, Wampler, Verschelde and others (see the texts [9,27] for references and background). They first considered the easier characterization of complex solution components of each possible dimension, by slicing the solution set with appropriate random

V.P. Gerdt et al. (Eds.): CASC 2017, LNCS 10490, pp. 377–391, 2017.
DOI: 10.1007/978-3-319-66320-3_27

planes, that intersected the solution components in complex points called *witness points*. The complex points are computed by homotopy continuation solvers. For example, a one dimensional circle, $x^2+y^2-1=0$ in $\mathbb{C}^2$ is intersected by a random line in two such witness points, but this method obviously fails for $(x,y) \in \mathbb{R}^2$ since a real line can miss the circle.

Instead the method in [31,32] yields real witness points as critical points of the distance from a random hyperplane to the real variety. The reader can easily see this yields two real witness points for the circle example. An alternative numerical approach where the witness points are critical points of the distance from a random point to the real variety has been developed in [15]. The works [15,31,32] use Lagrange multipliers to set up the critical point problem.

A contribution of our current paper is to remove the assumptions in [15, 31,32] by developing a penalty function based critical point method where the singularities are removed by embedding systems in a higher dimensional space. The method has two stages. In the first stage it finds approximate solution points of the input system such that there is at least one real point on each connected solution component. In the second stage it refines the points by a homotopy continuation or traditional Newton iteration. We also analyze the convergence rate and give an error analysis of our method. Experimental results are given and shown to closely agree with the theory.

Critical point methods in Lagrange form appeared previously in important symbolic works [24–26]. In those works, the systems are analyzed using Gröbner Bases. Ultimately numerical methods have to be used to approximate points on components, but only after application of symbolic algorithms to the systems, instead of the fully numerical methods used here and in [15,31,32]. Also see the early related symbolic works [1,6] and the recent work [5,11].

More distantly related symbolic approaches for computational real algebraic geometry include cylindrical algebraic decomposition (CAD) introduced by Collins [13] and improved by many others. Recent improvement of CAD by using triangular decompositions are given in [12] for solving semi-algebraic systems. But the double exponential cost of the CAD algorithm [14] is the main barrier to its application.

Numerical methods based on moment matrices and semi-definite programming techniques have been developed to approximate real radical ideals of zero dimensional systems, e.g. [20,21]. For a positive dimensional system, an approach is given in [7] which combines numerical algebraic geometry and sums of squares programming to test whether the input is real radical or not. Also see [23,33], based on moment matrices, and [22].

As a development of critical point approaches [15,24–26,31], this article will propose an approximation method to compute real witness points of polynomial systems without any regularity assumption [31,32] or pure dimension assumption [15]. In Sect. 2, we will describe how the polynomial systems are embedded in a higher dimensional space. In Sect. 3, we will describe error control with a rank assumption. In Sect. 4, this rank assumption is removed and error control is

provided for general systems. In Sect. 5, our method is illustrated with examples and concluding remarks are given in Sect. 6.

2 Augmented System

In this section we introduce our augmented system, that involves adding a variable to each equation, so the original system is obtained when these slack variables are set to zero. The resulting augmented system has solution set that is a smooth real manifold. We alert the reader that this is different to the embedding systems of (complex) numerical algebraic geometry. To avoid confusion with the well known embedding systems of the complex case we have used a different name for our systems, Augmented System.

Let $x = (x_1, \ldots, x_n)$. Let $f = \{f_1, \ldots, f_k\}$ be a set of polynomials in the ring $\mathbb{R}[x]$. We construct the following augmented system g for f with slack variables $z = (z_1, \ldots, z_k)$:

$$g = \{f_1 + z_1, f_2 + z_2, \ldots, f_k + z_k\}. \tag{1}$$

Note that $g \subset \mathbb{R}[x, z]$ holds.

Lemma 1. *The Jacobian matrix of g w.r.t. the variables $(x_1, \ldots, x_n, z_1, \ldots, z_k)$ has rank k at any point of $V_{\mathbb{R}}(g)$ and $V_{\mathbb{R}}(g)$ is a smooth submanifold of $\mathbb{R}^{n+k}$ with dimension n.*

Proof. Firstly, $V_{\mathbb{R}}(g) \neq \emptyset$ since $\{x_1 = 0, \ldots, x_n = 0, z_1 = -f_1(0), \ldots, z_k = -f_k(0)\}$ is a real solution. Secondly, it is easy to see that the Jacobian matrix $\frac{\partial g}{\partial (x,z)}$ has full rank k at any solution $(x^*, z^*) \in V_{\mathbb{R}}(g)$, which implies that $V_{\mathbb{R}}(g)$ is a smooth submanifold of $\mathbb{R}^{n+k}$ with dimension n by the regular level set theorem (see pp. 113–114 of [19]). □

By Lemma 1, the augmented system g satisfies the regularity assumptions A_1 and A_2 of [31]. Moreover, any point on $V_{\mathbb{R}}(g)$ being smooth is a crucial property for numerical stability of numerical methods applied to $V_{\mathbb{R}}(g)$.

Using the critical point technique [24], we choose a random point $\mathfrak{a} = (\mathfrak{a}_1, \ldots, \mathfrak{a}_n)$, where $\mathfrak{a} \notin V_{\mathbb{R}}(f)$, in x-space and consider the minimal distance from $V_{\mathbb{R}}(f)$ to this point. As the norm of the slack variables z approaches zero, the corresponding point of $V_{\mathbb{R}}(g)$ approaches $V_{\mathbb{R}}(f)$. To force the slack variables z to be very small, we introduce a penalty function $\beta \cdot (z_1^2 + \cdots + z_k^2)/2$ with penalty factor $\beta \gg 0$ and formulate the following optimization problem

$$\begin{aligned} \min\ \mu = (\beta \cdot (z_1^2 + \cdots + z_k^2) + \sum\nolimits_{i=1}^{n} (x_i - \mathfrak{a}_i)^2)/2 \\ s.t. \qquad g = 0. \end{aligned} \tag{2}$$

To solve the optimization problem above, we can use Lagrange multiplier techniques:

$$\begin{pmatrix} x_1 - \mathfrak{a}_1 \\ \vdots \\ x_n - \mathfrak{a}_n \\ \beta\, z_1 \\ \vdots \\ \beta\, z_k \end{pmatrix} = \begin{pmatrix} \partial f_1/\partial x_1 & \cdots & \partial f_k/\partial x_1 \\ \vdots & \ddots & \vdots \\ \partial f_1/\partial x_n & \cdots & \partial f_k/\partial x_n \\ 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}_{(n+k)\times k} \cdot \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix}. \tag{3}$$

Then (3) implies that $\lambda_i = \beta z_i = -\beta f_i$. Substituting this solution back into the Eq. (3) above yields a square system with n variables

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \beta \cdot \mathcal{J}^t \cdot \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} = \begin{pmatrix} \mathfrak{a}_1 \\ \vdots \\ \mathfrak{a}_n \end{pmatrix}. \tag{4}$$

where the $n \times k$ matrix $\mathcal{J}^t$ is the transpose of the Jacobian of f.

The optimization problem (2) is equivalent to the following unconstrained optimization problem (will be used in the next two sections):

$$\min\ \mu = (\beta \cdot (f_1^2 + \cdots + f_k^2) + \textstyle\sum_{i=1}^{n}(x_i - \mathfrak{a}_i)^2)/2. \tag{5}$$

Setting the gradient of μ to be zero, we also obtain Eq. (4).

Note that the left hand side of Eq. (4) defines a smooth mapping $M : \mathbb{R}^n \to \mathbb{R}^n$.

Lemma 2. *For a random point $\mathfrak{a} = (\mathfrak{a}_1, \ldots, \mathfrak{a}_n) \notin V_{\mathbb{R}}(f)$, Problem (2) has solutions and $M^{-1}(\mathfrak{a}) \neq \emptyset$. Moreover, every point of the real variety $M^{-1}(\mathfrak{a})$ is a regular point of M with probability* 1.

Proof. Let $z_i = w_i/\sqrt{\beta}$, $i = 1, \ldots, k$, and substitute them into (2). Let $h = \{\sqrt{\beta} f_1 + w_1, \ldots, \sqrt{\beta} f_k + w_k\}$. We obtain another equivalent form of Problem (2), where the objective function is now formulated as a distance function:

$$\begin{aligned} \min\ & (w_1^2 + \cdots + w_k^2 + \sum\nolimits_{i=1}^{n}(x_i - \mathfrak{a}_i)^2)/2 \\ s.t.\quad & h = 0. \end{aligned} \tag{6}$$

By Lemma 1, $V_{\mathbb{R}}(g)$ is a smooth submanifold of $\mathbb{R}^{n+k}$ with dimension n. So is $V_{\mathbb{R}}(h)$. For any $\mathfrak{a} \notin V_{\mathbb{R}}(f)$, the point $(x = \mathfrak{a}, w = 0)$ does not belong to $V_{\mathbb{R}}(h)$. Thus, (6) always has minimum distance from $(\mathfrak{a}, 0)$ to $V_{\mathbb{R}}(h)$ by completeness of the real numbers, which implies that the minimal value of (2) can always be attained. Since the Jacobian matrix of g has full rank at any point of $V_{\mathbb{R}}(g)$, a solution $\{x^*, z^*\}$ of Problem (2) must be a solution of Eq. (3), which implies that x^* is a solution of Eq. (4).

Thus $M^{-1}(\mathfrak{a}) \neq \emptyset$. By Sard's Theorem [30], for almost all $\mathfrak{a}$, every point of the real variety $M^{-1}(\mathfrak{a})$ is regular point of M. □

Lemma 2 implies that all the solutions of Eq. (4) can be obtained by applying homotopy continuation methods.

Among these solutions, we look for solutions with small residuals i.e. $\|z\| \ll 1$. It is possible that such points do not exist, which then provides strong evidence that $V_{\mathbb{R}}(f)$ is empty. Intuitively, this is because if $V_{\mathbb{R}}(f)$ is not empty, increasing the penalty factor β will force $\|z\|$ to be close to zero.

A theoretical study on the relationship between the magnitude of the residual $\|z\|$ and the emptiness of $V_{\mathbb{R}}(f)$ is out of the scope of this paper and will be treated in a future work. **In the rest of this paper, we always assume that** $V_{\mathbb{R}}(f) \neq \emptyset$. A natural question is how to estimate the distances of the local minima of Problem (5) to $V_{\mathbb{R}}(f)$. We divide this problem into two cases w.r.t. the rank of the Jacobian and will address them in the next two sections.

Note that, throughout this paper, the norm $\|\cdot\|$ always means the 2-norm.

3 Error Control with Rank Assumption

Since Problem (5) with penalty function is different from the goal of finding real witness points of the original system $f = 0$, it is of great importance to study the difference between their solutions. In this section, we will give an error estimate of the approximate answer given by solving Problem (5) under a rank assumption. In the next section, we will remove this assumption and give an error estimate for general systems.

For a smooth point x on $V_{\mathbb{R}}(f)$, let the local dimension of $V_{\mathbb{R}}(f)$ at point x be ℓ. The Jacobian matrix at x is a $k \times n$ matrix denoted by $\mathcal{J}_x$. Suppose its rank is m, where $m \leq \min\{k, n\}$. Then we say that x satisfies the **rank condition**, if $m = n - \ell$, i.e.

$$rank\, \mathcal{J}_x = n - \dim V_{\mathbb{R}}(f)_x \tag{7}$$

If any smooth point on $V_{\mathbb{R}}(f)$ satisfies the rank condition, then we say the system f satisfies the rank condition. For example this occurs if $f = \{x - y, x^2 - y^2\}$. This means that f can be an over-determined system and even generate a non-radical ideal (e.g. consider $f = \{(x^2+1)^2(x-y)\}$). Note that $f = (x-y)^2$ does not satisfy the rank condition, although its graph is a smooth line. Such systems will be discussed in the next section.

For a random point $\mathfrak{a} \in \mathbb{R}^n$, there is at least one point on each connected component of $V_{\mathbb{R}}(f)$ with minimal distance to $\mathfrak{a}$ satisfying the following problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} (x_i - \mathfrak{a}_i)^2 \\ s.t. \quad & x \in V_{\mathbb{R}}(f). \end{aligned} \tag{8}$$

Let us consider such a point p of (8) with local minimal distance to $\mathfrak{a}$. For this point, there exists a constant c and we have $\|f_i(p + \Delta x)\| < c\|\Delta x\|$ for each polynomial f_i when Δx is sufficiently small. The value of the target function μ at p of (5) is $D^2/2$ where $D = \|p - \mathfrak{a}\|$. If we move p towards $\mathfrak{a}$ with a sufficiently small distance Δx to p' then

$$\mu(p') = \beta\|f(p')\|^2/2 + (p' - \mathfrak{a})^2/2 < \beta c^2\,\Delta x^2/2 + (D - \Delta x)^2/2 < D^2/2. \quad (9)$$

It means that p of Problem (8) is not a local minimum of Problem (5). Let p' be a local minimum of (5) for a given β. Consequently, $p' \notin V_{\mathbb{R}}(f)$. But we have the following result.

Corollary 1. *Let p be a local minimum of (8). There exists a local minimum p' of (5) for sufficiently large β, such that $\|p - p'\|$ can be arbitrarily small.*

Proof. For any small $\delta > 0$, consider the sphere S of a ball centered at p with radius δ. Let $D = \|p - \mathfrak{a}\|$, where $\mathfrak{a} \notin V_{\mathbb{R}}(f)$ is the given point for both problems (5) and (8). The sphere S can be divided into two sets: $S_1 = \{x \in S : \|x - \mathfrak{a}\| \leq D\}$ and $S_2 = \{x \in S : \|x - \mathfrak{a}\| > D\}$. Since p obtains the local minimum distance from $V_{\mathbb{R}}(f)$ to $\mathfrak{a}$, we have $S_1 \cap V_{\mathbb{R}}(f) = \emptyset$ for a small enough δ. Let $s = \min_{x \in S_1}(\sum_j f_j(x)^2)$. So $s > 0$. When $\beta s + (D - \delta)^2 > D^2$, i.e. $\beta > \frac{2\delta D - \delta^2}{s}$, we have $\mu(x) > D^2/2 = \mu(p)$ for any point x on the sphere S. Since the ball is a compact set, the local minimal value of μ must be attained at p' inside this ball. □

We now consider how to estimate the error $\|p - p'\|$ for a given β. First let us consider a simple case when a local minimum p of (8) satisfies the rank condition (7). Then, we have the following result.

Theorem 2. *Suppose p is a local minimum of (8) satisfying the rank condition (7). Then there is at least one real solution p' of Eq. (4) such that $\|p - p'\| < \frac{D}{\beta\sigma_m^2 + 1}$, where $D = \|p - \mathfrak{a}\|$ and σ_m is the smallest nonzero singular value of $\mathcal{J}_p$.*

Proof. Without loss of generality, we assume that p is the origin o. Because of the rank condition, the local dimension at p is equal to $n - rank\mathcal{J}_p = n - m$. Moreover, the null-space of $\mathcal{J}_p$ is the tangent space T at p. Let N be the orthogonal complement of T in $\mathbb{R}^n$. Since $p \in T$ has the minimum distance to $\mathfrak{a}$, the vector $\mathfrak{a}$ belongs to N.

Let $U^T\mathcal{J}_pV = \Sigma_{k\times n} = diag(\sigma_1, \ldots, \sigma_m, 0, \ldots, 0)$ be the singular value decomposition of the Jacobian matrix at p, where $U = ([u_1|\cdots|u_k]) \in \mathbb{R}^{k\times k}$ and $V = ([v_1|\cdots|v_n]) \in \mathbb{R}^{n\times n}$. Then, the space N is spanned by $\{v_1, \ldots, v_m\}$.

By Corollary 1 for sufficiently large β, there exists a local minimum p' of (5) such that $\|p' - p\| = \delta \ll 1$ and $f(p') = f(p) + \mathcal{J}_p \cdot p' + O(\delta^2)$ since p is the origin.

Let $p' = t + b$, where $t \in T, b \in N$. Recall that $\mathfrak{a} \in N$ and $N \perp T$. Then we have $\mathcal{J}_p \cdot p' = \mathcal{J}_p \cdot b$ and $\|p' - \mathfrak{a}\|^2 = \|t\|^2 + \|\mathfrak{a} - b\|^2$. Since $\mathfrak{a}, b \in N$, we choose $\{v_1, \ldots, v_m\}$ as the coordinates of N and suppose $\mathfrak{a} = (a_1, \ldots, a_m)$, $b = (b_1, \ldots, b_m)$. Thus, ignoring high order errors we have

$$\|f(p')\|^2 = \|\mathcal{J}_p \cdot b\|^2 = b^T V \Sigma^2 V^T b = \sum_{i=1}^{m} \sigma_i^2 b_i^2.$$

Hence, p' is a point near p satisfying the following problem

$$\min_{t,b} \ \mu = \big(\beta(\sum_{i=1}^{m} \sigma_i^2 b_i^2) + \sum_{i=1}^{m}(a_i - b_i)^2 + \|t\|^2\big)/2.$$

It is straightforward to show that when $b_i = \frac{a_i}{\beta\sigma_i^2+1}$ and $t = 0$, the function μ attains the minimum $\sum_i \frac{\beta\sigma_i^2}{\beta\sigma_i^2+1} a_i^2/2$, which is less than $\mu(p) = \sum_i a_i^2/2 = D^2/2$.

Therefore,

$$\|p' - p\|^2 = \sum_i b_i^2 = \sum_i (\frac{a_i}{\beta\sigma_i^2+1})^2 \leq \sum_i (\frac{a_i}{\beta\sigma_m^2+1})^2 = (\frac{D}{\beta\sigma_m^2+1})^2.$$

Moreover, p' can be found by solving Eq. (4). □

Example 1. Consider the system $f = \{x^2 + y^2 - 2x, 2x^2 + 2y^2 - 4x\}$ and $\mathfrak{a} = (-0.8, 0.6)$. In this case $m < k$ holds. The real variety is a circle centered at $(1, 0)$ with radius 1. The point $p = (1 - \frac{3\sqrt{10}}{10}, \frac{\sqrt{10}}{10})$ has the minimal distance to $\mathfrak{a}$. Consequently, we have $D = 1, \sigma_m = 4.472$ and $r = \|p - p'\| \leq \mathfrak{e} = \frac{1}{20\beta+1}$. The behaviors of both the actual error r and the estimated error $\mathfrak{e}$ with increasing of β are given in Fig. 1, where the differences between the log of estimated errors and the log of actual errors are greater than 0.047 and less than 0.048. That is $0.895\mathfrak{e} < r < 0.897\mathfrak{e}$. Thus, the theoretical estimation is quite sharp.

Here increasing the value of β and producing more and more accurate roots aim to verify Theorem 2. Since the local minimum satisfies the rank condition, we can simply apply Gauss-Newton iteration [4] to improve the accuracy.

Remark 1. Since we only have a local minimum p' which is an approximation of p, a good estimate of σ_m can be obtained by $\mathcal{J}_{p'}$ for sufficiently small $\|p - p'\|$ because of Weyl's theorem [28].

This theorem only works for $\sigma_m > 0$. However, if σ_m is close to zero because of singularity of p or non-radicalness of the system f, the convergence will be very slow as $\beta \to \infty$. But Corollary 1 still applies.

Example 2. In an example of [31], $f = \{x_2^2 + x_3^2 - (2x_1 - x_1^2)^3\}$ and $\mathfrak{a} = (-0.5, -1, 0.1)$. The point $p = (0, 0, 0)$ with the minimal distance to $\mathfrak{a}$ is singular in $V_{\mathbb{R}}(f)$. To see the asymptotic behavior of the error given in Corollary 1, we plot the magnitude of the actual error against the magnitude of the penalty factor β in Fig. 2. Applying the `CurveFitting[LeastSquares]` command in Maple yields $\log(r) \doteq -0.538 - 0.202 \log(\beta)$.

4 Error Control for General Systems

Previously, we know that if a local minimum of (8) satisfies the rank condition $m = n - \ell$, the estimated error is of order $O(1/\beta)$ as can be observed in Fig. 1.

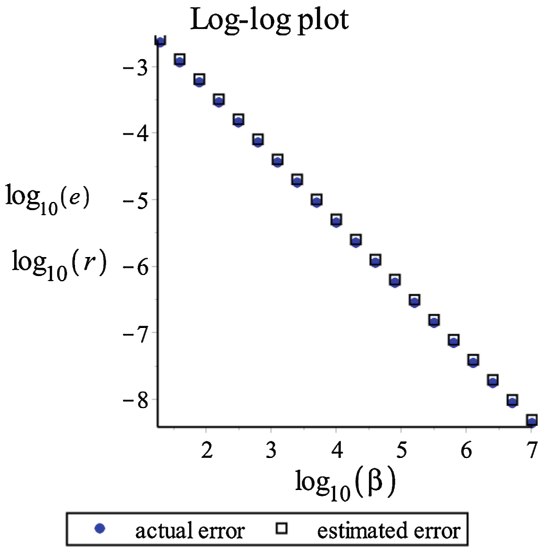

Fig. 1. For Example 1, the log of the estimated error, $log(\mathfrak{e})$ (resp. actual error, $log(r)$) is decreasing linearly with the increase of the magnitude of penalty factor β.

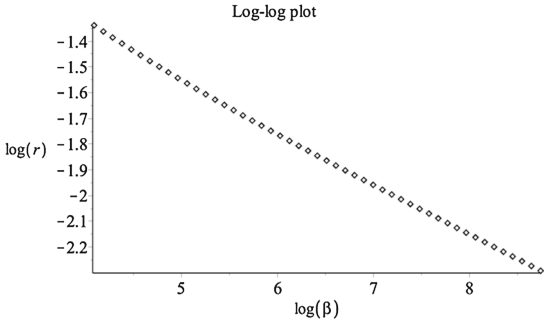

Fig. 2. The actual error $log(r)$ for Example 2 is decreasing slowly with slope -0.202.

For this case, although increasing the value of β can improve the accuracy, the Gauss-Newton iteration will improve convergence.

However, whether the input system satisfies the regularity assumptions or not, a local minimum of Problem (5) could be close to singularities (as in Example 2) which means $m < n - \ell$. Since such singularities are unavoidable, we will address the convergence for general systems with no assumptions in this section.

4.1 Degree Index

In Problem (5), the residual of f is amplified by a penalty factor β when x does not belong to $V_{\mathbb{R}}(f)$. In this section we will give a lower bound of the residual first. Then it leads to an error control of our method for general systems.

Let $f \in \mathbb{R}[x_1, \ldots, x_n]$. Suppose $f(0) = 0$. If we write $f = f_m + f_{m+1} + \ldots$ with f_α homogeneous of degree α and $f_m \neq 0$, then m is the multiplicity of f at the origin. For example $f = x^2 + y^3$ has a cusp at the origin with multiplicity 2.

Here we will consider the value of f near the origin. Let a direction be $v = (a, b)$ with $a^2 + b^2 = 1$. Then the bivariate polynomial $f = x^2 + y^3$ becomes a univariate polynomial $a^2t^2 + b^3t^3 = t^2(a^2 + b^3t)$ by substituting $(x = at, y = bt)$. For a generic direction v, the magnitude of f will be $O(t^2)$ as $t \to 0$. Here the degree 2 coincides with the multiplicity. But a lower bound is obtained for the direction $v = (0, 1)$ where the value of $f = t^3$ is even smaller of order 3.

In general we define **degree index** to study multiplicity discussed above.

Definition 3. *Let $f_v = f(vt)$ which is a polynomial in $\mathbb{R}[t]$ by substituting $x = vt$ into f with $v \neq 0 \in \mathbb{R}^n$. The lowest degree of nonzero terms of f_v is denoted by $\deg_{\min}(f_v)$. We define the **degree index** of f to be*

$$\deg_{ind}(f) = \max_v \deg_{\min}(f_v) \tag{10}$$

*Furthermore, for any polynomial f and a point $p \in \mathbb{R}^n$, if $f(p) = 0$ then we define the **degree index** of f at p to be $\deg_{ind}(f(x + p))$.*

For instance, $\deg_{ind}(x^2 - y^2) = 2$, $\deg_{ind}(x^2 + y^2) = 2$, and $\deg_{ind}(x^2 + y^3) = 3$, etc. But it is difficult to compute the degree index of an arbitrary multivariate polynomial f. It can be reduced to finding a nonzero common real root of the sequence $\{f_m = 0, f_{m+1} = 0, \ldots\}$. However, by definition, if $\deg(f) = d > 0$, then we have $1 \leq \deg_{ind}(f) \leq d$, which gives a simple bound.

Suppose $f_v(t) = a_0t^{\alpha_0} + a_1t^{\alpha_1} + \cdots + a_kt^{\alpha_k}$ is not a zero polynomial and $\deg_{\min}(f_v) = \alpha_0 < \alpha_1 < \cdots < \alpha_k$. The lowest degree term is $a_0t^{\alpha_0}$ which is the dominant term when $t \ll 1$. Thus, we have the following result.

Proposition 4. *Let $f \in \mathbb{R}[x_1, \ldots, x_n]$ and $f(0) = 0$. For any direction $v \neq 0 \in \mathbb{R}^n$, if $f_v(t) = f(vt)$ is not a zero polynomial, then there is a constant $c > 0$ such that $|f_v(t)| > c\, t^{\deg_{ind}(f)}$ for sufficiently small $t > 0$.*

As in Sect. 3, suppose the point $p \in V_{\mathbb{R}}(f)$ minimizing distance to a random point $\mathfrak{a}$ is singular. Let p' be the corresponding local minimum close to p of Problem (5). We have the following estimation for $\|p - p'\|$.

Theorem 5. *For a random point $\mathfrak{a} \in \mathbb{R}^n$ and a sufficiently large β, suppose $p \in V_{\mathbb{R}}(f)$ attains the local minimal distance to $\mathfrak{a}$. Then there is a solution p' of Eq. (4) such that $\|p' - p\| \leq O(\sqrt[2I-1]{1/\beta})$, where $I = \max\{\deg_{ind}(f_i(x + p)), i = 1, \ldots, k\}$.*

Proof. By Eq. (5), $\mu(p') = (\beta \sum_i f_i(p')^2 + \|p' - \mathfrak{a}\|^2)/2$. Let $D = \|p - \mathfrak{a}\|$. The relationship between points p and p' is shown in Fig. 3, where $r = \|p' - p\|$. Since D is the local minimal distance from $V_{\mathbb{R}}(f)$ to $\mathfrak{a}$ and $\mu(p') < \mu(p) = D^2/2$, we have $\|p' - \mathfrak{a}\| < D$ and $p' \notin V_{\mathbb{R}}(f)$ and the angle θ between $\overline{pp'}$ and $\overline{p\mathfrak{a}}$ is less

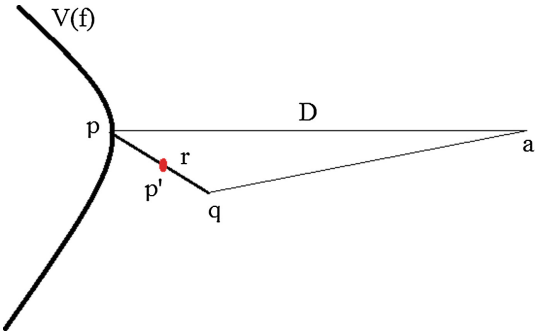

Fig. 3. The minimum value μ is attained at p' which is not a zero of f.

than $\pi/2$. Then $2\mu(p') = (D - r\cos\theta)^2 + r^2\sin^2\theta + \beta\sum_i f_i(p')^2 < D^2 = 2\mu(p)$, which is equivalent to $\beta\sum_i f_i(p')^2 + r^2 < 2r\cos\theta D$.

Thus $\beta\sum_i f_i(p')^2 < 2rD$ holds. Since $f(p') \neq 0$, there is at least one nonzero $f_i(p')$. Recall that $r = \|p' - p\|$, thus for the direction $v = (p' - p)/\|p - p'\|$, $f_i(vr + p) = f_i(p') \neq 0$, which implies that $f_i(vt + p)$ is a nonzero polynomial in t. By Proposition 4, when r is small enough, we have

$$f_i(p')^2 = f_i(vr + p)^2 > c^2\, r^{2\deg_{ind}(f_i(x+p))} \geq c^2\, r^{2I}.$$

Thus, we get $\beta c^2\, r^{2I} < 2rD \Rightarrow r^{2I-1} < O(1/\beta)$. □

Remark 2. If the input system f satisfies the rank condition, then we have $I = 1$ and $\|p' - p\| \leq O(1/\beta)$ in Theorem 5, which is in consistent with Theorem 2. But Theorem 2 provides a more precise estimation in this case.

Recall Example 2 in Sect. 3, $I = \deg_{ind}(f) = 3$. By Theorem 5, $r = C\sqrt[2I-1]{1/\beta}$ for some constant C. Then $\log(r) = \log(C) - \frac{1}{2I-1}\log(\beta) = \log(C) - 0.2\log(\beta)$ which is in close agreement with the experimental results.

4.2 Improve Accuracy

In contrast to Sect. 3, the input polynomial system may not satisfy the rank assumption. Consequently it is difficult to apply local methods such as Newton iteration to improve accuracy. For example if $f = x^2 + y^2$ and an approximate root of $f = 0$ close to 0 is given, it is still difficult to determine how to update the root because there is only one equation in f.

By Corollary 1, theoretically we can use Eq. (4) to update the approximate root x' by increasing β. But introducing a very large β will lead to numerical instability. To ease this difficulty, we substitute $\beta = 1/t$ into Eq. (4). Multiplying by t gives

$$t\begin{pmatrix} x_1 - \mathfrak{a}_1 \\ \vdots \\ x_n - \mathfrak{a}_n \end{pmatrix} + \mathcal{J}^t \cdot \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} = 0. \tag{11}$$

which can be considered as a homotopy with initial points in the form of (t_0, x_0), where $t_0 = 1/\beta$ with $\beta \gg 0$ and x_0 is a real solution of Eq. (4). When $t \to 0$, the homotopy path $x(t)$ will approach $V_{\mathbb{R}}(f)$. The invertibility of the Jacobian along the homotopy path is guaranteed by Lemma 2.

Let us reduce the value of t by a half at each step i.e. $t = \frac{1}{2^s \beta}$ after s steps. Combining with the result of Theorem 5, we have the following result.

Corollary 6. *Let $\tau = 2^{-1/(2d-1)} < 1$. After s steps of path tracking, the error of root is reduced to $O(\tau^s r)$, where r is the initial error $\|p' - p\|$.*

Example 3. Next we consider a sum of squares $f = x^2 + y^2$ with $\mathfrak{a} = (-1, 0.5)$. When $\beta = 1000$, the solution of (4) is the real point $(x = -0.0719, y = 0.0359)$ with a small residual $f(x, y) = 0.00646$. By tracking the path of the homotopy (11), it yields a sequence of points shown in Fig. 4. After 30 steps, we obtain the point $(x = 0.0000517, y = 0.000103)$ with residual $f(x, y) = 1.34 \times 10^{-8}$.

The `CurveFitting[LeastSquares]` command in Maple gives the formula $\log(r) \doteq -0.101 - 0.331 \log(\beta)$, where the coefficient -0.331 is very consistent with the formula $-\frac{1}{2I-1} = -0.333$ in Theorem 5, where $I = \deg_{ind}(x^2 + y^2) = 2$.

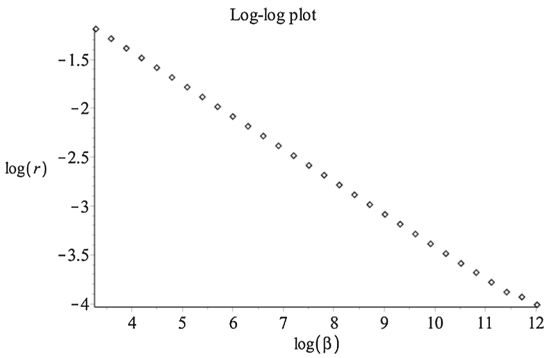

Fig. 4. The error for $f = x^2 + y^2$ in Example 3.

5 Examples

In this section, we demonstrate the generalized critical point method for finding real witness points of a general system on several examples. The numerical tool for solving the zero dimensional system (4) can be found in [18].

5.1 ISSAC 2016 System

Let us consider an interesting example $f = \{xyz, z(x^2 + y^2 + z^2 + y), y(y + z)\}$ in [7]. The real variety consists of a line $\{y = 0, z = 0\}$ and an isolated point $(0, -1/2, 1/2)$. We verify this result by `RealTriangularize` [12] in Maple 18.

To obtain the initial approximation, we set $\beta = 10000, \mathfrak{a} = (1, 0.5, 2)$ and solve the corresponding system (4) by `Hom4ps2` [18] with the output $\{[x = 0.00159, y = -0.499, z = 0.500], [x = 0.999, y = 0.0286, z = 0.000169]\}$. The first is an isolated solution with full rank Jacobian, so it can be refined simply. But the Jacobian at the second point is close to singular with singular values $\{1.03, 0.057, 0.0000045\}$. Using the homotopy (11), we can refine this point to $(1.0, 0.000044, 0.0)$ after 30 steps of path tracking. At the exact solution $(1, 0, 0)$ the degree index is 2. Hence, by Corollary 6 we have $\tau = 0.794$ and $\tau^{30} \times 0.0286 = 0.000028$ which has the same magnitude with 0.000044.

5.2 Seiler System

The system is $f = \{x_3^2 + x_2x_3 - x_1^2, x_1x_3 + x_1x_2 - x_3, x_2x_3 + x_2^2 + x_1^2 - x_1\}$ whose real variety is a curve. We use `RealTriangularize` to obtain a triangular set

$$\{(x_2 + x_3)\, x_1 - x_3, x_2^3 + 3\, x_3x_2^2 + 3\, x_3^2x_2 + x_3^3 - x_3\}.$$

Our method gives three initial points when $\beta = 10000$ and $\mathfrak{a} = (1, 1, 1)$, namely $p_1 = (0.233, 0.37, 0.113)$, $p_2 = (0.0546, -0.22, -0.013)$ and $p_3 = (1.12, 0.13, -1.19)$ with residuals $\|f(p_1)\| = 0.000096, \|f(p_2)\| = 0.00013, \|f(p_3)\| = 0.000014$. The rank of the Jacobian at p_1 is 2 with singular values $\{3.47, 2.06, 4.16 \times 10^{-6}\}$ which means f satisfies the rank condition. By Theorem 2, the accuracy can be improved quickly as β approaches infinity by the homotopy (11). In our experiment, the new residual becomes $\|f(p_1')\| = 1.35 \times 10^{-11}$ after 20 steps of path tracking. A similar situation happens for p_2 and p_3.

5.3 Larger Examples

Let f_1, f_2 and f_3 be random linear polynomials in variables $\{x_1, \ldots, x_n\}$ (where $n \geq 4$) and $f = \{(x_1^2 - x_2)^2 + f_1^2 + f_2^2, f_2^2 - f_3^2\}$. Since the first polynomial is a sum of squares, it implies that $x_1^2 - x_2 = 0, f_1 = 0, f_2 = 0, f_3 = 0$ which defines an $n - 4$ dimensional real variety of degree two. Let $f_1 = 97x_1 - 67x_2 + 58x_3 + 29x_4 + 37, f_2 = 5x_1 - 36x_2 - 57x_3 + 85x_4 + 80, f_3 = 90x_1 + 74x_2 + 27x_3 + 9x_4 - 91$.

Applying `RealTriangularize` to f yields a triangular set T in 3.6 s: $\{876997x_1 + 665882x_4 + 70645 = 0, 876997x_2 - 321399x_4 - 932414 = 0, 876997x_3 - 1046403x_4 - 635783 = 0, 443398837924x_4^2 - 187783491023x_4 - 812733564733 = 0\}$. Thus, there are two isolated solutions:

$$(0.799123750840176, 0.638598769156873, -0.657417515791706, -1.15857484076095),$$
$$(-1.28179034942671, 1.64298649988345, 2.61264348189493, 1.58208404954058).$$

Let $\beta = 10000$ and $\mathfrak{a} = (0, 1, 0.21, -0.053)$ and solve the corresponding square system (4) numerically by `Hom4ps2` in 3.09 s to obtain 4 real solutions, which are $(0.775, 0.642, -0.613, -1.12)$, $(0.775, 0.642, -0.613, -1.12)$, $(0.781, 0.650, -0.614, -1.12)$ and $(-1.24, 1.60, 2.53, 1.49)$.

We cannot refine these roots directly since f consists of only two equations. To improve the accuracy we apply the techniques introduced in Sect. 4.2 and after 30 steps of path tracking the refined roots are

$(0.799096390527587, 0.638616146725934, -0.657352004837882, -1.15851537630920)$,
$(0.799085595521195, 0.638625438788475, -0.657323985904906, -1.15848904058360)$,
$(0.799096848516397, 0.638616471399574, -0.657351954804033, -1.15851596503370)$,
$(-1.28174450424170, 1.64292754794384, 2.61255064161436, 1.58197011638993)$.

Apparently, the first three are multiple roots.

When $n = 5$, numerical solving for approximate points costs 17.6 s and refinement costs 0.26 s. It gives two real witness points. On the other hand, if we still use `RealTriangularize` to compute the triangular set of f, after 2902 s Maple 18 displayed an Error message and indicated that "Maple was unable to allocate enough memory to complete this computation".

Moreover, numerical solving stage costs 129 and 559 s for $n = 6$ and $n = 7$ respectively. Since it is difficult to compute the exact solutions, we verify the output by substituting the refined witness points back to f and the residuals are less than 10^{-7}.

6 Conclusions

This paper is part of a series in which we develop algorithms for numerical algebraic geometry. In current paper, we present a new formulation with penalty function, which is a development of critical point techniques. Comparing with existing numerical critical point methods, this method does not require the input system to have pure dimension or satisfy regularity assumptions. It leads to a kind of penalty function approximation method. The convergence rate of the method is given and it is in close agreement with our experimental results.

We plan to apply our method to larger systems, and those with approximate coefficients, which are beyond limitations of current (e.g. symbolic computation) based approaches. Since a non-radical polynomial system does not satisfy the rank condition, it is still difficult to move the obtained approximate real witness points on positive dimensional components to detect more geometric information. Extracting such information will be a focus of future work. In the second stage of our method, we assume local convergence of Newton iteration. An interesting research problem is to compare our approach with the certified path tracking approach [10].

Acknowledgements. The authors would like to thank the anonymous reviewers for their constructive comments that greatly helped improving the paper. This work is partially supported by the projects NSFC (11471307, 11671377, 61572024), cstc2015jcyjys40001, and the Key Research Program of Frontier Sciences of CAS (QYZDB-SSW-SYS026).

References

1. Aubry, P., Rouillier, F., El Din, M.S.: Real solving for positive dimensional systems. J. Symb. Comput. **34**(6), 543–560 (2002)
2. Basu, S., Pollack, R., Roy, M.-F.: Algorithms in Real Algebraic Geometry. Algorithms and Computation in Mathematics, vol. 10, 2nd edn. Springer, Heidelberg (2006). doi:10.1007/3-540-33099-2
3. Besana, G.M., DiRocco, S., Hauenstein, J.D., Sommese, A.J., Wampler, C.W.: Cell decomposition of almost smooth real algebraic surfaces. Numer. Algorithms **63**(4), 645–678 (2013)
4. Bjorck, A.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
5. Bank, B., Giusti, M., Heintz, J.: Point searching in real singular complete intersection varieties - algorithms of intrinsic complexity. Math. Comput. **83**(286), 873–897 (2014)
6. Bank, B., Giusti, M., Heintz, J., Mbakop, G.-M.: Polar varieties, real equation solving, and data structures: the hypersurface case. J. Complex. **13**, 5–27 (1997)
7. Brake, D.A., Hauenstein, J.D., Liddell, A.C.: Numerically validating the completeness of the real solution set of a system of polynomial equations. ISSAC **2016**, 143–150 (2016)
8. Bates, D.J., Hauenstein, J.D., Sommese, A.J., Wampler, C.W.: Adaptive multiprecision path tracking. SIAM J. Numer. Anal. **46**(2), 722–746 (2008)
9. Bates, D.J., Hauenstein, J.D., Sommese, A.J., Wampler, C.W.: Numerically Solving Polynomial Systems with the Software Package Bertini. SIAM, Philadelphia (2013)
10. Beltrán, C., Leykin, A.: Robust Certified Numerical Homotopy Tracking. Found. Comput. Math. **13**(2), 253–295 (2013)
11. Basu, S., Roy, M.-F., El Din, M.S., Schost, É.: A baby step-giant step roadmap algorithm for general algebraic sets. Found. Comput. Math. **14**(6), 1117–1172 (2014)
12. Chen, C., Davenport, J.H., May, J.P., Moreno Maza, M., Xia, B., Xiao, R.: Triangular decomposition of semi-algebraic systems. J. Symb. Comput. **49**, 3–26 (2013)
13. Collins, G.E.: Quantifier elimination for real closed fields by cylindrical algebraic decompostion. In: Brakhage, H. (ed.) GI-Fachtagung 1975. LNCS, vol. 33, pp. 134–183. Springer, Heidelberg (1975). doi:10.1007/3-540-07407-4_17
14. Davenport, J.H., Heintz, J.: Real quantifier elimination is doubly exponential. J. Symb. Comp. **5**, 29–35 (1988)
15. Hauenstein, J.: Numerically computing real points on algebraic sets. Acta Appl. Math. **125**(1), 105–119 (2013)
16. Hauenstein, J., Sommese, A.: What is numerical algebraic geometry? J. Symb. Comp. **79**, 499–507 (2017). Part 3
17. Hong, H.: Improvement in CAD-Based Quantifier Elimination. Ph.D. thesis. Ohio State University, Columbus, Ohio (1990)
18. Li, T.Y., Lee, T.L.: Homotopy method for solving Polynomial Systems software. http://www.math.msu.edu/~li/Software.htm
19. Lee, J.M.: Introduction to Smooth Manifolds, vol. 218. Springer, Heidelberg (2003). doi:10.1007/978-0-387-21752-9
20. Lasserre, J.B., Laurent, M., Rostalski, P.: Semidefinite characterization and computation of zero-dimensional real radical ideals. Found. Comput. Math. **8**(5), 607–647 (2008)

21. Lasserre, J.B., Laurent, M., Rostalski, P.: A prolongation-projection algorithm for computing the finite real variety of an ideal. Theoret. Comput. Sci. **410**(27–29), 2685–2700 (2009)
22. Lu, Y.: Finding all real solutions of polynomial systems. Ph.D thesis. University of Notre Dame (2006). Results of this thesis appear. In: (with Bates, D.J., Sommese, A.J., Wampler, C.W.), Finding all real points of a complex curve, Contemp. Math. vol. 448, pp. 183–205 (2006)
23. Ma, Y., Zhi, L.: Computing Real Solutions of Polynomial Systems via Low-rank Moment Matrix Completion. In: ISSAC, pp. 249–256 (2012)
24. Rouillier, F., Roy, M.-F., El Din, M.S.: Finding at least one point in each connected component of a real algebraic set defined by a single equation. J. Complex. **16**(4), 716–750 (2000)
25. El Din, M.S., Schost, É.: Polar varieties and computation of one point in each connected component of a smooth real algebraic set. In: ISSAC 2013, pp. 224–231 (2003)
26. El Din, M.S., Schost, É.: Properness defects of projection functions and computation of at least one point in each connected component of a real algebraic set. J. Discrete Comput. Geom. **32**(3), 417–430 (2004)
27. Sommese, A.J., Wampler, C.W.: The Numerical Solution of Systems of Polynomials Arising in Engineering and Science. World Scientific Press (2005)
28. Stewart, G.W.: Perturbation theory for the singular value decomposition. In: SVD and Signal processing, II: Algorithms, Analysis and Applications, pp. 99–109. Elsevier (1990)
29. Sommese, A.J., Verschelde, J., Wampler, C.W.: Introduction to numerical algebraic geometry. In: Bronstein, M., et al. (eds.) Solving Polynomial Equations. AACIM, vol. 14, pp. 339–392. Springer, Heidelberg (2005). doi:10.1007/3-540-27357-3_8
30. Sternberg, S.: Lectures on Differential Geometry. Prentice-Hall, Englewood Cliffs (1964)
31. Wu, W., Reid, G.: Finding points on real solution components and applications to differential polynomial systems. In: ISSAC, pp. 339–346 (2013)
32. Wu, W., Reid, G., Feng, Y.: Computing real witness points of positive dimensional polynomial systems. Accepted by Theoretical Computer Sciences (2017). http://doi.org/10.1016/j.tcs.2017.03.035
33. Yang, Z., Zhi, L., Zhu, Y.: Verified error bounds for real solutions of positive-dimensional polynomial systems. In: ISSAC, pp. 371–378 (2013)