非線形方程式の近似解の精度保証における候補者集合の生成

柏木 雅英

1 はじめに

精度保証付き数値計算はうんぬんかんぬん。

非線形方程式の解の精度保証を行う場合、精度保証付きで無い何らかの数値計算法で得られた解 (近似解) に対して、その誤差評価を行いたいことがある。この場合、その近似解を元にして何らかの「中に解を含みそうな」候補者集合を生成し、その集合に対して解の一意性定理を適用することになる。本論文では、この候補者集合の適切な生成法について議論する。

2 Krawczyk 法

精度保証付き数値計算で非線形方程式で解の存在を保証する場合、元の方程式 $f(x)=0$ を不動点形式に変換し、その成立を確かめるという手法を用いる。代表的な方法として、Krawczyk 法を簡単に説明しておく。

解きたい非線形方程式を

$$f(x)=0, \quad f:\boldsymbol{R}^n \to \boldsymbol{R}^n$$

解の存在を調べたい領域 (区間ベクトル) を $I \subset \boldsymbol{R}^n$ とする。c を I の中心、$R \simeq f'(c)^{-1}$ を正則行列とし、

$$c - Rf(c) + (E - Rf'(I))(I - c) \quad \subset \quad I \tag{1}$$

$$||E - Rf'(I)|| \quad < \quad 1 \tag{2}$$

が成立すると、I 内に $f(x)=0$ の解が唯一存在することが保証される。

これは、$g:\boldsymbol{R}^n \to \boldsymbol{R}^n$ を

$$g(x) = x - Rf(x)$$

と定義したとき、上記の条件が成立するならば g が I から I への縮小写像となることにより示される。

このように、区間 I を先に与え、そこに解が存在するかどうかを検証する定理なので、近似解が与えられてそれを元に誤差評価を行う場合は、 候補者集合を先に決めることが必要になる。

近似解を中心にしたある領域を候補者集合にする場合、その領域が小さすぎればその中に真の解が含まれず存在検証は失敗してしまう。一方、領域が大きすぎた場合は、その中に複数解を含んでしまったり、そうでなくても領域内の非線形性が強くなれば、やはり存在検証は失敗する。すなわち、候補者集合は大きすぎず小さすぎず適切な大きさに決める必要がある。

3　ニュートン法の修正量の 2 倍

適切な大きさの候補者集合を決めるためには、近似解と真解との大体の距離が必要になる。ニュートン法を一回行ってそのときに修正された距離が目安としてよく用いられる。近似解を中心とし、その修正量の「2 倍」を半径とした候補者集合を作るとうまく行くことが知られている。すなわち、

$$I = c + 2||f'(c)^{-1}f(c)|| \begin{pmatrix} [-1,1] \\ \vdots \\ [-1,1] \end{pmatrix} \tag{3}$$

のように候補者集合 I を作成する。

このことは、以下のように説明できる。Krawczyk 法の条件のうち、式 (2) は式 (1) が成立すればほぼ自動的に成立する。実際、(1) の左辺が右辺の内部に含まれる場合、適切なノルムを取れば (2) が成立することが示せる。以下、式 (1) について考える。

$$\begin{aligned} & c - Rf(c) + (E - Rf'(I))(I - c) \subset I \\ \Leftrightarrow \quad & -Rf(c) + (E - Rf'(I))(I - c) \subset I - c \\ \Leftrightarrow \quad & |-Rf(c)| + |E - Rf'(I))|u \leq u \end{aligned} \tag{4}$$

と変形できる。但し、$|\cdot|$ は区間行列または区間ベクトルの各成分の絶対値を並べた行列またはベクトルであり、$u = |I - c|$ とする。

以下、簡単のため「大雑把に」議論する。式 (4) を「1 変数化」して考える。更に、$|E - Rf'(I))|$ は I の幅 u が小さくなれば 0 に近付く量であるが、これが u に比例する、すなわち、

$$|E - Rf'(I))| \sim ku$$

と仮定する。また、修正量 $r = |-Rf(c)|$ と置く。

$$r + ku^2 \leq u$$

k が未知という状況で、この式が成立しやすいように u を決めれば良い。

$$k \leq \frac{u - r}{u^2}$$

と変形し、右辺がなるべく大きくなるように u を決めれば良いことになる。右辺は、

$$u = 2r$$

で最大値を取る (図 1)。これが、「ニュートン法の修正量の 2 倍」の理由である。

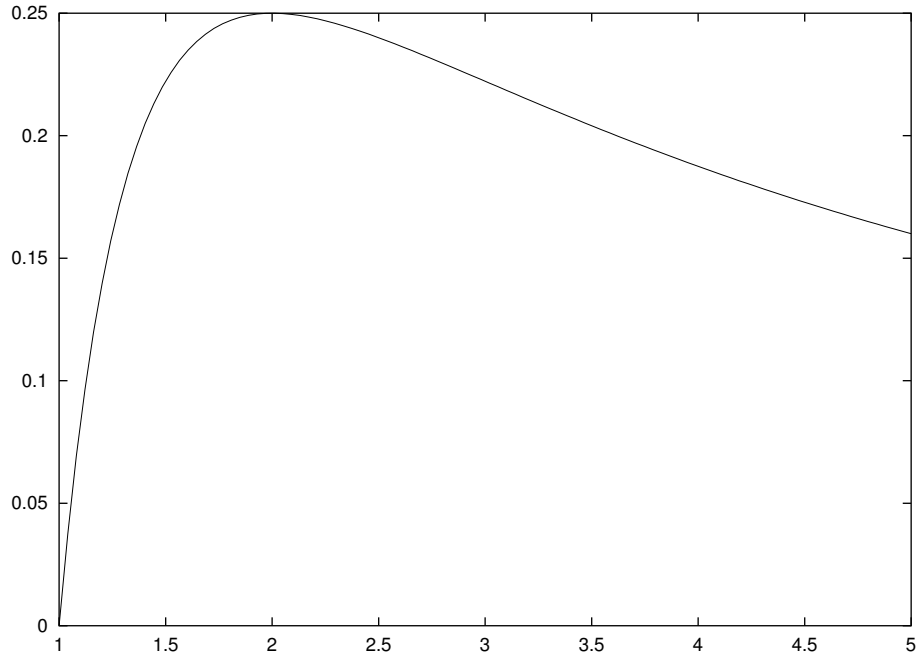

図 1: u の関数の概形 $(r = 1)$

4 成分毎に考える

第 3 節で示した方法は、必ず候補者集合として必ず超立方体領域を作成する。このことは、 スケーリングの問題を考えるとやや不自然な感がある。例えば、2 つの変数を持つ方程式

$$f(x, y) = 0$$

に対して、$x' = ax$ と変数変換しただけの実質的に同一の方程式

$$f(x'/a, y) = 0$$

を考えると、両者は実質的に同一の方程式であるにもかかわらず、作られる候補者集合の形状は異なることになる。

また、例えば方程式が未知数として電流 i と電圧 v を含むような場合、両者に同一の区間幅を設定するということは、電流と電圧という物理的に異なる量を比較していることになり、それは不自然であると言える。

そこで、ニュートン法の修正量に対して、最大値を取らずに成分毎の修正量の 2 倍で候補者集合を生成することが考えられる。すなわち、

$$I = c + 2|f'(c)^{-1} f(c)|[-1, 1] \tag{5}$$

のようにする。

しかしこの方法は、ニュートン法の修正量がある成分について移動量が 0 または 0 に非常に近い値だった場合、区間が潰れてしまったり、極端に幅の狭い区間になってしまった

りすることになる。区間が完全に潰れてしまっては内点を持たない集合になってしまってそもそも不動点定理の適用が難しく、また幅が極端に狭い場合も、Krawczyk 写像の出力がその中に入ることが難しくなってしまう。

5 提案手法

このように、今まで知られていた 2 つの方法は一長一短であり、どちらがいいとも言えず、どちらも問題を抱えている。そこで、あまり乱暴に議論せず少し丁寧に議論することにより、新しい手法を提案する。

解きたい非線形方程式を

$$f(x) = 0, \quad f : \boldsymbol{R}^n \to \boldsymbol{R}^n$$

$c \in \boldsymbol{R}^n$ を近似解、$R \simeq f'(c)^{-1}$ を正則行列とする。候補者集合 $I \subset \boldsymbol{R}^n$ を、$u \in \boldsymbol{R}^n, u \geq 0$ して

$$I = c + [-u, u]$$

で定めるとするとき、u をどのように決めればよいか考える。

式 (1) について、

$$\begin{aligned} & c - Rf(c) + (E - Rf'(I))(I - c) \subset I \\ \Leftrightarrow \quad & -Rf(c) + (E - Rf'(I))(I - c) \subset I - c \\ \Leftrightarrow \quad & r + |E - Rf'(I))|u \leq u \end{aligned} \tag{6}$$

と変形する。但し修正量 $r = |-Rf(c)|$ 、$M = |E - Rf'(I))|$ である。

ここで、$u \to 0$ としたとき、

$$M_{ij} \sim \sum_k K_{ijk} u_k$$

と仮定する。すなわち、漸近的に入力区間幅に比例すると仮定する。すると、式 (6) は、

$$r + (Ku)u \leq u \tag{7}$$

と書ける。r が既知、K が未知という状況で、これが成立しやすいように u を決めればよい。

このままでは難しいので K を簡略化し、$K_{ijk} = z$ (全成分が等しい) と仮定する。すると

$$(Ku)u = z(\sum_i u_i)^2 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

なので、 式 (7) は

$$z \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \leq \frac{u - r}{(\sum_i u_i)^2}$$

となる。z は未知なので、右辺 (の最小値) がなるべく大きくなるように u を決めればよい。すなわち、

$$\frac{\min(u_1 - r_1, u_2 - r_2, \cdots, u_n - r_n)}{(\sum_i u_i)^2}$$

が最大となるような u を決めればよい。

$(\sum_i u_i)^2$ を固定して考えれば、これが最大になるとき

$$u_1 - r_1 = u_2 - r_2 = \cdots = u_n - r_n$$

となることは自明。この値を α と置く。このとき、

$$\sum_i u_i = n\alpha + \sum_i r_i$$

なので、

$$\frac{\alpha}{(n\alpha + \sum_i r_i)^2}$$

を最大にする α を求める問題に帰着する。この式は、

$$\alpha = \frac{1}{n}\sum_i r_i$$

で最大となる。よって、最適な u は、

$$u_i = r_i + \frac{1}{n}\sum_k r_k$$

となる。

この値は、(その成分の r_i) + (r_i の平均) であり、前半が成分毎、後半が共通の値で、3 節、4 節で与えた 2 つの方法を混合したようなものになっている。よって、それぞれの欠点をうまく克服してくれることが期待できる。

6 数値実験による比較

多少精密になったとは言え、前節の方法でもいくらかの仮定が行われているので、どのような場合でも必ずうまくいくことを保証できるようなものではない。従って、数値実験によりその性能を検証する必要がある。以下、方法 1~3 を以下の表の通りとする。

方法 1	3 節の方法	$u_i = 2\max_k r_k$
方法 2	4 節の方法	$u_i = 2r_i$
方法 3	5 節の方法	$u_i = r_i + \frac{1}{n}\sum_k r_k$

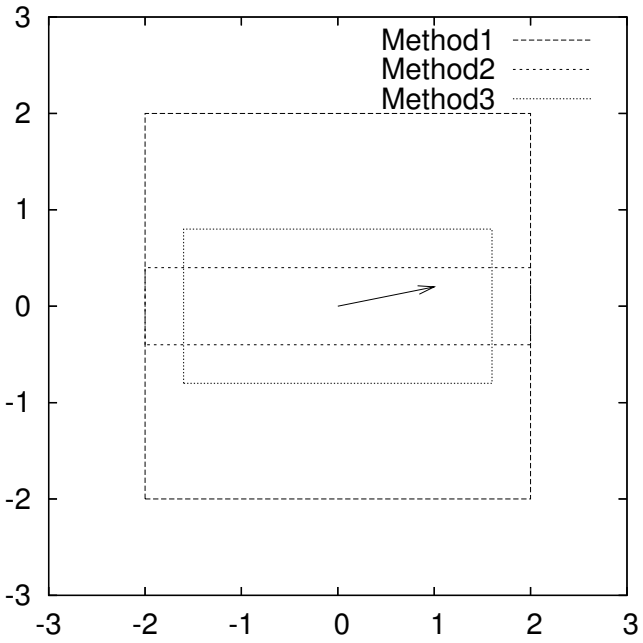

図 2: 各手法で生成される区間の違い

各手法で生成される区間の違いを簡単に説明する。2 次元の問題で、近似解から Newton 法を一回行い、修正ベクトルが $(1, 0.2)$ であったとしよう。このとき、方法 1 は半径 $(2, 2)$ の区間、方法 2 は半径 $(2, 0.4)$ の区間、方法 3 は半径 $(1.6, 0.8)$ となる。図示すると、図 2 のようになる。方法 3 が方法 1、方法 2 の中間になっていることが分かる。

性能比較の例として、まず例題 Himmelblau[1]:

$$\begin{aligned} -42x_1 + 2x_2^2 + 4x_1x_2 + 4x_1^3 - 14 &= 0 \\ -26x_2 + 2x_1^2 + 4x_1x_2 + 4x_2^3 - 22 &= 0 \end{aligned}$$

の、解 $(x_1, x_2) = (-0.127961, -1.95371)$ の周辺を調べた例を示す。その解から ± 0.2 の範囲の正方形内に近似解を乱数で 10000 点発生させ、その近似解を元に方法 1, 2, 3 で候補者集合を作成し、その候補者集合で Krawczyk 法が成功するかどうかを調べる。

方法 1, 2, 3 それぞれで Krawkzyk 法が成功した近似解をプロットしたものを図 3、図 4、図 5 に示す。

これを見ると、方法 2 では特定の方向でうまく行っていないこと、方法 3 は広い範囲で成功していることが分かる。

近似解の真解からの距離と成功率の関係をグラフにしたものを図 6 に示す。方法 2 では特定の方向に弱いためかなり真解に近い場合でも成功率が 100%に達していない、方法 3 ではかなり遠い距離でも一定の成功率を示している、全ての距離に渡って方法 3 がもっとも成功率が高い、などのことが分かる。

次に、簡単な方程式に対して、スケーリングの影響を調べてみる。図 7 は、方程式

$$x_1^2 + x_2^2 - 1 = 0$$

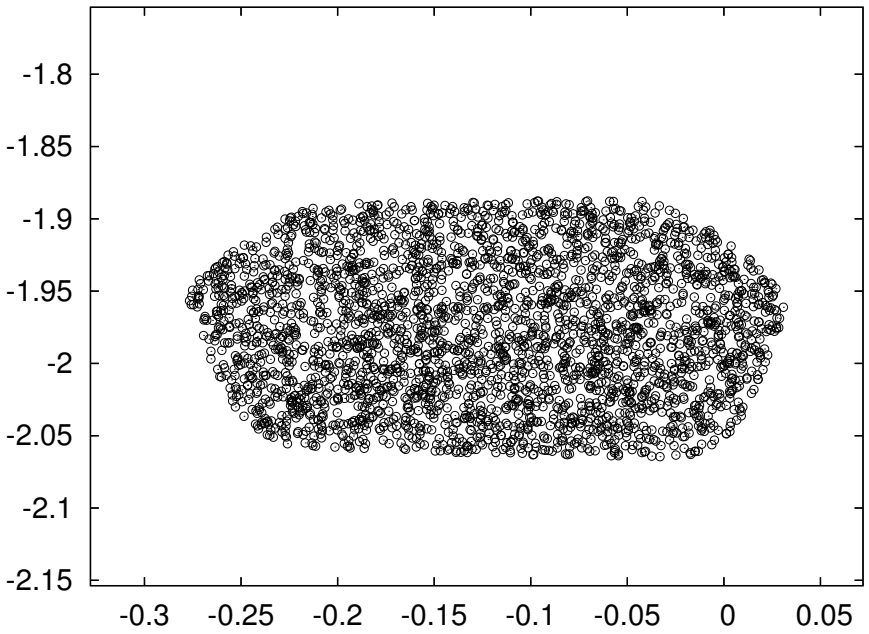

図 3: 方法 1 の成功した近似解

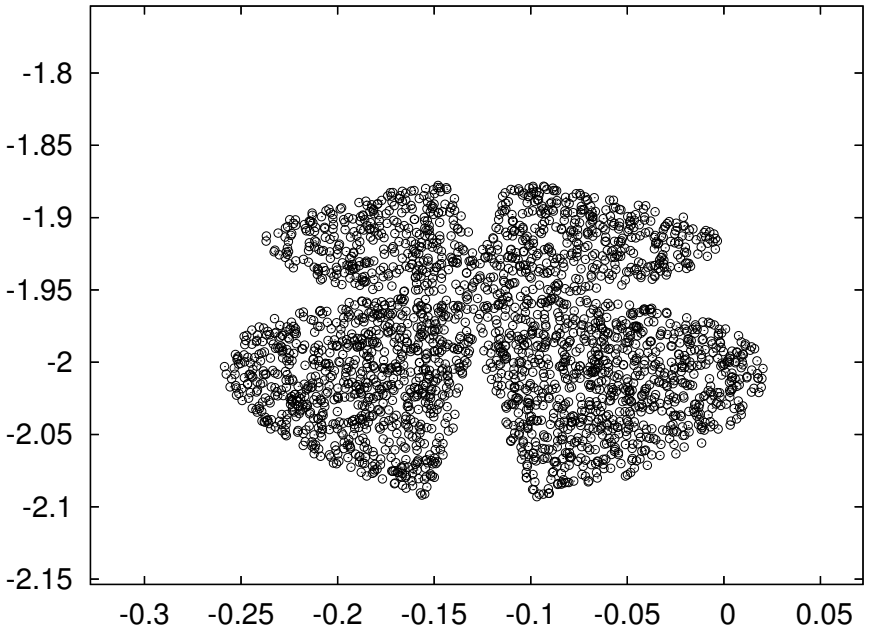

図 4: 方法 2 の成功した近似解

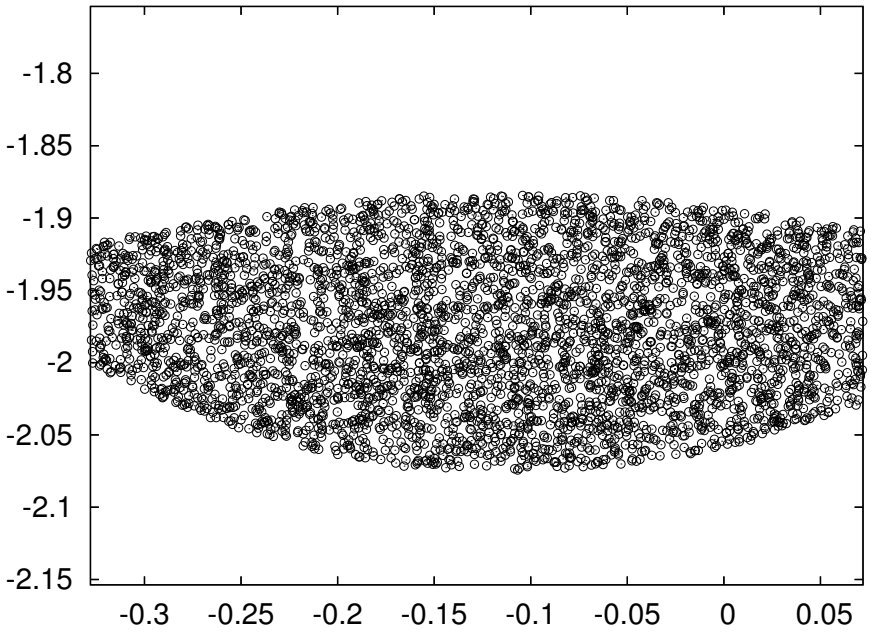

図 5: 方法 3 の成功した近似解

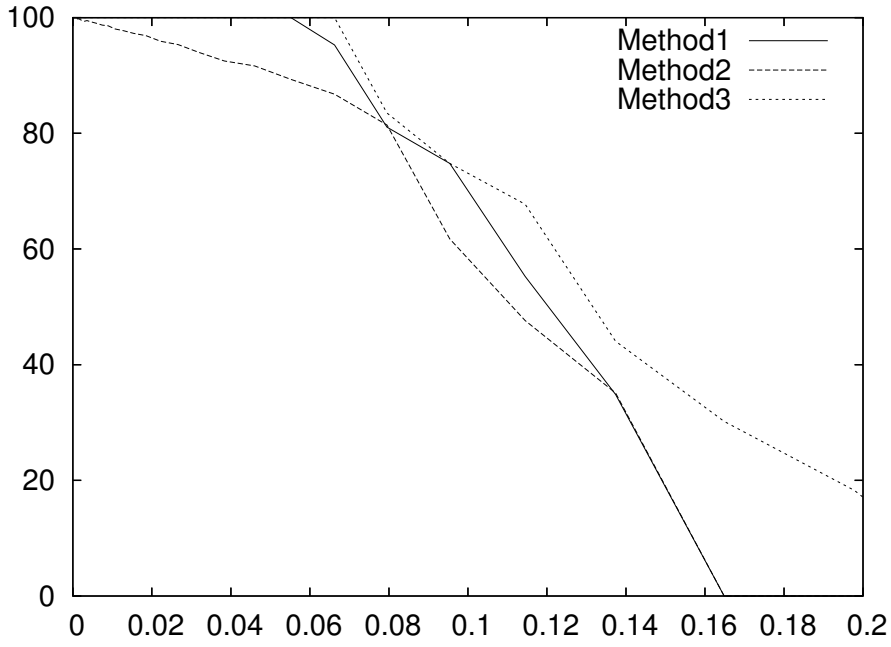

図 6: 真解からの距離と成功率 (Himmelblau)

$$x_1 - x_2 = 0$$

の、解 $(x_1, x_2) = (\sqrt{2}/2, \sqrt{2}/2)$ からの距離と成功率の関係である。方法 2 ははっきり悪く、方法 3 が一番よい。

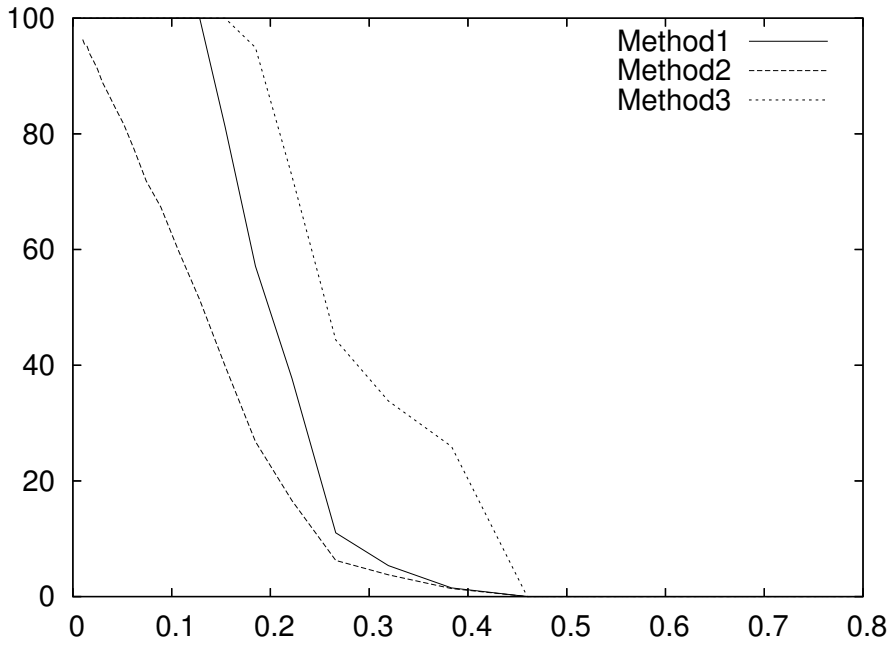

図 7: 真解からの距離と成功率 (Fn1)

図 8 は、方程式

$$\begin{aligned} x_1^2/4 + x_2^2 - 1 &= 0 \\ x_1/2 - x_2 &= 0 \end{aligned}$$

の、解 $(x_1, x_2) = (\sqrt{2}, \sqrt{2}/2)$ からの距離と成功率の関係である。図 7 とはスケーリングを変化させただけで実質的に同じ問題であるが、方法 1 の性能が低下し、方法 2 と逆転している箇所があることが分かる。方法 3 には大きな性能の低下は見られず、スケーリングの変化にも強いことが分かる。

図 9 は、方程式

$$\begin{aligned} 2(x_1^2 + x_2^2 - 1) &= 0 \\ x_1 - x_2 &= 0 \end{aligned}$$

の、解 $(x_1, x_2) = (\sqrt{2}/2, \sqrt{2}/2)$ からの距離と成功率の関係である。図 7 とは片方の式を 2 倍しただけの違いで、この場合は図 7 と変化が見られなかった。

次に、もう少し高次元の問題として、文献 [1] から Kincox(4 次元) と Bellido(9 次元) に対して試してみる。

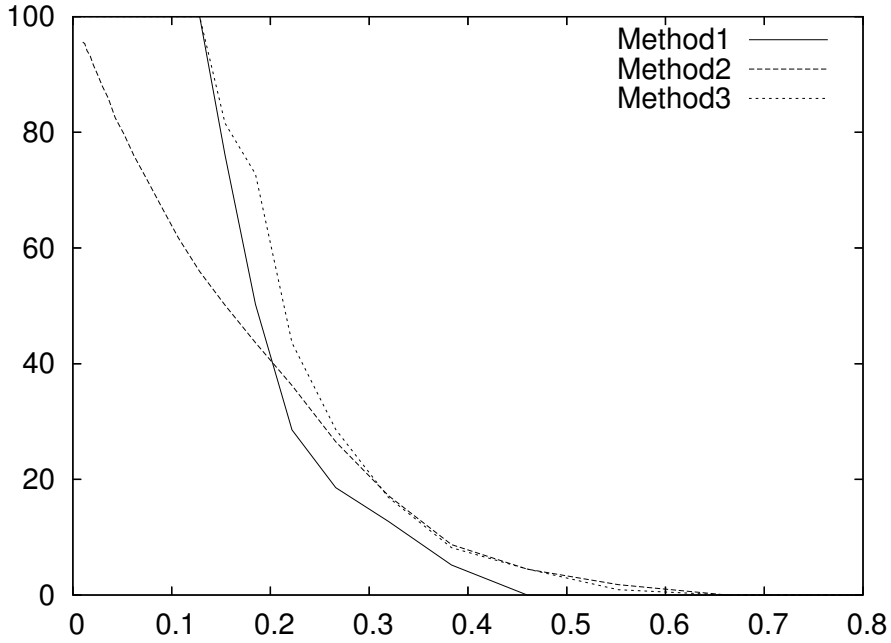

図 8: 真解からの距離と成功率 (Fn1s)

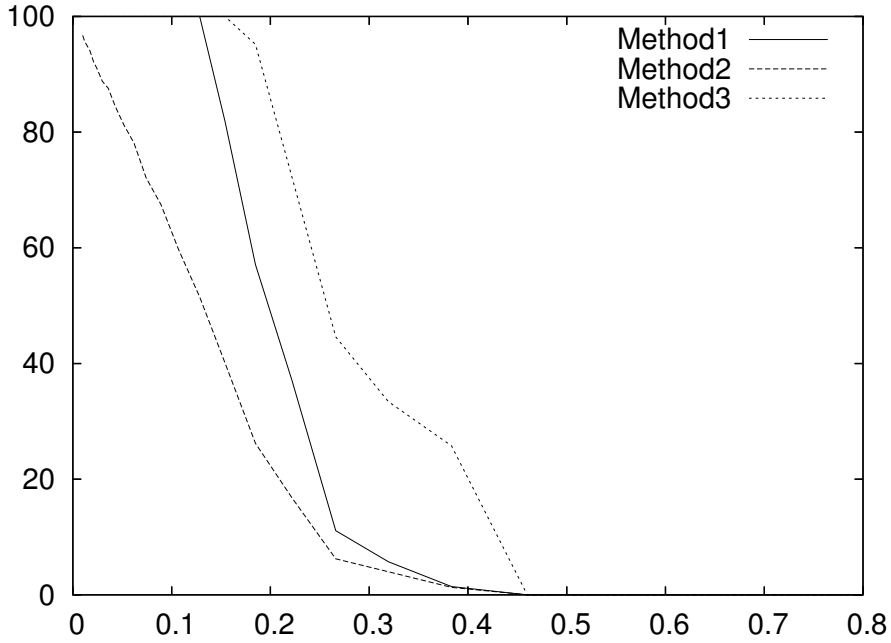

図 9: 真解からの距離と成功率 (Fn1s2)

図 10 は、Kincox[1]:

$$\begin{aligned} -1+6(c_1c_2-s_1s_2)+10c_1 &= 0 \\ -4+6(c_1s_2+c_2s_1)+10s_1 &= 0 \\ c_1^2+s_1^2-1 &= 0 \\ c_2^2+s_2^2-1 &= 0 \end{aligned}$$

の、解 $(c_1, c_2, s_1, s_2) = (0.420113, -0.991667, 0.907472, 0.12883)$ からの距離と成功率の関係、図 11 は、Bellido[1]:

$$\begin{aligned} (z_1-6)^2+z_2^2+z_3^2-104 &= 0 \\ z_4^2+(z_5-6)^2+z_6^2-104 &= 0 \\ z_7^2+(z_8-12)^2+(z_9-6)^2-80 &= 0 \\ z_1(z_4-6)+z_5(z_2-6)+z_3z_6-52 &= 0 \\ z_1(z_7-6)+z_8(z_2-12)+z_9(z_3-6)+64 &= 0 \\ z_4z_7+z_8(z_5-12)+z_9(z_6-6)-6z_5+32 &= 0 \\ 2z_2+2z_3-2z_6-z_4-z_5-z_7-z_9+18 &= 0 \\ z_1+z_2+2z_3+2z_4+2z_6-2z_7+z_8-z_9-38 &= 0 \\ z_1+z_3+z_5-z_6+2z_7-2z_8-2z_4+8 &= 0 \end{aligned}$$

の、解

$$\begin{aligned} &(z_1, \cdots, z_9) \\ &= (9.39167, 9.24763, 2.64156, 7.96267, 5.195, 6.32043, 5.01331, 6.50454, 10.9666) \end{aligned}$$

からの距離と成功率の関係である。

いずれも、方法 3 が優秀なことが分かる。

7 おわりに

提案手法 (方法 3) は、従来の方法 1、方法 2 に対して特に計算時間を要するわけでもなく、単に計算方法を変えるだけで大きな効果を発揮する。与えられた近似解の精度保証を行う場合には是非使ってみて欲しい。

参考文献

[1] The COPRIN example page (`http://www-sop.inria.fr/coprin/logiciels/ALIAS/Benches/benches.html`).

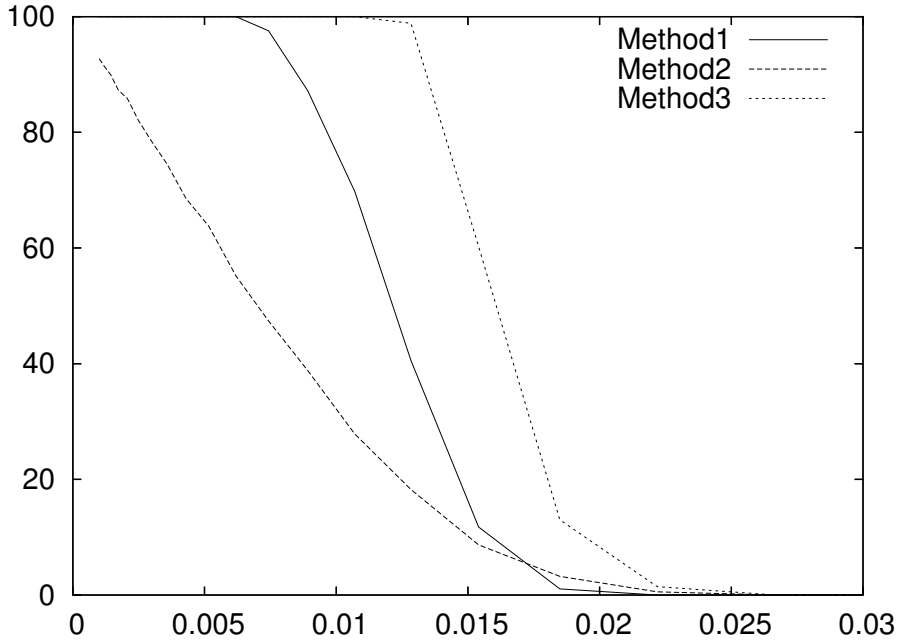

図 10: 真解からの距離と成功率 (Kincox)

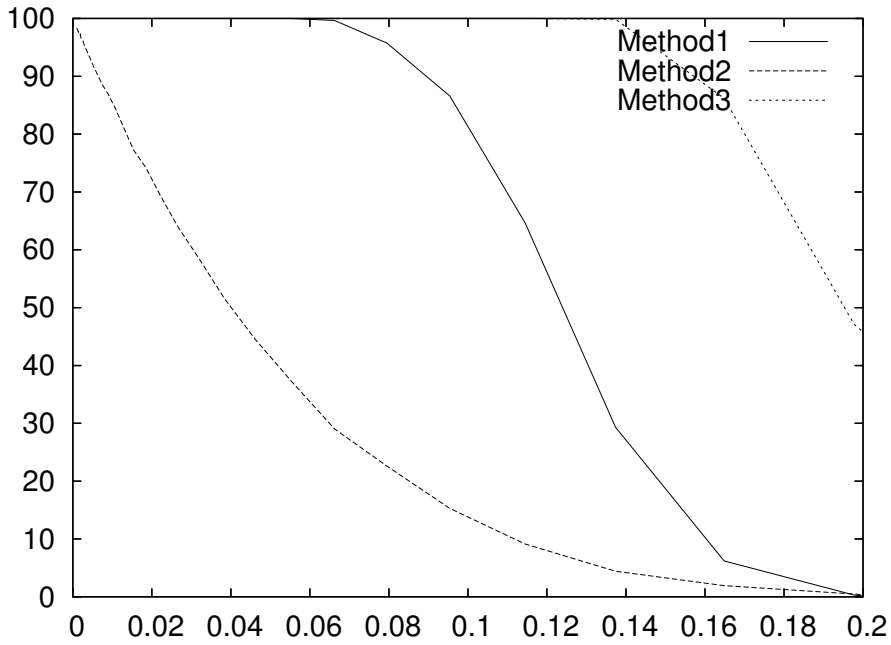

図 11: 真解からの距離と成功率 (Bellido)