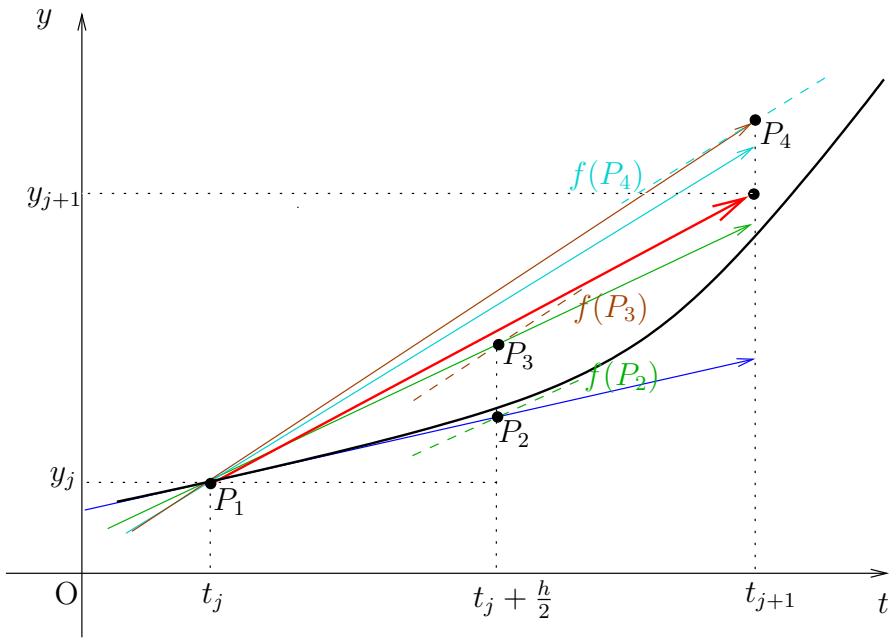
Εθνικο & Καποδιστριακο Πανεπιστημιο Αθηνων
Τμημα Πληροφορικης & Τηλεπικοινωνιων
Τομεας Θεωρητικης Πληροφορικης

Προηγμένοι Επιστημονικοί Υπολογισμοί

Σημειώσεις μεταπτυχιακού μαθήματος



Φίλιππος Τζαφέρης
ftzaf@di.uoa.gr

Παν. έτος 2023-24

Περιεχόμενα

Κεφάλαιο 1

Εντοπισμός ριζών πολυωνύμου

Ο αριθμητικός υπολογισμός των ριζών ενός πολυωνύμου είναι ένα πολύ ενδιαφέρον θέμα σε πολλούς τομείς της Αριθμητικής ανάλυσης όπως κατά την επίλυση διαφορικών εξισώσεων (ή συστημάτων διαφορικών εξισώσεων), ή όπου αλλού απαιτείται η εύρεση των ριζών του χαρακτηριστικού πολυωνύμου ενός πίνακα (πρόβλημα ιδιοτιμών).

Για τον αριθμητικό υπολογισμό των πραγματικών ριζών πολυωνύμου μπορούν να εφαρμοσθούν οι γνωστές μέθοδοι διχοτόμησης, τέμνουσας, Newton-Raphson. Ειδικότερα, λόγω του ενδιαφέροντος του προβλήματος και της ανάγκης να αναπτυχθούν μέθοδοι για τον υπολογισμό και των μιγαδικών ριζών ενός πολυωνύμου μελετούμε αναλυτικότερα το πρόβλημα αυτό.

1.1 Πολυώνυμα

Έστω το πολυώνυμο n βαθμού

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

όπου $n \geq 1,\ a_n \neq 0$ και $a_i \in \mathbb{C}$. Αναφέρουμε στη συνέχεια ορισμένες βασικές προτάσεις για τα πολυώνυμα, πολύ χρήσιμες για τον αριθμητικό υπολογισμό των ριζών αυτών.

Πρόταση 1.1. *Το πολυώνυμο $P(x)$ έχει τουλάχιστον μία ρίζα στο $\mathbb{C}$. Επομένως έχει ακριβώς n ρίζες $\rho_1, \rho_2, \ldots, \rho_n \in \mathbb{C}$ από τις οποίες μερικές ή ακόμη και όλες μπορεί να συμπίπτουν. Το πολυώνυμο γράφεται*

$$P(x) = a_n(x - \rho_1)(x - \rho_2)\cdots(x - \rho_n)$$

Αν ισχύει $P(x) = (x-\rho)^k Q(x)$ με $Q(\rho) \neq 0, k \in \mathbb{N}^*$, τότε λέμε ότι η ρ είναι ρίζα πολλαπλότητας k.

Πρόταση 1.2. *Αν $\rho_1, \rho_2, \ldots, \rho_n$ είναι οι ρίζες του $P(x)$, τότε ισχύουν οι γνωστοί τύποι Vieta:*

$$\begin{aligned} \rho_1 + \rho_2 + \cdots + \rho_n &= -\frac{a_{n-1}}{a_n} \\ \rho_1\rho_2 + \rho_1\rho_3 + \cdots + \rho_{n-1}\rho_n &= \frac{a_{n-2}}{a_n} \\ &\vdots \\ \rho_1\rho_2\rho_3\cdots\rho_n &= (-1)^n\frac{a_0}{a_n} \end{aligned}$$

Πρόταση 1.3. *Κάθε πολυώνυμο με πραγματικούς συντελεστές, αν έχει μια μιγαδική ρίζα $\rho = \alpha + \beta i,\ \beta \neq 0$ τότε θα έχει ως ρίζα και τη συζυγή αυτής $\overline{\rho} = \alpha - \beta i$. Επομένως ένα πολυώνυμο περιττού βαθμού έχει τουλάχιστον μια πραγματική ρίζα.*

Πρόταση 1.4. *Κανόνας του Descarte: Έστω το πολυώνυμο $P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ με $a_i \in \mathbb{R}$. Αν μ ο αριθμός των μεταβολών του προσήμου στην ακολουθία των συντελεστών $a_n,\ a_{n-1},\ a_{n-2}, \ldots,\ a_1,\ a_0$ (οι μηδενικοί συντελεστές, αν υπάρχουν, παραλείπονται), τότε ο αριθμός των θετικών ριζών του πολυωνύμου $P(x)$ ισούται με $k = \mu - 2\lambda$, όπου λ φυσικός με $0 \leq \lambda \leq \frac{\mu}{2}$.*

Εφαρμογή 1.1. Έστω $P(x) = x^7 - 2x^6 + x^4 - 3x^3 + 4$. Να βρείτε το μέγιστο αριθμό θετικών και αρνητικών ριζών του.

Έχουμε:

a_7	a_6	a_5	a_4	a_3	a_2	a_1	a_0
1	-2	0	1	-3	0	0	4
	1		2		3		4

'ρα $\mu = 4$ και $0 \leq \lambda \leq 2$, οπότε $\lambda = 0, 1, 2$.

- Για $\lambda = 0 \ \Rightarrow\ k = \mu - 2\lambda = 4 - 2\cdot 0 = \boxed{4}$
- Για $\lambda = 1 \ \Rightarrow\ k = \mu - 2\lambda = 4 - 2\cdot 1 = 2$
- Για $\lambda = 2 \ \Rightarrow\ k = \mu - 2\lambda = 4 - 2\cdot 2 = 0$

Επομένως το $P(x)$ έχει το πολύ 4 θετικές ρίζες.

Παρατήρηση 1.1. Η *Πρόταση 1.4* μπορεί να εφαρμοσθεί για να δώσει πληροφορίες σχετικά με τον αριθμό των αρνητικών ριζών ενός πολυωνύμου, αρκεί να θέσουμε στο $P(x)$ όπου x το $-x$: Είναι $P(x) = 0 \iff P_1(x) \equiv P(-x) = 0$, οπότε αν ρ είναι θετική ρίζα του P_1 τότε το $-\rho$ είναι αρνητική ρίζα του $P(x)$.

Εφαρμογή 1.2. Θέτουμε $P_1(x) = P(-x) = -x^7 - 2x^6 + x^4 + 3x^3 + 4$. Τότε στην ακολουθία των συντελεστών $-1 \quad -2 \quad 0 \quad 1 \quad 3 \quad 0 \quad 0 \quad 4$ έχουμε $\mu = 1$ αλλαγή προσήμου και $0 \leq \lambda \leq \frac{1}{2}$, οπότε $\lambda = 0$. 'ρα $k = 1 - 2 \cdot 0 = \boxed{1}$.

Επομένως το $P_1(x)$ έχει μία το πολύ θετική ρίζα, οπότε το $P(x)$ έχει μία το πολύ αρνητική ρίζα *.

Πρόταση 1.5. *Για κάθε ρίζα ρ του $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ ισχύει ο παρακάτω τύπος φράγματος:*

$$|\rho| \leq \frac{|a_0| + |a_1| + \cdot + |a_n|}{|a_n|}$$

1.2 Αλυσίδα του Sturm

Ορισμός 1.1. *Αν οι συναρτήσεις $f_0(x),\ f_1(x),\ f_2(x), \ldots,\ f_n(x)$ είναι ορισμένες στο $[\alpha, \beta]$ και ικανοποιούν τις προϋποθέσεις:*

α΄) $f_i \in C([\alpha, \beta]),\ \forall\, i = 0(1)n$

β΄) $f_0(\alpha) \neq 0,\ f_0(\beta) \neq 0$ *και* $f_n(x) \neq 0,\ \forall\, x \in [\alpha, \beta]$

γ΄) Αν για $\xi \in [\alpha, \beta]$ ισχύει $f_i(\xi) = 0$, τότε $f_{i-1}(\xi) \cdot f_{i+1}(\xi) < 0$, όπου $1 \leq i \leq n-1$

δ΄) Αν για $\xi \in [\alpha, \beta]$ ισχύει $f_0(\xi) = 0$ τότε υπάρχει $h > 0$ οσοδήποτε μικρό τέτοιο ώστε $f_0(\xi - h) f_1(\xi) < 0$ και $f_0(\xi + h) f_1(\xi) > 0$.

Τότε θα λέμε ότι οι συναρτήσεις $f_i(x),\ i = 0(1)n$ αποτελούν μια αλυσίδα Sturm στο $[\alpha, \beta]$.

Πρόταση 1.6. *Αν οι συναρτήσεις $f_i(x),\ i = 0(1)n$ αποτελούν μια αλυσίδα Sturm στο διάστημα $[\alpha, \beta]$ και συμβολίσουμε με $\mu(\xi)$ το πλήθος των μεταβολών του προσήμου στην ακολουθία των αριθμών:*

$$f_0(\xi),\ f_1(\xi), \ldots,\ f_n(\xi)$$

Τότε η διαφορά $\mu(\alpha) - \mu(\beta)$ ισούται με τον αριθμό των ριζών της $f_0(x)$ στο $[\alpha, \beta]$.

*Το εν λόγω πολυώνυμο έχει στην πραγματικότητα δυο ζεύγη συζυγών μιγαδικών ριζών, 2 θετικές πραγματικές ρίζες και μια αρνητική.

Πρόταση 1.7. *Έστω $f_0(x) = P_n(x)$ ένα πολυώνυμο $n-$βαθμού χωρίς πολλαπλές ρίζες. Θέτουμε $f_1(x) = f_0'(x)$ και κατασκευάζουμε την ακολουθία των διαδοχικών υπολοίπων με τον αλγόριθμο του Ευκλείδη ως εξής:*

$$\begin{aligned} f_0(x) &= f_1(x)\pi_1(x) - f_2(x) \\ f_1(x) &= f_2(x)\pi_2(x) - f_3(x) \\ &\vdots \\ f_{n-2}(x) &= f_{n-1}(x)\pi_{n-1}(x) - f_n(x) \end{aligned}$$

όπου τα $\pi_i(x)$ είναι πολυώνυμα πρώτου βαθμού.

Τότε η ακολουθία των συμαρτήσεων $f_0(x),\ f_1(x), \ldots,\ f_n(x)$ αποτελεί μια αλυσίδα Sturm σε κάθε διάστημα $[\alpha, \beta]$ με $f_0(\alpha) \neq 0$ και $f_0(\beta) \neq 0$.

Παρατήρηση 1.2. Αν το πολυώνυμο $f_0(x)$ έχει στο (α, β) πολλαπλές ρίζες τότε μπορούμε να κατασκευάσουμε όπως παραπάνω τις συναρτήσεις $f_1,\ f_2, \ldots,\ f_k$ όπου $k < n$ και $f_{k+1} = 0$. Τότε όμως δεν σχηματίζεται αλυσίδα Sturm. Ισχύει όμως πάλι ότι η διαφορά $\mu(\alpha) - \mu(\beta)$ ισούται με το πλήθος των ριζών της $f_0(x)$ στο $[\alpha, \beta]$, όπου η κάθε μια ρίζα λαμβάνεται μόνο μια φορά ανεξάρτητα από την πολλαπλότητά της.

Εφαρμογή 1.3. Έστω $P(x) = x^4 - 2x^2 + 3x - 1$. Τότε έχουμε:

$$\begin{aligned} f_0(x) &= P(x) = x^4 - 2x^2 + 3x - 1 \\ f_1(x) &= f_0'(x) = 4x^3 - 4x + 3 \\ f_2(x) &= x^2 - \frac{9}{4}x + 1 \\ f_3(x) &= -\frac{49}{4}x + 6 \\ f_4(x) &= -\frac{331}{2401} \end{aligned}$$

και άρα έχουμε τον πίνακα:

x	$f_0(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$\mu(x)$	Συμπέρασμα
$-\infty$	$+\infty$	$-\infty$	$+\infty$	$+\infty$	$-\frac{331}{2401}$	3	$3-3=0$ καμία ρίζα στο $(-\infty,-2)$
-2	13	-21	$\frac{19}{2}$	$-\frac{37}{2}$	$-\frac{331}{2401}$	3	$3-2=1$ ρίζα στο $[-2,0]$
0	-1	3	1	6	$-\frac{331}{2401}$	2	$2-1=1$ ρίζα στο $[0,1]$
1	1	3	$-\frac{1}{4}$	$-\frac{25}{4}$	$-\frac{331}{2401}$	1	$1-1=0$ καμία ρίζα στο $(1,+\infty)$
$+\infty$	$+\infty$	$+\infty$	$+\infty$	$-\infty$	$-\frac{331}{2401}$	1	

Για τη διευκόλυνση στην εκτέλεση των διαδοχικών διαιρέσεων του Ευκλειδείου αλγορίθμου, δίνουμε έναν αλγόριθμο με τον οποίο εκτελείται η διαίρεση ενός πολυωνύμου

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

με ένα άλλο πολυώνυμο:

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0 \qquad \text{όπου} \qquad m \leq n$$

Αλγόριθμος 1.1. *Διαίρεση Πολυωνύμων.*

1. *Διάβασε* $n,\ a_i,\ i = 0(1)n,\ m,\ b_i,\ i = 0(1)m$
2. *Για* $k = n - m(-1)0$
   - 2.1. $c_k = \dfrac{a_{m+k}}{b_m}$
   - 2.2. *Για* $j = m + k - 1(-1)k$
     $a_j = a_j - c_k b_{j-k}$
3. *Τύπωσε[Πηλίκο:* $c_k,\ k = n - m(-1)0$*, Υπόλοιπο:* $a_j,\ j = m - 1(-1)0$*]*

Στον ανωτέρω αλγόριθμο θέτουμε τους συντελεστές των διαδοχικών υπολοίπων στις θέσεις των συντελεστών a_j για λόγους οικονομίας μνήμης. Έτσι προκύπτουν τελικά οι συντελεστές $a_j,\ j = m - 1(-1)0$ του τελικού υπολοίπου. Τα $c_k,\ k = n - m(-1)0$ είναι οι συντελεστές του πηλίκου.

1.3 Φράγματα των ριζών πολυωνύμου

Έστω το πολυώνυμο $P(x) = x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$, όπου $a_k \in \mathbb{R}$ ή $a_k \in \mathbb{C}$. Για την εύρεση διαστήματος του $\mathbb{R}$ ή δίσκου του $\mathbb{C}$, μέσα στο οποίο βρίσκονται όλες οι ρίζες ρ_k του $P(x)$, έχει δοθεί ένα πλήθος τύπων. Παρακάτω παραθέτουμε ορισμένους σταθερούς αριθμούς r τέτοιους ώστε να ισχύει: $|\rho_k| \leq r\ ,\ \forall\, k = 1(1)n$

1. $r = 1 + A\ \ ,\ \ A = \max\limits_{k=1(1)n} \{|a_k|\}$
2. $r = (1 + |a_1|^2 + \cdots + |a_n|^2)^{\frac{1}{2}}$
3. $r = |a_1| + \sqrt{|a_2|} + \sqrt[3]{|a_3|} + \cdots + \sqrt[n]{|a_n|}$

4. $r = \max\left\{B, \sqrt[n]{B}\right\} = \begin{cases} B & , B \geq 1 \\ \sqrt[n]{B} & , B < 1 \end{cases}, \quad B = |a_1| + |a_2| + \cdots + |a_n|$

5. $r = \min\{C, D\}, \quad C = \max\{1, B\}, \quad D = \max\{1 + |a_1|, 1 + |a_2|, \ldots, 1 + |a_{n-1}|, |a_n|\}$

6. $r = \max\limits_{k=1(1)n} \left\{\sqrt[k]{n|a_k|}\right\}$

7. $r = \frac{1}{2}\left(1 + \sqrt{1 + 2E}\right), \quad E = \max\limits_{k=1(1)n}\{|a_1 a_k - a_{k+1}|\}, \quad a_{n+1} = 0$

8. $r = 1 + \sqrt{F}, \quad F = \max\limits_{k=1(1)n}\{|(1 - a_1)a_k - a_{k+1}|\}, \quad a_{n-1} = 0$

9. $r = \frac{1}{2}\left(1 + |a_1| + \sqrt{(1 - a_1)^2 + 4A_1}\right) \quad , \quad A_1 = \max\limits_{k=2(1)n}\{|a_k|\}$

10. $r = 1 + \left(1 - \frac{1}{(1 + A)^n}\right) A, \quad A = \max\limits_{k=1(1)n}\{|a_k|\}$

Εκτός από τα ανωτέρω φράγματα, που ορίζονται μέσω των σταθερών αριθμών r, ένα (άνω και κάτω) φράγμα πολύ χρήσιμο στην πράξη, το οποίο φράσσει το μέτρο όλων των ριζών ρ του πολυωνύμου $P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$, $a_k \in \mathbb{R}$ ή $a_k \in \mathbb{C}$, $a_0 \neq 0$ είναι το εξής:

$$\frac{|a_n|}{|a_n| + A_2} \leq |\rho| \leq \frac{|a_0| + A_1}{|a_0|}$$

όπου $A_1 = \max\limits_{k=1(1)n}\{|a_k|\}$, $A_2 = \max\limits_{k=0(1)n-1}\{|a_k|\}$.

Αν το δοθέν πολυώνυμο $P(x) = x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ έχει πραγματικούς συντελεστές και όλες του οι ρίζες είναι πραγματικές τότε αυτές βρίσκονται μέσα στο διάστημα, του οποίου τα άκρα είναι οι ρίζες της εξίσωσης

$$nz^2 + 2a_1 z + 2(n - 1)a_2 - (n - 2)a_1^2 = 0$$

Παρατήρηση 1.3. Αν θεωρήσουμε το πολυώνυμο $Q(x) = x^n P(\frac{1}{x}) = a_n x^n + \cdots + a_1 x + 1$ τότε για κάθε ρίζα ρ_k του $P(x)$ ισχύει

$$\frac{1}{r} \leq |\rho_k| \quad , \quad \forall\, k = 1(1)n$$

όπου r η σταθερά που προσδιορίζεται από τους ανωτέρω τύπους 1 έως 10 σε σχέση με το πολυώνυμο $Q(x)$.

Έτσι λοιπόν π.χ. σύμφωνα με τον τύπο 1 και για το πολυώνυμο $Q(x)$ θα έχουμε:

$$r_q = 1 + A_q \quad \text{όπου} \quad A_q = \max_{k=0(1)n-1}\left\{\left|\frac{a_k}{a_n}\right|\right\} \ , \ a_0 = 1$$

οπότε για κάθε ρίζα t_k του $Q(x)$ θα ισχύει

$$|t_k| \leq r_q \quad , \quad \forall\, k = 1(1)n$$

ενώ για κάθε ρίζα του $P(x)$, αφού $\rho_k = \dfrac{1}{t_k}$, έπεται

$$|\rho_k| \geq \frac{1}{r_q} = \frac{1}{1 + A_q}$$

1.4 Μέθοδος του Bernoulli

Με τη μέθοδο Bernoulli μπορούμε να υπολογίσουμε τις δυο πρώτες μικρότερες κατά μέτρο ρίζες ενός πολυωνύμου. Θεωρούμε το πολυώνυμο $n-$βαθμού

$$P_n(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0 \quad , \quad c_i \in \mathbb{C},\ c_n \neq 0$$

και υποθέτουμε χωρίς περιορισμό της γενικότητας, ότι το πολυώνυμο αυτό έχει ακριβώς n διακεκριμένες μη μηδενικές ρίζες, τις $\rho_1, \rho_2, \ldots, \rho_n$.

Η αρίθμηση έχει γίνει κατ΄ αύξουσα σειρά ως προς το μέτρο, δηλαδή $0 < |\rho_1| < |\rho_2| \leq \cdots \leq |\rho_n|$. Θεωρούμε τη ρητή συνάρτηση

$$R(x) = \frac{1}{P_n(x)}$$

Προφανώς οι πόλοι της $R(x)$ είναι οι ρίζες του πολυωνύμου $P_n(x)$ και αντίστροφα. ’ρα το πρόβλημα ανάγεται στην εύρεση των πόλων της συνάρτησης $R(x)$.

Αλγόριθμος 1.2. *Μέθοδος Bernoulli*

1. *Αναπτύσσουμε σε δυναμοσειρά με κέντρο το 0 και ακτίνα ρ_1 τη συνάρτηση $R(x)$ και έχουμε:*

$$R(x) = \sum_{k=0}^{\infty} \beta_k x^k \quad \Longleftrightarrow \quad \frac{1}{c_0 + c_1 x + \cdots + c_n x^n} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots$$

Επειδή $P_n(0) \neq 0$ έχουμε $c_0 \neq 0$ και χωρίς περιορισμό της γενικότητας υποθέτουμε $c_0 = 1$.

’ρα $1 = (1 + c_1 x + c_2 x^2 + \cdots + c_n x^n)(\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots)$. Εξισώνοντας τους συντελεστές των ομοιοβαθμίων όρων προκύπτει

$$
\begin{aligned}
& \beta_0 = 1 \\
& \beta_1 + \beta_0 c_1 = 0 \\
& \beta_2 + \beta_1 c_1 + \beta_0 c_2 = 0 \\
& \beta_3 + \beta_2 c_1 + \beta_1 c_2 + \beta_0 c_3 = 0 \\
& \vdots \\
& \beta_n + \beta_{n-1} c_1 + \beta_{n-2} c_2 + \cdots + \beta_0 c_n = 0 \\
\textit{για } k > n \qquad & \beta_k + \beta_{k-1} c_1 + \beta_{k-2} c_2 + \cdots + \beta_{k-n} c_n = 0 \\
& \vdots
\end{aligned}
$$

Το σύστημα αυτό λύνεται εύκολα ως προς $\beta_0, \beta_1, \ldots$ *και είναι*

$$\beta_k = -\sum_{\lambda=1}^{\mu} c_\lambda \beta_{k-\lambda} \qquad \textit{όπου} \qquad \mu = \min\{k, n\}$$

2. *Υπολογίζουμε τους όρους* $q_k,\ \varepsilon_k,\ p_k$ *από τους τύπους*

$$q_k = \frac{\beta_k}{\beta_{k-1}} \quad , \quad \varepsilon_k = q_{k+1} - q_k \quad , \quad p_k = \frac{\varepsilon_k}{\varepsilon_{k-1}} q_k \quad , \qquad k = 1, 2, \ldots$$

3. *Οι ακολουθίες* $q_k,\ p_k$ *συγκλίνουν αντίστοιχα προς τις αντίστροφες τιμές των ριζών:* $\frac{1}{\rho_1},\ \frac{1}{\rho_2}$. *Αν επιπλέον ισχύει* $|\rho_2| < |\rho_3|$ *η σύγκλιση είναι γραμμική με συντελεστή σύγκλισης* $\frac{\rho_1}{\rho_2}$.

Παρατήρηση 1.4. Αν οι ρίζες ρ_1 και ρ_2 είναι μιγαδικές συζυγείς, δηλαδή ισχύει $0 < |\rho_1| = |\rho_2| \leq \cdots \leq |\rho_n|$ τότε οι ακολουθίες q_k και p_k δεν συγκλίνουν προς τις αντίστροφες τιμές των ριζών $\rho_1,\ \rho_2$ (διότι δεν ισχύει $|\rho_1| < |\rho_2|$), αλλά αποδεικνύεται ότι οι ρίζες της δευτεροβάθμιας εξίσωσης

$$q_{k-1} p_k x^2 - (q_k + p_k) x + 1 = 0$$

συγκλίνουν για $k \to \infty$ προς τις ρίζες ρ_1 και ρ_2 αντίστοιχα.

Αξίζει να αναφερθεί ότι ο αλγόριθμος QD (Πηλίκων-Διαφορών) είναι μια επέκταση της μεθόδου Bernoulli για τον υπολογισμό όλων των ριζών ενός πολυωνύμου(με την υπόθεση $0 < |\rho_1| < |\rho_2| < \cdots < |\rho_n|$).

Αν αναλύσουμε τη συνάρτηση $R(x) = \frac{1}{P_n(x)}$ σε απλά κλάσματα έχουμε:

$$R(x) = \frac{a_1}{x - \rho_1} + \frac{a_2}{x - \rho_2} + \cdots + \frac{a_n}{x - \rho_n} \qquad \text{όπου} \qquad a_i \in \mathbb{C},\ a_i \neq 0$$

Για κάθε x με $|x| < |\rho_i|$ ισχύει (από τον τύπο της γεωμετρικής σειράς)

$$\frac{a_i}{x - \rho_i} = -\frac{a_i}{\rho_i - x} = -\frac{a_i}{\rho_i} \cdot \frac{1}{1 - \frac{x}{\rho_i}} = -\frac{a_i}{\rho_i} \left(1 + \frac{x}{\rho_i} + \frac{x^2}{\rho_i^2} + \cdots\right) = -a_i \left(\frac{1}{\rho_i} + \frac{x}{\rho_i^2} + \frac{x^2}{\rho_i^3} + \cdots\right)$$

Επομένως για $|x| < |\rho_1|$ η $R(x)$ αναπτύσσεται στη δυναμοσειρά

$$R(x) = \sum_{k=0}^{\infty} \beta_k x^k \quad \text{όπου} \quad \beta_k = -\left(\frac{a_1}{\rho_1^{k+1}} + \frac{a_2}{\rho_2^{k+1}} + \cdots + \frac{a_n}{\rho_n^{k+1}} \right) \tag{1.1}$$

Έστω τώρα ότι ισχύει $|\rho_1| < |\rho_2|$ τότε από την (1.1) έχουμε:

$$q_k := \frac{\beta_k}{\beta_{k-1}} = \frac{\frac{a_1}{\rho_1^{k+1}} + \frac{a_2}{\rho_2^{k+1}} + \cdots + \frac{a_n}{\rho_n^{k+1}}}{\frac{a_1}{\rho_1^{k}} + \frac{a_2}{\rho_2^{k}} + \cdots + \frac{a_n}{\rho_n^{k}}} = \frac{1}{\rho_1} \cdot \frac{1 + \frac{a_2}{a_1}\left(\frac{\rho_1}{\rho_2}\right)^{k+1} + \cdots + \frac{a_n}{a_1}\left(\frac{\rho_1}{\rho_n}\right)^{k+1}}{1 + \frac{a_2}{a_1}\left(\frac{\rho_1}{\rho_2}\right)^{k} + \cdots + \frac{a_n}{a_1}\left(\frac{\rho_1}{\rho_n}\right)^{k}} \tag{1.2}$$

Επειδή $|\rho_1| < |\rho_i|$ έχουμε $\left|\frac{\rho_1}{\rho_i}\right| < 1$ για κάθε $i = 2, 3, \ldots, n$ και άρα:

$$\lim_{k\to\infty} q_k = \frac{1}{\rho_1}$$

Ας μελετήσουμε τώρα την ταχύτητα σύγκλισης της ακολουθίας q_k, η οποία είναι $\lim_{k\to\infty} \frac{\varepsilon_k}{\varepsilon_{k-1}}$, όπου $\varepsilon_k = q_{k+1} - q_k$. Από την (1.2) έχουμε

$$\frac{1}{\rho_1} - q_k = \frac{1}{\rho_1} \cdot \frac{\frac{a_2}{a_1}\left(1 - \frac{\rho_1}{\rho_2}\right)\left(\frac{\rho_1}{\rho_2}\right)^{k} + \frac{a_3}{a_1}\left(1 - \frac{\rho_1}{\rho_3}\right)\left(\frac{\rho_1}{\rho_3}\right)^{k} + \cdots + \frac{a_n}{a_1}\left(1 - \frac{\rho_1}{\rho_n}\right)\left(\frac{\rho_1}{\rho_n}\right)^{k}}{1 + \frac{a_2}{a_1}\left(\frac{\rho_1}{\rho_2}\right)^{k} + \frac{a_3}{a_1}\left(\frac{\rho_1}{\rho_3}\right)^{k} + \cdots + \frac{a_n}{a_1}\left(\frac{\rho_1}{\rho_n}\right)^{k}}$$

ή

$$\frac{\frac{1}{\rho_1} - q_k}{\left(\frac{\rho_1}{\rho_2}\right)^{k}} = \frac{1}{\rho_1} \cdot \frac{\frac{a_2}{a_1}\left(1 - \frac{\rho_1}{\rho_2}\right) + \frac{a_3}{a_1}\left(1 - \frac{\rho_1}{\rho_3}\right)\left(\frac{\rho_2}{\rho_3}\right)^{k} + \cdots + \frac{a_n}{a_1}\left(1 - \frac{\rho_1}{\rho_n}\right)\left(\frac{\rho_2}{\rho_n}\right)^{k}}{1 + \frac{a_2}{a_1}\left(\frac{\rho_1}{\rho_2}\right)^{k} + \frac{a_3}{a_1}\left(\frac{\rho_1}{\rho_3}\right)^{k} + \cdots + \frac{a_n}{a_1}\left(\frac{\rho_1}{\rho_n}\right)^{k}} \tag{1.3}$$

Αν υποθέσουμε ότι επιπλέον ισχύει $|\rho_2| < |\rho_3|$ και θέσουμε στην (1.3) όπου k το $k+1$ τότε διαιρώντας κατά μέλη και παίρνοντας το όριο καθώς $k \to \infty$ έχουμε:

$$\lim_{k\to\infty} \frac{\frac{1}{\rho_1} - q_{k+1}}{\frac{1}{\rho_1} - q_k} = \frac{\rho_1}{\rho_2}$$

'ρα η σύγκλιση είναι γραμμική με συντελεστή σύγκλισης $\frac{\rho_1}{\rho_2}$. Αν θέσουμε $\varepsilon_k = q_{k+1} - q_k$ τότε με απλούς μετασχηματισμούς βρίσκουμε ότι:

$$\lim_{k\to\infty} \frac{\varepsilon_k}{\varepsilon_{k-1}} = \frac{\rho_1}{\rho_2}$$

'ρα για την ακολουθία

$$p_k = \frac{\varepsilon_k}{\varepsilon_{k-1}} q_k$$

ισχύει

$$\lim_{k\to\infty} p_k = \frac{1}{\rho_2}$$

1.5 Μέθοδος QD (Πηλίκων-Διαφορών)

Η σύγκλιση της μεθόδου QD είναι γραμμική (και συνεπώς αργή) και ευαίσθητη ως προς την επίδραση σφαλμάτων στρογγύλευσης, αλλά είναι πολύ χρήσιμη για την εύρεση αρχικών προσεγγίσεων των πραγματικών ριζών και των δευτεροβαθμίων παραγόντων που αντιστοιχούν σε κάθε ζεύγος συζυγών μιγαδικών ριζών. Αυτές οι προσεγγίσεις είναι απαραίτητες ως αρχικές τιμές σε μια πιο γρήγορα συγκλίνουσα μέθοδο, π.χ. τη μέθοδο Newton-Raphson.

Στη μέθοδο QD για την προσέγγιση των ριζών της πολυωνυμικής εξίσωσης $P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$ δημιουργούνται οι παρακάτω ακολουθίες πραγματικών αριθμών

$$\begin{aligned} &\left\{e_i^{(k)}\right\}_{i=1}^{\infty} \quad \forall\, k = 1, 2, \ldots, n+1 \qquad (n+1 \text{ το πλήθος}) \\ &\left\{q_i^{(k)}\right\}_{i=1}^{\infty} \quad \forall\, k = 1, 2, \ldots, n \qquad (n \text{ το πλήθος}) \end{aligned}$$

όπου

$$\begin{aligned} e_i^{(1)} &= 0 && \text{για } i = 1, 2, \ldots \\ e_i^{(n+1)} &= 0 && \text{για } i = 1, 2, \ldots \\ e_1^{(k)} &= \tfrac{a_{n-k}}{a_{n-k+1}} && \text{για } k = 2, 3, \ldots, n \\ q_1^{(1)} &= -\tfrac{a_{n-1}}{a_n} && \\ q_1^{(k)} &= 0 && \text{για } k = 2, 3, \ldots, n \\ q_{i+1}^{(k)} &= e_i^{(k+1)} + q_i^{(k)} - e_i^{(k)} && \text{για } k = 1, 2, \ldots, n \text{ και } i = 1, 2, \ldots \\ e_{i+1}^{(k)} &= \tfrac{q_{i+1}^{(k)} \cdot e_i^{(k)}}{q_{i+1}^{(k-1)}} && \text{για } k = 2, 3, \ldots, n \text{ και } i = 1, 2, \ldots \end{aligned}$$

Αν και η κατασκευή των ακολουθιών φαίνεται πολύπλοκη, στην πράξη μπορεί να υλοποιηθεί πολύ απλά με τη χρήση ενός πίνακα που κατασκευάζουμε εισάγοντας αρχικά όλες τις τιμές για τα $q_1^{(k)}$, $e_1^{(k)}$, $e_i^{(1)}$ και $e_i^{(n+1)}$:

i	$e_i^{(1)}$	$q_i^{(1)}$	$e_i^{(2)}$	$q_i^{(2)}$	$e_i^{(3)}$	$q_i^{(3)}$	$\cdots$	$e_i^{(n)}$	$q_i^{(n)}$	$e_i^{(n+1)}$
1	0	$-\frac{a_{n-1}}{a_n}$	$\frac{a_{n-2}}{a_{n-1}}$	0	$\frac{a_{n-3}}{a_{n-2}}$	0	$\cdots$	$\frac{a_0}{a_1}$	0	0
2	0									0
3	0									0
$\vdots$	$\vdots$									$\vdots$

Το επόμενο βήμα κατασκευάζει τις εισόδους $q_2^{(k)}$ στη δεύτερη γραμμή παίρνοντας το στοιχείο αμέσως πάνω και δεξιά, δηλαδή το $e_1^{(k+1)}$, προσθέτοντας το αμέσως από πάνω στοιχείο $q_1^{(k)}$ και

αφαιρώντας το στοιχείο πάνω αριστερά $e_1^{(k)}$:

i	$e_i^{(1)}$	$q_i^{(1)}$	$e_i^{(2)}$		$q_i^{(2)}$		$e_i^{(3)}$	$q_i^{(3)}$	$\cdots$	$e_i^{(n)}$	$q_i^{(n)}$	$e_i^{(n+1)}$
1	0	$-\frac{a_{n-1}}{a_n}$	$\frac{a_{n-2}}{a_{n-1}}$	$\xleftarrow{-}$	0	$\xleftarrow{+}$	$\frac{a_{n-3}}{a_{n-2}}$	0		$\frac{a_0}{a_1}$	0	0
				$\searrow^{=}$								
2	0	$q_2^{(1)}$			$\boxed{q_2^{(2)}}$			$q_2^{(3)}$			$q_2^{(n)}$	0
3	0											0
$\vdots$	$\vdots$											$\vdots$

Οι είσοδοι $e_2^{(k)}$ τώρα προκύπτουν παίρνοντας το δεξιό στοιχείο $q_2^{(k)}$, πολλαπλασιάζοντας με το από πάνω στοιχείο $e_1^{(k)}$ και διαιρώντας με το αριστερό στοιχείο $q_2^{(k-1)}$.

i	$e_i^{(1)}$	$q_i^{(1)}$	$e_i^{(2)}$	$q_i^{(2)}$		$e_i^{(3)}$		$q_i^{(3)}$	$\cdots$	$e_i^{(n)}$	$q_i^{(n)}$	$e_i^{(n+1)}$
1	0	$-\frac{a_{n-1}}{a_n}$	$\frac{a_{n-2}}{a_{n-1}}$	0		$\frac{a_{n-3}}{a_{n-2}}$		0		$\frac{a_0}{a_1}$	0	0
					$\div\swarrow$		$\nwarrow\times$					
2	0	$q_2^{(1)}$	$e_2^{(2)}$	$q_2^{(2)}$	$\xrightarrow{=}$	$\boxed{e_2^{(3)}}$		$q_2^{(3)}$	$\cdots$	$e_2^{(n)}$	$q_2^{(n)}$	0
3	0											0
$\vdots$	$\vdots$											$\vdots$

Παρατήρηση 1.5. Αν

$$\lim_{i\to\infty} e_i^{(k)} = \lim_{i\to\infty} e_i^{(k+1)} = 0$$

για κάποιο $k = 1, 2, \ldots, n$ τότε αποδεικνύεται ότι υπάρχει το $\lim_{i\to\infty} q_i^{(k)}$, το οποίο είναι μια ρίζα του πολυωνύμου $P(x)$. Επιπλέον, αν η ακολουθία $\left\{e_i^{(k)}\right\}_{i=1}^{\infty}$ δεν συγκλίνει στο 0 για κάποιο k τότε οι ακολουθίες $\left\{\tau_i^{(k)}\right\}_{i=1}^{\infty}$ και $\left\{s_i^{(k)}\right\}_{i=1}^{\infty}$ όπου

$$\begin{aligned} \tau_i^{(k)} &= q_i^{(k-1)} + q_i^{(k)} \quad , \quad i = 1, 2, \ldots \\ s_i^{(k)} &= q_{i-1}^{(k-1)} \cdot q_i^{(k)} \quad , \quad i = 2, 3, \ldots \end{aligned}$$

συγκλίνουν προς τους αριθμούς $\tau^{(k)}$ και $s^{(k)}$, αντίστοιχα, όπου το

$$x^2 - \tau^{(k)} x + s^{(k)}$$

είναι ένας δευτεροβάθμιος παράγοντας του $P_n(x)$ που αντιστοιχεί σε ένα ζεύγος συζυγών μιγαδικών ριζών.

Αλγόριθμος 1.3. *Μέθοδος QD*

1. *Διάβασε* n , $a_i, i = 0(1)n$, M

2. $e_1^{(1)} = 0$, $e_1^{(n+1)} = 0$, $q_1^{(1)} = -\frac{a_{n-1}}{a_n}$, $i = 2$,
   $IND_1 = 1$, $IND_{n+1} = 1$ *(*$IND_k = 1$ *δηλώνει σύγκλιση της* $e_i^{(k)}$ *στο 0)*

3. *Για* $k = 2(1)n$
   $q_1^{(k)} = 0$, $e_1^{(k)} = \frac{a_{n-k}}{a_{n-k+1}}$, $IND_k = 0$, $\tau_1^{(k)} = 0$, $s_1^{(k)} = 0$

4. *Όσο* $i \leq M$ *επανάλαβε*

   4.1. $e_i^{(1)} = 0$, $e_i^{(n+1)} = 0$, $q_i^{(1)} = e_{i-1}^{(2)} + q_{i-1}^{(1)} - e_{i-1}^{(1)}$

   4.2. *Για* $k = 2(1)n$
   $q_i^{(k)} = e_{i-1}^{(k+1)} + q_{i-1}^{(k)} - e_{i-1}^{(k)}$
   $e_i^{(k)} = (q_i^{(k)} \cdot e_{i-1}^{(k)})/q_i^{(k-1)}$
   $\tau_i^{(k)} = q_i^{(k-1)} + q_i^{(k)}$
   $s_i^{(k)} = q_{i-1}^{(k-1)} \cdot q_i^{(k)}$

   4.3. *Για* $k = 2(1)n$
   Αν $e_j^{(k)} \to 0$ *τότε* $IND_k = 1$
   Αν $e_j^{(k)} \not\to 0$ *τότε* $IND_k = -1$ *(αν η εκλογή δεν είναι φανερή τότε* $IND_k = 0$*)*

   4.4. $NS = 0$ *(δηλώνει ότι όλες οι* $e_j^{(k)}$ *συγκλίνουν)*

   4.5. *Για* $k = 2(1)n$, *Αν* $IND_k = 0$ *τότε* $NS = 1$

   4.6. *Αν* $NS = 0$ *τότε*

      4.6.1. *Για* $k = 2(1)n + 1$
      Αν $IND_k = 1$ *και* $IND_{k-1} = 1$ *τότε Τύπωσε("Προσεγγιστική ρίζα: ",* q_i^{k-1}*)*
      αλλιώς Τύπωσε("Προσεγγιστικός δευτεροβάθμιος παράγοντας:", $x^2 - \tau_i^{(k)}x + s_i^{(k)}$*)*

      4.6.2. *Τέλος.*

   4.7. $i = i + 1$

5. *Τύπωσε("Υπέρβαση μέγιστου αριθμού επαναλήψεων:" ,* M*),* *Τέλος.*

Παρατήρηση 1.6. Είναι φανερό ότι η μέθοδος QD απαιτεί να μην υπάρχουν μηδενικοί συντελεστές στο πολυώνυμο. Έστω $P_n(x) = \sum_{k=0}^{n} a_k x^k$ και υπάρχει δείκτης k τέτοιος ώστε $a_k = 0$. Αν θεωρήσουμε το ανάπτυγμα $Taylor$ του $P_n(x)$ με κέντρο κάποιο $x_0 \in \mathbb{R}$ τότε ισχύει

$$P_n(x) = \sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n} c_k (x - x_0)^k = \tilde{P}_n(\underbrace{x - x_0}_{y}) = \tilde{P}_n(y)$$

όπου $c_k = \frac{1}{k!} P_n^{(k)}(x_0) = \Upsilon_{n-k}(x_0)$ (τα υπόλοιπα των διαδοχικών διαιρέσεων του $P_n(x)$ δια $x - x_0$, π.χ. $c_n = \Upsilon_0(x_0)$).

Επομένως όταν υπάρχουν μηδενικοί συντελεστές η μέθοδος QD μπορεί να εφαρμοστεί στο πολυώνυμο $\tilde{P}_n(y)$, με συντελεστές $c_k \neq 0$.

1.6 Εύρεση ριζών πραγματικών πολυωνύμων - Μέθοδος Bairstow

Η ευστάθεια των ηλεκτρικών ή μηχανικών συστημάτων σχετίζεται με το πραγματικό μέρος των μιγαδικών ριζών ορισμένων πολυωνύμων βαθμού τουλάχιστον 30 με πραγματικούς συντελεστές. Έχει αποδειχθεί ότι δεν υπάρχουν γενικοί τύποι για τον υπολογισμό των ριζών πολυωνύμων βαθμού μεγαλύτερου από 4. Έτσι δεν υπάρχει άλλη επιλογή παρά μόνο η χρήση μιας αριθμητικής μεθόδου για την εύρεση των ριζών.

Αν μετατρέψουμε τις πραγματικές μεταβλητές σε μιγαδικές (τροποποιώντας κατάλληλα τις παραγράφους εισόδου/εξόδου) σε ένα πρόγραμμα FORTRAN της μεθόδου Newton-Raphson, μπορούμε να βρούμε προσεγγιστικές τιμές των ριζών. Όμως αυτή η εργασία είναι αρκετά επίπονη αν δεν είναι αυτόματα διαθέσιμη μια αριθμητική μιγαδικών. Στη συνέχεια παρουσιάζουμε μια περισσότερο επιθυμητή διαδικασία.

Η βασική ιδέα είναι να βρούμε τους δευτεροβάθμιους παράγοντες που αντιστοιχούν στις μιγαδικές ρίζες και έπειτα να βρούμε τις ρίζες αυτών των δευτεροβαθμίων πολυωνύμων με τους γνωστούς τύπους.

Γνωρίζουμε ότι αν μια πολυωνυμική εξίσωση

$$P(x) = a_1 x^n + a_2 x^{n-1} + \cdots + a_{n-1} x^2 + a_n x + a_{n+1} = 0$$

έχει μια μιγαδική ρίζα $\alpha + \beta i$ τότε θα έχει ρίζα και τη συζυγή της $\alpha - \beta i$, οπότε το πολυώνυμο:

$$R(x) = (x - \alpha - \beta i)(x - \alpha + \beta i) = x^2 - 2\alpha x + \alpha^2 + \beta^2$$

είναι παράγοντας του $P(x)$.

Επειδή το $R(x)$ έχει μόνο πραγματικούς συντελεστές, είναι καταλληλότερο να εργασθούμε με αυτό παρά με τις ρίζες του.

Η ταυτότητα της διαίρεσης ενός πολυωνύμου $P(x)$ με ένα δευτεροβάθμιο πολυώνυμο $R(x)$ είναι

$$P(x) = R(x) \cdot Q(x) + Y(x)$$

όπου $Y(x) = ax + b$ και $Q(x)$ ένα πολυώνυμο βαθμού $n-2$. Αν στην παραπάνω ταυτότητα διαίρεσης του $P(x)$ με το δευτεροβάθμιο παράγοντα $R(x)$ συμβολίσουμε:

$$\begin{aligned} R(x) &= x^2 - rx - s \\ Y(x) &= b_n(x - r) + b_{n+1} \\ Q(x) &= b_1 x^{n-2} + b_2 x^{n-3} + \cdots + b_{n-2}x + b_{n-1} \end{aligned}$$

τότε η ταυτότητα γράφεται:

$$\underbrace{a_1 x^n + \cdots + a_n x + a_{n+1}}_{P(x)} = \underbrace{(x^2 - rx - s)}_{R(x)} \cdot \underbrace{(b_1 x^{n-2} + \cdots + b_{n-2}x + b_{n-1})}_{Q(x)} + \underbrace{b_n(x - r) + b_{n+1}}_{Y(x)}$$

Εξισώνοντας τους συντελεστές των ομοβάθμιων όρων προκύπτουν οι ισότητες:

$$\begin{array}{rcl} a_1 = b_1 & & b_1 = a_1 \\ a_2 = b_2 - rb_1 & & b_2 = a_2 + rb_1 \\ a_3 = b_3 - rb_2 - sb_1 & & b_3 = a_3 + rb_2 + sb_1 \\ a_4 = b_4 - rb_3 - sb_2 & \Longleftrightarrow & b_4 = a_4 + rb_3 + sb_2 \\ \vdots & & \vdots \\ a_n = b_n - rb_{n-1} - sb_{n-2} & & b_n = a_n + rb_{n-1} + sb_{n-2} \\ a_{n+1} = b_{n+1} - rb_n - sb_{n-1} & & b_{n+1} = a_{n+1} + rb_n + sb_{n-1} \end{array}$$

Σχηματικά έχουμε

	a_1	a_2	a_3	a_4	$\cdots$	a_{n-1}	a_n	a_{n+1}
$\times s$	0	0	sb_1	sb_2	$\cdots$	sb_{n-3}	sb_{n-2}	sb_{n-1}
$\times r$	0	rb_1	rb_2	rb_3	$\cdots$	rb_{n-2}	rb_{n-1}	rb_n
	b_1	b_2	b_3	b_4	$\cdots$	b_{n-1}	b_n	b_{n+1}

Παράδειγμα 1.1. Έστω $P(x) = x^5 - 2x^4 + 7x^3 - 4x^2 + 11x - 2, \quad R(x) = x^2 - 2x + 3.$ Είναι $r = 2,\ s = -3.$

	1	-2	7	-4	11	-2
$\times(-3)$	0	0	-3	0	-12	-12
$\times 2$	0	2	0	8	8	14
	1	0	4	4	7	0
	b_1	b_2	b_3	b_4	b_5	b_6

'ρα $Q(x) = x^3 + 4x + 4, \quad Y(x) = 7(x-2) + 0 = 7x - 14.$

Η μέθοδος του Bairstow βρίσκει δευτεροβάθμιους παράγοντες $R(x)$ του πολυωνύμου $P(x)$. Για να είναι το $Y(x) = b_n(x-r) + b_{n+1} = 0$, οπότε το $R(x) = x^2 - rx - s$ είναι δευτεροβάθμιος παράγοντας του $P(x)$ πρέπει να βρεθούν οι συντελεστές r και s έτσι ώστε: $b_n = 0$ και $b_{n+1} = 0$. Έστω $\overline{r}$ και $\overline{s}$ οι τιμές που μηδενίζουν τα $b_n,\ b_{n+1}$ και ας υποθέσουμε ότι $r,\ s$ είναι προσεγγίσεις των $\overline{r},\ \overline{s}$.

Αν οι διαφορές $dr = \overline{r} - r$ και $ds = \overline{s} - s$ είναι «μικρές» τότε μπορούμε να χρησιμοποιήσουμε τα ολικά διαφορικά των $b_n,\ b_{n+1}$ για να πάρουμε τις προσεγγίσεις:

$$0 = b_n(\overline{r}, \overline{s}) = b_n(r + dr, s + ds) \cong b_n(r, s) + \frac{\partial b_n}{\partial r} dr + \frac{\partial b_n}{\partial s} ds$$

$$0 = b_{n+1}(\overline{r}, \overline{s}) = b_{n+1}(r + dr, s + ds) \cong b_{n+1}(r, s) + \frac{\partial b_{n+1}}{\partial r} dr + \frac{\partial b_{n+1}}{\partial s} ds$$

όπου οι μερικές παράγωγοι υπολογίζονται στο (r, s).

Επομένως αν r_k, s_k είναι οι τιμές των r και s στην k επανάληψη και συμβολίσουμε με $dr_k,\ ds_k$ τη λύση του παραπάνω συστήματος (αν αντικαταστήσουμε τα $\cong$ με $=$), τότε ορίζονται οι επόμε-

νες προσεγγιστικές τιμές

$$\begin{aligned} r_{k+1} &= r_k + dr_k \\ s_{k+1} &= s_k + ds_k \end{aligned}$$

οι οποίες πρέπει να είναι καλύτερες προσεγγίσεις των $\overline{r},\ \overline{s}$. Για να χρησιμοποιήσουμε αυτή τη μέθοδο πρέπει να γνωρίζουμε τις τέσσερις τιμές των μερικών παραγώγων στο σημείο (r, s).

Αν εφαρμόσουμε τη συνθετική διαίρεση του $P(x)$ δια $R(x)$, αντικαθιστώντας τα a_i με τα b_i τότε προκύπτουν οι νέοι συντελεστές $c_1, c_2, \ldots, c_n$ ως εξής:

$$\begin{array}{cccccc} & b_1 & b_2 & b_3 & \ldots & b_n \\ \times r & 0 & 0 & c_1 s & \ldots & c_{n-2}s \\ \times s & 0 & c_1 r & c_2 r & \ldots & c_{n-1}r \\ \hline & c_1 & c_2 & c_3 & \ldots & c_n \end{array}$$

Αποδεικνύεται (με επαγωγή) ότι οι μερικές παράγωγοι μπορούν να υπολογιστούν από τους τύπους:

$$\frac{\partial b_n}{\partial r} = c_{n-1} \quad , \quad \frac{\partial b_n}{\partial s} = c_{n-2}$$

$$\frac{\partial b_{n+1}}{\partial r} = c_n \quad , \quad \frac{\partial b_{n+1}}{\partial s} = c_{n-1}$$

Τελικά τα dr_k και ds_k προκύπτουν από τη λύση του γραμμικού συστήματος:

$$\begin{aligned} c_{n-1}dr + c_{n-2}ds &= -b_n \\ c_n dr + c_{n-1}ds &= -b_{n+1} \end{aligned}$$

η οποία είναι

$$dr_k = \frac{b_n c_{n-1} - b_{n+1}c_{n-2}}{c_n c_{n-2} - c_{n-1}^2} \quad \text{και} \quad ds_k = \frac{b_{n+1}c_{n-1} - b_n c_n}{c_n c_{n-2} - c_{n-1}^2}$$

Οι αρχικές τιμές $r_0,\ s_0$ λαμβάνονται συνήθως ίσες με 0. Καλύτερες προσεγγίσεις επιτυγχάνονται επιλέγοντας

- Για μεγάλες κατά μέτρο ρίζες (και εφόσον $a_1 \neq 0$), $r_0 = -\dfrac{a_2}{a_1}$ και $s_0 = -\dfrac{a_3}{a_1}$.
- Για μικρές κατά μέτρο ρίζες (και εφόσον $a_{n-1} \neq 0$), $r_0 = -\dfrac{a_n}{a_{n-1}}$ και $s_0 = -\dfrac{a_{n+1}}{a_{n-1}}$.

Αυτό γιατί για πολύ μεγάλες ρίζες $\overline{x}$ θα ισχύει:

$$0 = P(\overline{x}) \cong a_1\overline{x}^n + a_2\overline{x}^{n-1} + \overline{x}^{n-2} \quad \overset{\overline{x}\neq 0}{\Longrightarrow} \quad \overline{x}^2 + \frac{a_2}{a_1}\overline{x} + \frac{a_3}{a_1} \cong 0$$

και παρόμοια για πολύ μικρές ρίζες θα είναι:

$$a_{n-1}\overline{x}^2 + a_n\overline{x} + a_{n+1} \cong 0 \quad \overset{\overline{x}\neq 0}{\Longrightarrow} \quad \overline{x}^2 + \frac{a_n}{a_{n-1}}\overline{x} + \frac{a_{n+1}}{a_{n-1}} \cong 0$$

Αλγόριθμος 1.4. *Μέθοδος Bairstow (εύρεση δευτεροβάθμιου παράγοντα $R(x) = x^2 - rx - s$ του $n-$βαθμού πολυωνύμου $P(x) = a_1x^n + a_2x^{n-1} + \cdots + a_nx + a_{n+1},\ a_1 \neq 0$)*

1. *Διάβασε* $n,\ \ a_i,\ i = 1(1)n+1,\ \ M,\ NS,\ r_0,\ s_0$

2. $b_1 = a_1,\ c_1 = b_1,\ r = r_0,\ s = s_0,\ \varepsilon = 10^{-NS},\ k = 1$

3. *Όσο* $k \leq M$ *επανάλαβε*

   3.1. $b_2 = a_2 + b_1r,\ \ c_2 = b_2 + c_1r$

   3.2. *Για* $i = 3(1)n+1$
   $b_i = a_i + rb_{i-1} + sb_{i-2}$
   $c_i = b_i + rc_{i-1} + sc_{i-2}$

   3.3. $D = c_nc_{n-2} - c_{n-1}^2$
   $dr = (b_nc_{n-1} - b_{n+1}c_{n-2})/D$
   $ds = (b_{n+1}c_{n-1} - b_nc_n)/D$
   $r = r + dr,\ \ s = s + ds$

   3.4. *Αν* $|dr| \leq \varepsilon \cdot \max\{1, |r|\}$ *και* $|ds| \leq \varepsilon \cdot \max\{1, |s|\}$ *τότε*

      3.4.1. *Τύπωσε*($R(x) = x^2 - rx - s$ *είναι δευτεροβάθμιος παράγοντας του* $P(x)$ *και* $Q(x) = b_1x^{n-1} + \cdots + b_{n-2}x + b_{n-1}$ *είναι το πηλίκο της διαίρεσης* $P(x)/R(x)$)

      3.4.2. *Τέλος.*

   3.5. $k = k + 1$

4. *Τύπωσε("Όχι σύγκλιση μετά από* M *επαναλήψεις"), Τέλος.*

Αλγόριθμος 1.5. *Τετραγωνικός υποβιβασμός(Quadratic Deflation για τον υπολογισμό όλων των ριζών ενός πραγματικού πολυωνύμου* $P(x) = a_1x^n + a_2x^{n-1} + \cdots + a_nx + a_{n+1}, \quad a_1 \neq 0$*)*

1. *Διάβασε* $n, \quad a_i, \ i = 1(1)n+1$
2. *Όσο* $n \geq 2$ *επανάλαβε*
   - 2.1. *Εύρεση ενός πραγματικού δευτεροβάθμιου παράγοντα* $R(x)$ *ώστε* $P(x) = Q(x)R(x)$ *(μέθοδος Bairstow)*
   - 2.2. *Υπολογισμός των ριζών* $\rho_1, \ \rho_2$ *του* $R(x)$ *(τύποι δευτεροβάθμιας εξίσωσης)*
   - 2.3. *Τύπωσε*$(\rho_1, \ \rho_2)$
   - 2.4. $P(x) \leftarrow Q(x), \quad n \leftarrow n-2$
3. *Αν* $n = 1$ *τότε Τύπωσε("τελευταία ρίζα:",* $\rho = -\dfrac{a_{n-1}}{a_n}$*)*

Παράδειγμα 1.2. Έστω η εξίσωση $ax^2 + bc + c = 0$. Αν $b^2 - 4ac > 0$ τότε οι ρίζες της δίνονται από τον τύπο

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Έστω $b > 0$ και έστω ότι θέλουμε να υπολογίσουμε τη μικρότερη κατ΄ απόλυτο τιμή ρίζα εφαρμόζοντας τον τύπο

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

Αν το $4ac$ είναι μικρό συγκριτικά με το b^2 τότε η ποσότητα $\sqrt{b^2 - 4ac}$ προσεγγίζεται από το b με ακρίβεια ορισμένων δεκαδικών ψηφίων. Επομένως, δοθέντος ότι υπολογίζουμε με αυτήν την προσέγγιση το ριζικό, συμπεραίνουμε ότι ο αριθμητής του ανωτέρω τύπου και κατά συνέπεια η υπολογιζόμενη ρίζα θα είναι ακριβής σε ορισμένες θέσεις. Για να το αντιληφθούμε αυτό, δίνουμε το εξής παράδειγμα:

$$x^2 + 111.11x + 1.2121 = 0$$

Αν χρησιμοποιήσουμε αριθμητική κινητής υποδιαστολής με 5 σημαντικά ψηφία και στρογγύλευση τότε έχουμε:

$$\begin{aligned} b^2 &= 12345 \\ \sqrt{b^2 - 4ac} &= \sqrt{12345 - 4 \cdot 1 \cdot (1.2121)} = \sqrt{12340} = 111.09 \\ x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = -\frac{0.02}{2.1} = -0.01 \end{aligned}$$

ενώ η ακριβής τιμή της ρίζας (με 5 σημαντικά ψηφία) είναι $x_1 = -0.01091$.

Η απώλεια σημαντικών ψηφίων είναι δυνατό να αποφευχθεί αν χρησιμοποιήσουμε για τον υπολογισμό της μικρότερης κατ΄ απόλυτο τιμή ρίζας τον τύπο

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}} = \frac{-2 \cdot (1.2121)}{111.11 + 111.09} = -\frac{2.4242}{222.2} = -0.01091$$

Η μετάδοση σφάλματος μελετάται κατάλληλα με τη βοήθεια των εννοιών *συνθήκη*(condition) και *αστάθεια*(instability). Η λέξη *συνθήκη* χρησιμοποιείται για να περιγράψει την ευαισθησία της τιμής $f(x)$ μιας συνάρτησης σε μεταβολές της μεταβλητής x. Η συνθήκη συνήθως μετρείται ως εξής:

$$\mathrm{cond}(f) = \max\left\{ \left| \frac{f(x) - f(x^*)}{x - x^*} \right| \Big/ \left| \frac{x - x^*}{x} \right| \quad : \quad |x - x^*| \ll 1 \right\} \quad \cong \quad \left| \frac{x f'(x)}{x - x^*} \right|$$

Στην πράξη χρησιμοποιούμε την προσέγγιση

$$f(x) - f(x^*) \cong f'(x) \cdot (x - x^*)$$

Η έννοια της αστάθειας περιγράφει την ευαισθησία μιας αριθμητικής διαδικασίας για τον υπολογισμό του $f(x)$ σε αναπόφευκτα σφάλματα στρογγύλευσης που μεταφέρονται(διαδίδονται) κατά τη διάρκεια της εκτέλεσης αυτής σε αριθμητική πεπερασμένης ακρίβειας.

1.7 Μιγαδικές ρίζες και μέθοδος Müller

Οι μέθοδοι που εξετάσαμε μέχρι τώρα βρίσκουν μια απομονωμένη ρίζα μιας συνάρτησης, αν είναι γνωστή μια προσέγγιση της ρίζας. Οι μέθοδοι αυτές δεν είναι αρκετά ικανοποιητικές όταν απαιτείται ο υπολογισμός όλων των ριζών μιας συνάρτησης ή όταν δεν διατίθενται καλές αρχικές προσεγγίσεις. Όπως είδαμε σε προηγούμενες παραγράφους για τις πολυωνυμικές συναρτήσεις υπάρχουν μέθοδοι που δίνουν συγχρόνως προσεγγίσεις όλων των ριζών (π.χ. μέθοδος QD). Στη συνέχεια μπορούν να εφαρμοσθούν οι γνωστές επαναληπτικές μέθοδοι (π.χ. Newton-Raphson, ΤέμνουσαςSecant) για να πετύχουμε πιο ακριβείς προσεγγίσεις των ριζών.

Μια ενδιαφέρουσα μέθοδος, έχει προταθεί από τον Müller και έχει εφαρμοσθεί με αξιοσημείωτη επιτυχία. Η μέθοδος αυτή χρησιμοποιείται για την εύρεση ορισμένου αριθμού ριζών, πραγματικών ή μιγαδικών μιας οποιασδήποτε συνάρτησης. Η μέθοδος είναι επαναληπτική, συγκλίνει περίπου τετραγωνικά στην περιοχή μιας ρίζας, δεν απαιτεί τον υπολογισμό της παραγώγου της συνάρτησης και επιτυγχάνει την εύρεση και των μιγαδικών ριζών, ακόμα και στην περίπτωση που οι ρίζες δεν είναι απλές (δηλαδή έχουν πολλαπλότητα $k > 1$).

Επίσης η μέθοδος είναι γενική υπό την έννοια ότι ο χρήστης δε χρειάζεται αρχική προσέγγιση. Στην παράγραφο αυτή περιγράφουμε σύντομα τη μέθοδο, παραλείποντας τη μελέτη σύγκλισής της και συζητούμε τη χρήση της για την εύρεση των πραγματικών και μιγαδικών ριζών. Θα εξετάσουμε ειδικά το πρόβλημα της εύρεσης μιγαδικών ριζών πολυωνύμων με πραγματικούς συντελεστές, καθώς αυτό το πρόβλημα έχει μεγάλο ενδιαφέρον σε πολλούς κλάδους της Μηχανικής.

Η μέθοδος Müller είναι μια φυσική επέκταση της μεθόδου της Τέμνουσας. Υπενθυμίζουμε ότι στη μέθοδο της Τέμνουσας προσδιορίζουμε από τις προσεγγίσεις $x_i,\ x_{i+1}$ μιας ρίζας της $f(x)=0$, τη νέα προσέγγιση ως τη ρίζα του πρωτοβάθμιου πολυωνύμου $p(x)$, το οποίο διέρχεται από τα δυο σημεία $(x_i, f(x_i))$ και $(x_{i+1}, f(x_{i+1}))$.

Στη μέθοδο Müller, η επόμενη προσέγγιση x_{i+1} βρίσκεται ως η ρίζα της παραβολής που διέρχεται από τα τρία σημεία

$$\Big(x_i, f(x_i)\Big) \quad , \quad \Big(x_{i-1}, f(x_{i-1})\Big) \quad , \quad \Big(x_{i-2}, f(x_{i-2})\Big)$$

Σύμφωνα με τον τύπο παρεμβολής του Newton η παραβολή

$$p(x) = f(x_i) + f[x_i, x_{i-1}](x - x_i) + f[x_i, x_{i-1}, x_{i-2}](x - x_i)(x - x_{i-1})$$

είναι η μοναδική που προσεγγίζει τη συνάρτηση $f(x)$ στα τρία σημεία $x_i,\ x_{i-1},\ x_{i-2}$.

Επειδή είναι

$$(x - x_i)(x - x_{i-1}) = (x - x_i)^2 + (x - x_i)(x_i - x_{i-1})$$

το $p(x)$ γράφεται ως

$$\begin{aligned} p(x) &= f(x_i) + f[x_i, x_{i-1}](x - x_i) + f[x_i, x_{i-1}, x_{i-2}]\Big((x - x_i)^2 + (x - x_i)(x_i - x_{i-1})\Big) \\ &= f(x_i) + \Big(\underbrace{f[x_i, x_{i-1}] + f[x_i, x_{i-1}, x_{i-2}](x_i - x_{i-1})}_{c_i}\Big)(x - x_i) + f[x_i, x_{i-1}, x_{i-2}](x - x_i)^2 \\ &= f(x_i) + (x - x_i)c_i + f[x_i, x_{i-1}, x_{i-2}](x - x_i)^2 \end{aligned}$$

'ρα μια ρίζα ρ της παραβολής $p(x)$ ικανοποιεί τη σχέση

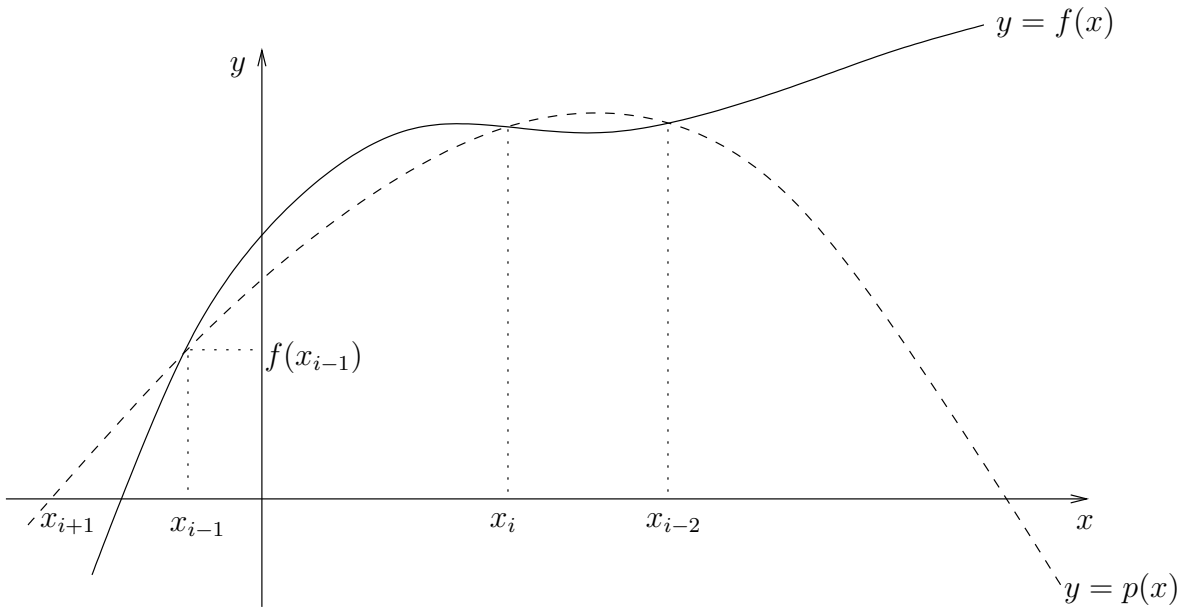
$$\rho - x_i = \frac{-2f(x_i)}{c_i \pm \Big(c_i^2 - 4f(x_i)f[x_i, x_{i-1}, x_{i-2}]\Big)^{1/2}} \tag{1.4}$$

σύμφωνα με το γνωστό τύπο για την εύρεση της μικρότερης κατά μέτρο ρίζας.

Αν επιλέξουμε το πρόσημο στην (1.4) έτσι ώστε ο παρονομαστής να είναι όσο το δυνατό μεγαλύτερος και αν ονομάσουμε h_{i+1} το δεξί μέλος της (1.4), τότε η επόμενη προσέγγιση της f είναι

$$x_{i+1} = x_i + h_{i+1}$$

Η διαδικασία επαναλαμβάνεται χρησιμοποιώντας τις τρεις βασικές προσεγγίσεις $x_{i-1},\ x_i,\ x_{i+1}$. Αν οι ρίζες που προκύπτουν από την (1.4) είναι πραγματικές, η κατάσταση φαίνεται γραφικά στο παρακάτω σχήμα:



Τονίζουμε ότι ακόμα και αν οι ρίζες είναι πραγματικές, μπορεί να προκύψουν μιγαδικές προσεγγίσεις, λόγω του ότι οι λύσεις που δίνει η (1.4) είναι μιγαδικές. Οπωσδήποτε, στις περιπτώσεις αυτές η φανταστική συντεταγμένη θα είναι τόσο μικρή ώστε να μπορεί να θεωρηθεί αμελητέα. Στην πράξη, στον αλγόριθμο που δίνεται παρακάτω κάποιες μιγαδικές συντεταγμένες που συναντώνται κατά την αναζήτηση μιας πραγματικής ρίζας παραλείπονται.

Δίνουμε τώρα τον αλγόριθμο της μεθόδου του Müller.

Αλγόριθμος 1.6. *Μέθοδος Müller*

1. *Διάβασε*

   $x_0,\ x_1,\ x_2$ *(αρχικές προσεγγίσεις της ρίζας ξ της f)*

   $\varepsilon_1,\ \varepsilon_2$ *(ανεκτικότητες)*

   M *(μέγιστος επιτρεπτός αριθμός επαναλήψεων)*

2. $h_1 = x_1 - x_0,\ h_2 = x_2 - x_1,$
   $f[x_1, x_0] = \frac{f(x_2)-f(x_1)}{h_1}, \quad f[x_2, x_1] = \frac{f(x_2)-f(x_1)}{h_2}, \quad i = 2$

3. *Επανάλαβε*

   3.1. $f[x_i, x_{i-1}, x_{i-2}] = \dfrac{f[x_i, x_{i-1}] - f[x_{i-1}, x_{i-2}]}{h_i + h_{i-1}}$

   3.2. $c_i = f[x_i, x_{i-1}] + h_i f[x_i, x_{i-1}, x_{i-2}]$

   3.3. $h_{i+1} = \dfrac{-2f(x_i)}{c_i \pm \sqrt{c_i^2 - 4f(x_i)f[x_i, x_{i-1}, x_{i-2}]}}$
   (επιλογή του προσήμου έτσι ώστε ο παρονομαστής να είναι κατ,απόλυτη τιμη μέγιστος)

   3.4. $x_{i+1} = x_i + h_{i+1}, \quad f[x_{i+1}, x_i] = \dfrac{f(x_{i+1}) - f(x_i)}{h_{i+1}}$

   3.5. $i = i + 1$

   Έως ότου $|x_i - x_{i-1}| < \varepsilon_1 \cdot |x_i| \quad$ *ή* $\quad |f(x_i)| < \varepsilon_2 \quad$ *ή* $\quad i > M$

4. *Αν* $i \leq M$ *Τύπωσε("Προσεγγιστική τιμή της ρίζας:",* x_i*)*
   αλλιώς Τύπωσε(" Όχι σύγκλιση μετά από M *επαναλήψεις")*

Κεφάλαιο 2

Μη Γραμμικά Συστήματα

Έστω $f_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$, n το πλήθος μιγαδικές συναρτήσεις n μεταβλητών $x_1, x_2, \ldots, x_n$ ορισμένες σε μια περιοχή $B \subseteq \mathbb{R}^n$. Θεωρούμε το σύστημα

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned}$$

ή υπό μορφή διανυσμάτων

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \tag{2.1}$$

όπου

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \quad , \quad \mathbf{f} = (f_1, f_2, \ldots, f_n)^T$$

Ονομάζουμε ρίζα (ή μηδενικό σημείο) του συστήματος κάθε σημείο $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)^T \in B$ τέτοιο ώστε $\mathbf{f}(\boldsymbol{\xi}) = \mathbf{0}$.

Μια ειδική περίπτωση συστημάτων της μορφής (2.1) είναι τα γραμμικά συστήματα. Για τα μη γραμμικά συστήματα η εύρεση κριτηρίων που εξασφαλίζουν την ύπαρξη και το μονοσήμαντο των λύσεων είναι ένα εξαιρετικά δύσκολο πρόβλημα και αντιμετωπίζεται αποτελεσματικά μόνο σε ειδικές περιπτώσεις. Αλλά και αν ακόμα εξασφαλισθεί η ύπαρξη λύσης, η εύρεσή της είναι αρκετά δύσκολο πρόβλημα.

2.1 Μέθοδος απαλοιφής

Η πλέον γνωστή μέθοδος επίλυσης συστημάτων της μορφής (2.1) είναι η μέθοδος της απαλοιφής. Κατά τη μέθοδο αυτή απαλείφονται οι άγνωστοι μεταξύ των εξισώσεων και το αρχικό σύστημα μετατρέπεται σε ένα πεπερασμένο σύνολο συστημάτων της μορφής

$$\begin{aligned} \varphi_1(x_1, x_2, \ldots, x_n) &= 0 \\ \varphi_2(x_2, x_3, \ldots, x_n) &= 0 \\ &\vdots \\ \varphi_n(x_n) &= 0 \end{aligned}$$

και στη συνέχεια λύνοντας αυτά τα (απλούστερα) συστήματα κατά τα γνωστά. Δηλαδή βρίσκουμε πρώτα τις ρίζες της τελευταίας εξίσωσης $\phi_n(x_n) = 0$ με μια από τις γνωστές μεθόδουω. Με αντικατάσταση των ριζών αυτών στις προηγούμενες εξισώσεις, προκύπτουν επίσης συστήματα τις ίδιας μορφής, αλλά με $n-1$ εξισώσεις/αγνώστους. Εργαζόμενοι με τον τρόπο αυτό βρίσκουμε τελικώς όλες τις ρίζες του συστήματος.

Παράδειγμα 2.1. Έστω το σύστημα

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

Με απαλοιφή του x_1 από τις εξισώσεις προκύπτει το ισοδύναμο σύστημα

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

Από τη δεύτερη εξίσωση προκύπτουν $x_2 = \frac{1}{2}$ ή $x_2 = -1$. Αν αντικαταστήσουμε στην πρώτη προκύπτουν τελικά οι παρακάτω λύσεις του αρχικού συστήματος:

$$\begin{aligned} x_1 = 0 \quad &, \quad x_2 = -1 \\ x_1 = \frac{\sqrt{3}}{2} \quad &, \quad x_2 = \frac{1}{2} \\ x_1 = -\frac{\sqrt{3}}{2} \quad &, \quad x_2 = \frac{1}{2} \end{aligned}$$

Η μέθοδος της απαλοιφής παρουσιάζει γενικά πολλές δυσκολίες ώστε να χρησιμοποιείται μόνο σε ειδικές περιπτώσεις και για συστήματα με μικρό αριθμό αγνώστων.

2.2 Γραφική μέθοδος

Έστω το σύστημα

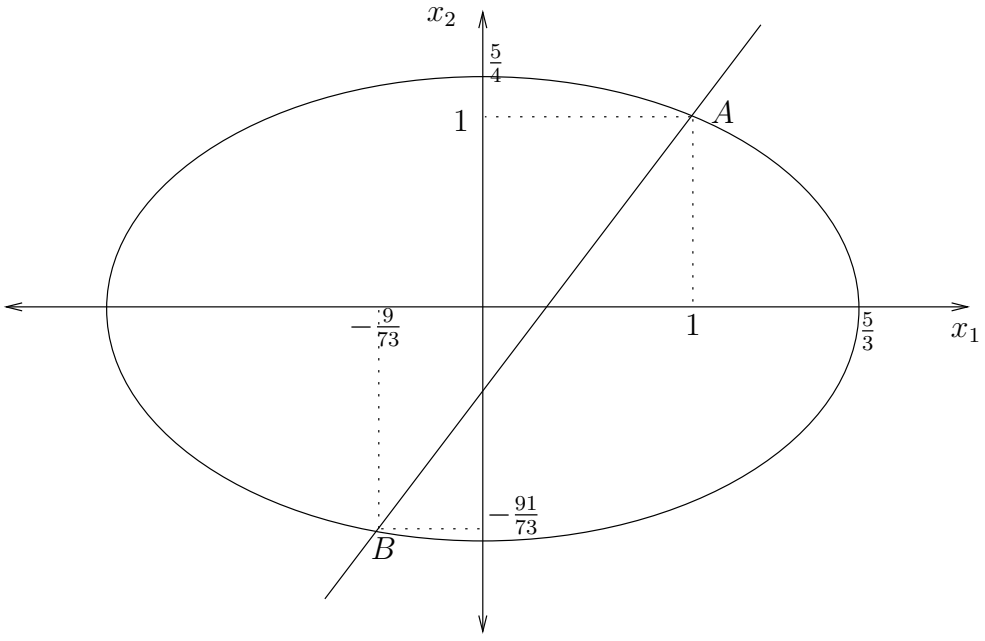
$$\begin{aligned} f_1(x_1, x_2) &= 0 \\ f_2(x_2, x_3) &= 0 \end{aligned}$$

Θεωρούμε ένα σύστημα ορθογωνίων συντεταγμένων $x_1 O x_2$ και σχεδιάζουμε τις γραφικές παραστάσεις των παραπάνω συναρτήσεων στο πεδίο ορισμού τους. Προφανώς οι συντεταγμένες των σημείων τιμής των δυο γραμμών ορίζουν τις πραγματικές λύσεις του συστήματος. Η μέθοδος αυτή είναι εφικτή εφόσον είναι δυνατό να γίνουν οι γραφικές αυτές παραστάσεις.

Παράδειγμα 2.2. Να λυθεί γραφικά το σύστημα

$$\begin{aligned} 9x_1^2 + 16x_2^2 - 25 &= 0 \\ 2x_1 - x_2 - 1 &= 0 \end{aligned}$$

Η πρώτη από τις εξισώσεις παριστά έλλειψη με κέντρο την αρχή των αξόνων και ημιάξονες $\frac{5}{3}$ και $\frac{5}{4}$. Η δεύτερη παριστά μια ευθεία.



Από το σχήμα φαίνεται ότι οι λύσεις του συστήματος είναι οι συντεταγμένες των σημείων τομής $A(1,1),\ B(-\frac{9}{73}, -\frac{91}{73})$.

2.3 Επαναληπτικές μέθοδοι

Υποθέτουμε ότι το σύστημα (2.1) μπορεί να μετασχηματιστεί με κατάλληλους μετασχηματισμούς στην ισοδύναμη μορφή

$$\mathbf{x} = \boldsymbol{\varphi}(\mathbf{x})$$

όπου $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \quad , \quad \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$.

Ορίζουμε τις παρακάτω επαναληπτικές μεθόδους:

Επαναληπτική μέθοδος ολικού βήματος: $\mathbf{x}^{(m+1)} = \boldsymbol{\varphi}(\mathbf{x}^{(m)}) \quad , \quad m = 0, 1, 2, \ldots$ όπου το $\mathbf{x}^{(0)}$ είναι μια πρώτη προσέγγιση του ζητούμενου σταθερού σημείου της $\boldsymbol{\varphi}(\mathbf{x})$.

Επαναληπτική μέθοδος απλού βήματος:
$\mathbf{x}_i^{(m+1)} = \boldsymbol{\varphi}_i(\mathbf{x}_1^{(m+1)}, \mathbf{x}_2^{(m+1)}, \ldots, \mathbf{x}_{i-1}^{(m+1)}, \mathbf{x}_{i+1}^{(m)}, \ldots, \mathbf{x}_n^{(m)})$, $i = 1(1)n$, όπου $m = 0, 1, 2, \ldots$ και δίνεται μια αρχική προσέγγιση $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$.

Στη συνέχεια δίνουμε μια πρόταση με την οποία εξασφαλίζεται η ύπαρξη και το μονοσήμαντο ενός σταθερού σημείου της $\boldsymbol{\varphi}(\mathbf{x})$ καθώς και τον τρόπο υπολογισμού αυτού.

Πρόταση 2.1. *Θεωρούμε το κλειστό ορθογώνιο $n-$διαστάσεων*

$$\Delta_n = \{\ \alpha_i \leq x_i \leq \beta_i \quad , \quad i = 1, 2, \ldots, n\ \}$$

και τις n πραγματικές συναρτήσεις n μεταβλητών

$$\varphi_i = \varphi_i(x_1, x_2, \ldots, x_n) \quad , \quad i = 1, 2 \ldots, n$$

που ορίζονται στο Δ_n και ικανοποιούν τις παρακάτω προϋποθέσεις:

1) $\varphi_i\Big|_{\Delta_n}$ *συνεχείς,* $i = 1, 2 \ldots, n$

2) *Για κάθε* $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \Delta_n$ *ισχύει* $\boldsymbol{\varphi}(\mathbf{x}) \in \Delta_n$

3) *Υπάρχει μια σταθερά* $L < 1$ *τέτοια ώστε για κάθε* $\mathbf{x}_1, \mathbf{x}_2 \in \Delta_n$ *ισχύει η συνθήκη του Lipschitz* $\|\boldsymbol{\varphi}(\mathbf{x}_1) - \boldsymbol{\varphi}(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$

Τότε ισχύουν τα παρακάτω

a) *Υπάρχει ακριβώς ένα* $\boldsymbol{\xi} \in \Delta_n$ *τέτοιο ώστε* $\boldsymbol{\xi} = \boldsymbol{\varphi}(\boldsymbol{\xi})$

β) Για κάθε $\mathbf{x}_0 \in \Delta_n$ *η ακολουθία* $\mathbf{x}_m,\ m = 0, 1, 2, \ldots$ *που ορίζεται με την επαναληπτική μέθοδο* $\mathbf{x}_{m+1} = \boldsymbol{\varphi}(\mathbf{x}_m),\ m = 0, 1, 2, \ldots$ *έχει νόημα, δηλαδή* $\mathbf{x}_m \in \Delta_n,\ \forall\, m = 0, 1, 2, \ldots$ *και* $\lim\limits_{m\to\infty} \mathbf{x}_m = \boldsymbol{\xi}$ *και επιπλέον ισχύει η ανισότητα*

$$\|\mathbf{x}_m - \boldsymbol{\xi}\| \leq \frac{L^m}{1-L}\|\mathbf{x}_1 - \mathbf{x}_0\|$$

Στη συνέχεια για λόγους συντομίας και απλής παρουσίασης περιοριζόμαστε στο χώρο $\mathbb{R}^2$. Είναι αυτονόητο ότι όλα τα συμπεράσματα μπορούν να επεκταθούν στο χώρο $\mathbb{R}^n$ χωρίς μεγάλη δυσκολία. Έστω λοιπόν η διανυσματική συνάρτηση

$$\boldsymbol{\varphi}(\mathbf{x}) = \Big(\ \varphi_1(x_1, x_2)\ ,\ \varphi_2(x_1, x_2)\ \Big)^T$$

για την οποία υποθέτουμε ότι υπάρχουν οι μερικές παράγωγοι πρώτης τάξης

$$\frac{\partial \varphi_j}{\partial x_i}\ \ ,\ \ i = 1, 2\ ,\ \ j = 1, 2$$

σε μια περιοχή ενός σημείου $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$.

Ορίζουμε τώρα τον Ιακωβιανό πίνακα (Jacobian) των $\varphi_1,\ \varphi_2$ στο $\boldsymbol{\xi}$ ως εξής:

$$J(\boldsymbol{\varphi}(\boldsymbol{\xi})) = \begin{bmatrix} \dfrac{\partial \varphi_1(\xi_1, \xi_2)}{\partial x_1} & \dfrac{\partial \varphi_1(\xi_1, \xi_2)}{\partial x_2} \\ \\ \dfrac{\partial \varphi_2(\xi_1, \xi_2)}{\partial x_1} & \dfrac{\partial \varphi_2(\xi_1, \xi_2)}{\partial x_2} \end{bmatrix}$$

Δίνουμε τώρα μια άλλη πρόταση, η οποία είναι επέκταση της γνωστής πρότασης σταθερού σημείου, στο διδιάστατο χώρο.

Πρόταση 2.2. *Έστω το σύστημα* $\mathbf{x} = \boldsymbol{\varphi}(\mathbf{x})$*, το οποίο υποθέτουμε ότι έχει ένα σταθερό σημείο* $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$*. Υποθέτουμε ότι υπάρχει* $\delta > 0$ *ώστε για κάθε* $\mathbf{x} \in \mathbb{R}^2$ *με* $\|\mathbf{x} - \boldsymbol{\xi}\| < \delta$ *υπάρχουν οι μερικές παράγωγοι* $\dfrac{\partial \varphi_j}{\partial x_i}\ ,\ i, j = 1, 2$ *και η φασματική ακτίνα του Ιακωβιανού πίνακα των* $\varphi_1,\ \varphi_2$ *στο* $\mathbf{x}$ *είναι μικρότερη της μονάδας, δηλαδή* $S\Big(J(\boldsymbol{\varphi}(\boldsymbol{\xi}))\Big) < 1$*. Τότε, για κάθε* $\mathbf{x}_0 \in \mathbb{R}^2$ *με* $\|\mathbf{x}_0 - \boldsymbol{\xi}\| < \delta$ *η επαναληπτική μέθοδος* $\mathbf{x}_{m+1} = \boldsymbol{\varphi}(\mathbf{x}_m),\ m = 0, 1, 2, \ldots$ *έχει έννοια και* $\lim\limits_{m\to\infty} \mathbf{x}_m = \boldsymbol{\xi}$*.*

Επίσης η $\boldsymbol{\varphi}(\mathbf{x})$ *έχει το μοναδικό σταθερό σημείο* $\boldsymbol{\xi}$ *στην περιοχή* $\|\mathbf{x} - \boldsymbol{\xi}\| < \delta$*.*

Μια άλλη χρήσιμη πρόταση για τις εφαρμογές είναι η παρακάτω, η οποία προκύπτει εύκολα από την προηγούμενη.

Πρόταση 2.3. *Έστω το παραπάνω σύστημα* $\mathbf{x} = \boldsymbol{\varphi}(\mathbf{x})$ *και οι συναρτήσεις* $\varphi_1,\ \varphi_2$ *ορισμένες στο ορθογώνιο*

$$\Delta_2 = \{\ \alpha_i \le x_i \le \beta_i \quad ,\quad i = 1, 2\ \}$$

με συνεχείς μερικές παραγώγους πρώτης τάξης.

Έστω $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$ *σταθερό σημείο της* $\boldsymbol{\varphi}(\mathbf{x})$ *που βρίσκεται στο εσωτερικό του* Δ_2*, τέτοιο ώστε* $J(\boldsymbol{\varphi}(\boldsymbol{\xi})) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

Τότε υπάρχει $\delta > 0$ *ώστε η επαναληπτική μέθοδος* $\mathbf{x}_{m+1} = \boldsymbol{\varphi}(\mathbf{x}_m),\ m = 0, 1, 2, \ldots$ *έχει έννοια για κάθε* $\mathbf{x}_0 \in \Delta_2$ *με* $\|\mathbf{x}_0 - \boldsymbol{\xi}\| < \delta$ *και ισχύει* $\lim\limits_{m\to\infty} \mathbf{x}_m = \boldsymbol{\xi}$.

Παράδειγμα 2.3. Έστω το σύστημα

$$\begin{aligned} x_1 &= x_1^2 - x_2^2 \\ x_2 &= x_1^2 + x_2^2 \end{aligned}$$

δηλαδή $\boldsymbol{\varphi}(\mathbf{x}) = (x_1^2 - x_2^2,\ \ x_1^2 + x_2^2)$.

Προφανώς ισχύει $\boldsymbol{\varphi}(\mathbf{0}) = \mathbf{0}$, δηλαδή το $(0, 0)$ είναι σταθερό σημείο της $\boldsymbol{\varphi}(\mathbf{x})$. Είναι

$$J(\boldsymbol{\varphi}(\mathbf{x})) = \begin{bmatrix} 2x_1 & -2x_2 \\ 2x_1 & 2x_2 \end{bmatrix}$$

άρα $J(\boldsymbol{\varphi}(\mathbf{0})) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. Θεωρούμε την επαναληπτική μέθοδο

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

Αν πάρουμε ώς αρχικό διάνυσμα $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}) = \left(\frac{1}{2},\ \frac{1}{2}\right)$ τότε έχουμε

m	$x_1^{(m)}$	$x_2^{(m)}$
0	0.5	0.5
1	0	0.5
2	−0.25	0.25
3	0	0.125
4	−0.015625	0.015625
5	0	0.00048826125

2.4 Επαναληπτική μέθοδος Newton-Raphson(N-R)

Για λόγους απλότητας θεωρούμε $\mathbf{x} = (x, y) \in \mathbb{R}^2 \quad , \quad \mathbf{f} = (f, g)$. Έστω λοιπόν η διανυσματική συνάρτηση

$$\mathbf{f}(\mathbf{x}) = \Big(\ f(x,y) \ , \ g(x,y) \ \Big)^T$$

ορισμένη στο ορθογώνιο $\Delta_2 = \{ \ \alpha \leq x \leq \beta \ , \ \gamma \leq y \leq \delta \ \}$. Υποθέτουμε ότι οι $f, \ g$ έχουν συνεχείς μερικές παραγώγους μέχρι δεύτερης τάξης στο Δ_2. Αν το σύστημα

$$\begin{aligned} f(x,y) &= 0 \\ g(x,y) &= 0 \end{aligned}$$

έχει μια λύση $\boldsymbol{\xi} = (\xi, \eta)^T$ στο εσωτερικό του Δ_2 και επιπλέον ισχύει $|J(\mathbf{f}(\boldsymbol{\xi}))| \neq 0$ τότε είναι γνωστό από τον απειροστικό λογισμό ότι το σύστημα δεν έχει άλλη λύση σε μια αρκετά μικρή περιοχή του $\boldsymbol{\xi}$.

Με τις ανωτέρω προϋποθέσεις ορίζουμε την επαναληπτική μέθοδο N-R:

$$\mathbf{x}_{m+1} = \mathbf{x}_m - J^{-1}(\mathbf{f}(\mathbf{x}_m)) \cdot \mathbf{f}(\mathbf{x}_m) \quad , \quad m = 0, 1, 2, \ldots \tag{2.2}$$

όπου $J^{-1}(\mathbf{f}(\mathbf{x}_m))$ ο αντίστροφος του Ιακωβιανού πίνακα $J(\mathbf{f}(\mathbf{x}_m))$.

Αποδεικνύεται ότι υπάρχει $\delta > 0$ ώστε για κάθε $\mathbf{x}_0 \in \Delta_2$ με $\|\mathbf{x}_0 - \boldsymbol{\xi}\| < \delta$ η ανωτέρω επαναληπτική μέθοδος έχει έννοια και ορίζει την ακολουθία $\mathbf{x}_m$ έτσι ώστε $\lim\limits_{m\to\infty} \mathbf{x}_m = \boldsymbol{\xi}$.

Για την καλύτερη κατανόηση της μεθόδου τη γράφουμε αναλυτικά υπό μορφή συντεταγμένων. Έστω $\mathbf{x}_m = (x_m, y_m)^T$ και

$$J(\mathbf{f}(\mathbf{x}_m)) = \begin{bmatrix} \dfrac{\partial f(x_m, y_m)}{\partial x} & \dfrac{\partial f(x_m, y_m)}{\partial y} \\ \\ \dfrac{\partial g(x_m, y_m)}{\partial x} & \dfrac{\partial g(x_m, y_m)}{\partial y} \end{bmatrix}$$

Αν αντιστρέψουμε τον πίνακα αυτό και αντικαταστήσουμε στην (2.2) τότε προκύπτουν μετά τη

διάσπαση σε συντεταγμένες οι ακόλουθες εξισώσεις της ε.μ. N-R

$$x_{m+1} = x_m + \frac{g(x_m, y_m)\dfrac{\partial f(x_m, y_m)}{\partial y} - f(x_m, y_m)\dfrac{\partial g(x_m, y_m)}{\partial y}}{\dfrac{\partial f(x_m, y_m)}{\partial x} \cdot \dfrac{\partial g(x_m, y_m)}{\partial y} - \dfrac{\partial f(x_m, y_m)}{\partial y} \cdot \dfrac{\partial g(x_m, y_m)}{\partial x}}$$

$$y_{m+1} = y_m + \frac{f(x_m, y_m)\dfrac{\partial g(x_m, y_m)}{\partial x} - g(x_m, y_m)\dfrac{\partial f(x_m, y_m)}{\partial x}}{\dfrac{\partial f(x_m, y_m)}{\partial x} \cdot \dfrac{\partial g(x_m, y_m)}{\partial y} - \dfrac{\partial f(x_m, y_m)}{\partial y} \cdot \dfrac{\partial g(x_m, y_m)}{\partial x}}$$

Παράδειγμα 2.4. Εφαρμόστε την επαναληπτική μέθοδο Newton-Raphson για τη λύση του συστήματος:

$$\begin{aligned} f(x, y) &= x - x^2 - y^2 \\ g(x, y) &= y - x^2 + y^2 \end{aligned}$$

και με αρχική τιμή $\mathbf{x}_0 = (x_0, y_0) = (0.8, 0.4)$.

Αν αντικαταστήσουμε στις ανωτέρω εξισώσεις της μεθόδου N-R προκύπτουν οι ακόλουθες εξισώσεις συντεταγμένων

$$x_{m+1} = x_m + \frac{(y_m - x_m^2 + y_m^2) \cdot (-2y_m) - (x_m - x_m^2 - y_m^2) \cdot (1 + 2y_m)}{(1 - 2x_m) \cdot (1 + 2y_m) - (-2y_m) \cdot (-2x_m)}$$

$$y_{m+1} = y_m + \frac{(x_m - x_m^2 - y_m^2) \cdot (-2x_m) - (y_m - x_m^2 + y_m^2) \cdot (1 - 2x_m)}{(1 - 2x_m) \cdot (1 + 2y_m) - (-2y_m) \cdot (-2x_m)}$$

Αν εφαρμόσουμε τους παραπάνω τύπους συντεταγμένων προκύπτουν:

m	$x_1^{(m)}$	$x_2^{(m)}$
0	0.8	0.4
1	0.772881359	0.420338983
2	0.771845967	0.419644283
3	0.771844506	0.419643377
4	0.771844506	0.419643377

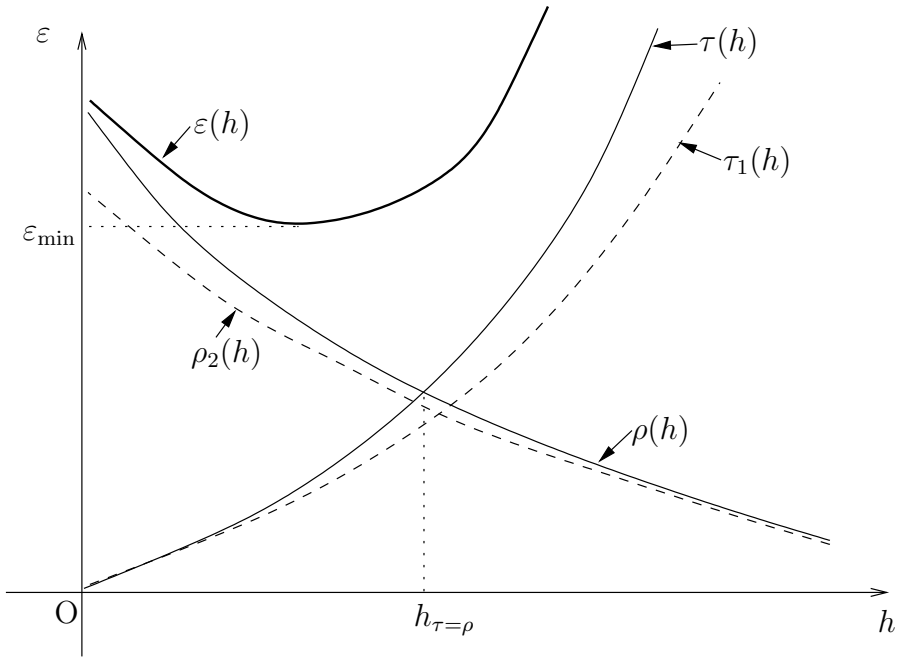
2.5 Το δίλημμα στην επιλογή μεγέθους βήματος h

Το πραγματικό σφάλμα $\varepsilon(h)$ σε μια αριθμητική λύση προέρχεται από δυο παράγοντες· το σφάλμα αποκοπής $\tau(h)$ και το σφάλμα στρογγύλευσης $\rho(h)$:

$$\varepsilon(h) = \tau(h) + \rho(h)$$

Το $\tau(h)$ εξαρτάται από τον τύπο της αριθμητικής μεθόδου που χρησιμοποιείται, και είναι $\tau(h) = O(h^n)$ για μια $n-$τάξης μέθοδο, δηλαδή $\tau(h) \to 0$ καθώς $h \to 0$. Το $\rho(h)$ εξαρτάται από τον τύπο καθώς και τον επεξεργαστή που χρησιμοποιείται, ενώ αυξάνει καθώς $h \to 0$. Επομένως θα υπάρχει κάποια τιμή $h_{\tau=\rho}$ τέτοια ώστε $\tau(h) = \rho(h)$ και για $h < h_{\tau=\rho}$ να είναι $\tau(h) < \rho(h)$).

Το πρόβλημα της εύρεσης ενός μεγέθους βήματος h αρκετά μικρού ώστε το $\tau(h)$ να είναι ¨μικρό' και το $\rho(h)$ να μην υπερέχει αυτού (δηλαδή $\tau(h) \leq \rho(h)$) αναφέρεται ως το *δίλημμα μεγέθους βήματος*. Σχηματικά:



Υπάρχουν δυο τρόποι ελάττωσης του $\varepsilon_{\min}$. Πρώτον η χρήση ενός τύπου υψηλότερης τάξης (αύξηση του n). Αυτό ελαττώνει το $\varepsilon_{\min}$ με ελάττωση της καμπύλης $\tau(h)$ σε $\tau_1(h)$, όπως φαίνεται στο γράφημα. Δεύτερον η χρήση αριθμητικής υψηλότερης ακρίβειας. Αυτό ελαττώνει την $\rho(h)$ προς τα κάτω και αριστερά ($\rho_2(h)$ στο γράφημα). Και οι δυο αυτοί τρόποι έχουν μια ατέλεια: απαιτούν να γίνει επανάληψη των υπολογισμών. Ο δεύτερος τρόπος μπορεί να χρειαστεί και διαφορετικό επεξεργαστή.

2.6 Βελτιωτικός τύπος του Richardson

Θεωρούμε το ανάπτυγμα της f σε πολυώνυμο Taylor $n-$βαθμού ως προς το x_0 (n άρτιος) και υπολογίζουμε της τιμές $f(x_0+h)$ και $f(x_0-h)$, οπότε έχουμε:

$$\begin{aligned} f(x_0+h) &= f(x_0)+hf'(x_0)+\frac{h^2}{2!}f''(x_0)+\cdots+\frac{h^n}{n!}f^{(n)}(x_0)+\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi_1) \\ f(x_0-h) &= f(x_0)-hf'(x_0)+\frac{h^2}{2!}f''(x_0)-\cdots+(-1)^n\frac{h^n}{n!}f^{(n)}(x_0)-\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi_{-1}) \end{aligned}$$

όπου $x_0-h<\xi_{-1}<x_0<\xi_1<x_0+h$.

Αφαιρώντας κατά μέλη και εφαρμόζοντας το Θεώρημα Ενδιάμεσης Τιμής προκύπτει

$$\frac{f(x_0+h)-f(x_0-h)}{2h}=f'(x_0)+\underbrace{\frac{h^2}{3!}f'''(x_0)+\frac{h^4}{5!}f^{(5)}(x_0)+\cdots+\frac{h^{n-2}}{(n-1)!}f^{(n-1)}(x_0)+\frac{h^n}{(n+1)!}f^{(n+1)}(\xi)}_{-\tau(h)}$$

όπου $\xi\in(x_0-h,x_0+h)$, ή

$$\underbrace{f'(x_0)}_{Q}=\underbrace{D_h(f(x_0))}_{F(h)}+\tau(h)$$

Υποθέτουμε ότι $F(h)$ είναι μια προσέγγιση τάξης $O(h^n)$ της προσεγγιζόμενης ποσότητας Q και παίρνουμε δυο προσεγγίσεις $F(h),\ F(h_{\max})$. Τότε δίνεται μια βελτιωμένη προσέγγιση της Q με τον τύπο:

$$F_1(h)=\frac{q^nF(h)-F(h_{\max})}{q^n-1},\qquad \text{όπου } h_{\max}=qh \tag{2.3}$$

Αν είναι γνωστό ότι $\tau(h)=ch^n+O(h^m)$ τότε ο $F_1(h)$ είναι τάξης $m>n$, δηλαδή:

$$Q-F_1(h)=Dh^m+\ (\text{όροι υψηλότερης τάξης})\ =O(h^m)$$

Στην περίπτωση αυτή μπορούμε να χρησιμοποιήσουμε τον τύπο (2.3) για να πάρουμε υψηλότερης τάξης προσεγγίσεις

$$F_2(h)=\frac{q^mF_1(h)-F_1(h_{\max})}{q^m-1},\qquad \text{τάξης } O(h^{m+2})$$

Παρατήρηση 2.1. Όταν εφαρμόζεται ο τύπος βελτίωσης σφάλματος του Richardson για ένα τύπο πεπερασμένων διαφορών (προς τα εμπρός ή προς τα πίσω) η τάξη σφάλματος αυξάνει κατά 1, ενώ για έναν τύπο κεντρικών διαφορών η τάξη σφάλματος αυξάνει κατά 2.

Εφαρμογή 2.1. Ας δούμε την τεχνική βελτίωσης σφάλματος του Richardson με ένα παράδειγμα. Γνωρίζουμε ότι

$$\underbrace{f'(x)}_{Q} = \underbrace{\frac{f(x+h)-f(x-h)}{2h}}_{F(h)=D_h(f(x))} + \underbrace{O(h^2)}_{\tau(h)} \tag{2.4}$$

Θεωρούμε τη συνάρτηση $f(x) = e^x$ και θέλουμε να προσεγγίσουμε την $f'(1) = e \cong 2.718282$. Αν χρησιμοποιήσουμε τον προσεγγιστικό τύπο των κεντρικών διαφορών με αριθμητική 7 σημαντικών ψηφίων προκύπτει ο ακόλουθος πίνακας:

h	$F(h) = \frac{e^{1+h}-e^{1-h}}{2h}$	$\varepsilon(h) = f'(x) - D_h(f(x))$	$\tau(h) = -\frac{f^{(3)}(\xi)}{6}h^2 \cong -\frac{1}{6}h^2$
0.2	$2.7\underline{36440}$	-0.018158	$-1.8 \cdot 10^{-2}$
0.02	$2.718\underline{475}$	-0.000193	$-1.8 \cdot 10^{-4}$
0.002	$2.71825\underline{0}$	-0.000032	$-1.8 \cdot 10^{-6}$

Μπορούμε να εκφράσουμε το βελτιωτικό τύπο του Richardson για $n = 2,\ m = 4$ παίρνοντας $q = 10$. Τότε έχουμε

$$F_1(h) = \frac{10^2 F(h) - F(10h)}{10^2 - 1} \qquad \text{και} \qquad F_2(h) = \frac{10^4 F_1(h) - F_1(10h)}{10^4 - 1} \tag{2.5}$$

Αν εφαρμόσουμε τους τύπους (2.5) για τις τιμές του $F(h)$ στα σημεία $h = 0.2,\ h = 0.02,\ h = 0.002$ στον πιο πάνω πίνακα τότε προκύπτει:

h	$F(h)$ $[O(h^2)]$		$F_1(h)$ $[O(h^4)]$		$F_2(h)$ $[O(h^6)]$
0.2	$2.7\underline{36440}$				
		$\searrow$			
0.02	$2.718\underline{475}$	$\longrightarrow$	$2.71829\underline{4}$		
		$\searrow$		$\searrow$	
0.002	$2.71825\underline{0}$	$\longrightarrow$	$2.71824\underline{8}$	$\longrightarrow$	$2.71824\underline{8}$

Παρατηρούμε ότι η $F_1(0.02)$ είναι πιο ακριβής από τις $F_1(0.002)$ και $F_2(0.002)$. Αυτό οφείλεται στο ότι η τιμή $F(0.002)$ που έχει το μικρότερο σφάλμα στρογγύλευσης βαρύνεται περισσότερο από την τιμή $F(0.02)$ στους τύπους (2.4) και (2.5).

2.7 Προσεγγιστικοί τύποι υψηλότερης τάξης για τις παραγώγους $f^{(k)}(x)$

Αρχίζουμε με την $f'(x)$. Αν θεωρήσουμε τον προσεγγιστικό τύπο $F(h) = \dfrac{\Delta f(x)}{h}$ τάξης $O(h)$ και εφαρμόσουμε τον τύπο βελτίωσης του Richardson για $q = 2$ τότε λαμβάνουμε τον προσεγγιστικό τύπο $F_1(h)$ τάξης $O(h^2)$ έτσι ώστε

$$f'(x) \cong F_1(h) = \frac{2F(h) - F(2h)}{2-1} = 2\frac{f(x+h) - f(x)}{h} - \frac{f(x+2h) - f(x)}{2h}$$

ή

$$f'(x) \cong \frac{1}{2h}(-3f(x) + 4f(x+h) - f(x+2h)), \quad \tau(h) = O(h^2)$$

ο οποίος είναι ο τύπος πεπερασμένων (προς τα πίσω) διαφορών τριών σημείων.

Αν τώρα θεωρήσουμε τον προσεγγιστικό τύπο $F(h) = \dfrac{\delta f(x)}{2h}$ τάξης $O(h^2)$ και εφαρμόσουμε τον τύπο βελτίωσης σφάλματος του Richardson για $q = 2$ λαμβάνουμε τον προσεγγιστικό τύπο $F_1(h)$ τάξης $O(h^4)$ έτσι ώστε

$$f'(x) \cong F_1(h) = \frac{2^2F(h) - F(2h)}{2^2 - 1} = \frac{1}{3}\left(4\frac{f(x+h) - f(x-h)}{2h} - \frac{f(x+2h) - f(x-2h)}{2h}\right)$$

ή

$$f'(x) \cong \frac{1}{12h}\left[f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)\right], \quad \tau(h) = O(h^2)$$

που είναι ο τύπος κεντρικών διαφορών τάξης $O(h^4)$.

Προσθέτοντας κατά μέλη τις σειρές των $f(x+h)$ και $f(x-h)$ και λύνοντας ως προς $f''(x)$ προκύπτει ο προσεγγιστικός τύπος των κεντρικών διαφορών τάξης $O(h^2)$:

$$f''(x) \cong \frac{\delta^2 f(x)}{h^2} = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}, \quad \tau(h) = -\frac{h^2}{12}f^{(4)}(x) + O(h^4)$$

Αν εφαρμόσουμε τον τύπο βελτίωσης σφάλματος του Richardson για $q = 2$ προκύπτει ο τύπος $F_1(h)$ τάξης $O(h^4)$ έτσι ώστε

$$f''(x) \cong F_1(h) = \frac{1}{12h^2}\left[-f(x-2h) + 16f(x-h) - 30f(x) + 16f(x+h) - f(x+2h)\right]$$

που είναι ο τύπος κεντρικών διαφορών τάξης $O(h^4)$.

Άλλοι προσεγγιστικοί τύποι υψηλής τάξης μπορούν να προκύψουν με αντικατάσταση τύπων τάξης $O(h)$ σε μια προσέγγιση Taylor της $f(x+h)$.

Παράδειγμα 2.5. Αν θεωρήσουμε την προσέγγιση Taylor της $f(x+h)$ τάξης $O(h^4)$

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + O(h^4) \tag{2.6}$$

και αντικαταστήσουμε στην (2.6) τις προσεγγίσεις τάξης $O(h)$:

$$f''(x) = \frac{f(x) - 2f(x+h) + f(x+3h)}{h^2} - hf'''(x) + O(h^2)$$

όπου* $f'''(x) = \dfrac{\Delta^3 f(x)}{h^3} + O(h)$, τότε λύνοντας ως προς $f'(x)$ προκύπτει ο τύπος πεπερασμένων διαφορών τεσσάρων σημείων:

$$f'(x) \cong \frac{1}{6h}\left[-11f(x) - 18f(x+h) - 9f(x+2h) + 2f(x+3h)\right], \quad \tau(h) = O(h^3)$$

Παρατήρηση 2.2. Βλέπουμε ότι εφαρμόζοντας τον τύπο βελτίωσης του Richardson για έναν τύπο διαφορών (προς τα εμπρός ή προς τα πίσω) αυξάνει η τάξη ακρίβειας κατά 1, ενώ για έναν τύπο κεντρικών διαφορών η τάξη ακρίβειας αυξάνει κατά 2.

*Αποδεικνύεται επαγωγικά ότι $f^{(k)}(x) \cong \dfrac{\Delta^k f(x)}{h^k} = \dfrac{1}{h}\left(\dfrac{\Delta^{k-1} f(x+h)}{h^{k-1}} - \dfrac{\Delta^{k-1} f(x)}{h^{k-1}}\right)$ με σφάλμα $O(h)$

Κεφάλαιο 3

Αριθμητικές μέθοδοι για Συνήθεις Διαφορικές Εξισώσεις

Θα περιοριστούμε σε συστήματα στα οποία όλες οι εξαρτημένες μεταβλητές (συμβ. y) εξαρτώνται από μια απλή ανεξάρτητη μεταβλητή (συμβ. t), δηλαδή στις *συνήθεις* διαφορικές εξισώσεις. Το πιο απλό Πρόβλημα Αρχικών Τιμών (Π.Α.Τ.) είναι το

$$\frac{dy}{dt} = f(t,y)\Big|_{I_1 \times I_2}, \qquad y(t_0) = y_0$$

όπου $y : I_1 \to I_2$ παραγωγίσιμη στο I_1.

Θα εξετάσουμε τις αριθμητικές μεθόδους Euler, Taylor, Runge-Kutta, Πρόβλεψης-Διόρθωσης (Predictor-Corrector) και στη συνέχεια θα γενικεύσουμε αυτές τις μεθόδους για την επίλυση συστημάτων $n-$πρωτοβαθμίων διαφορικών εξισώσεων εισάγοντας n εξαρτημένες μεταβλητές. Κατόπιν θα τις χρησιμοποιήσουμε για την επίλυση του $n-$τάξης Π.Α.Τ. της μορφής:

$$y^{(n)} = f(t, y, \ldots, y^{(n-1)}), \qquad y^{(k)}(t_0) = y_{ok}, \quad k = 0(1)n-1$$

Τέλος θα περιγράψουμε τις μεθόδους Βολής (ή σκόπευσης) (shooting) και τις μέθόδους πεπερασμένων διαφορών για την επίλυση λύση του προβλήματος συνοριακών τιμών (Π.Σ.Τ.) δυοσημείων

$$y'' = f(t, y, y'), \qquad \begin{matrix} y(a) = \alpha \\ y(b) = \beta \end{matrix}$$

Στη συνέχεια παρουσιάζουμε ένα σχεδιάγραμμα της δομής της ύλης στο παρόν κεφάλαιο.

Πρόβλημα αρχικών τιμών (Π.Α.Τ.)

$$\frac{dy}{dt} = f(t,y)\Big|_{I_1 \times I_2} \quad , \quad y(t_0) = y_0$$

όπου $y : I_1 \to I_2$ παραγωγίσιμη στο I_1

Αριθμητικές μέθοδοι:

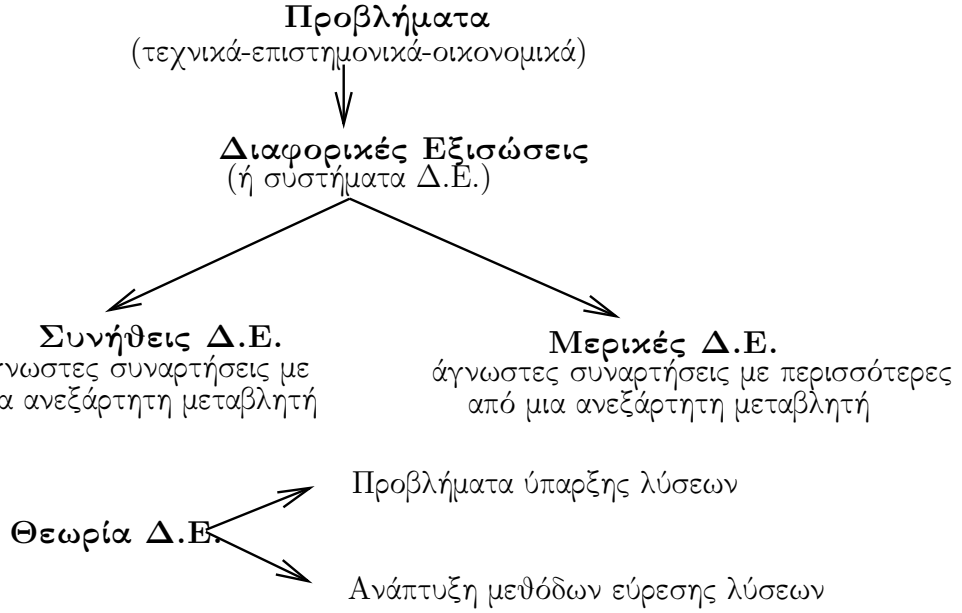
- Euler
- Runge-Kutta
- Predictor-Corrector

Πρόβλημα συνοριακών τιμών(Π.Σ.Τ.)

$$y'' = f(t,y,y') \quad , \quad \begin{matrix} y(a) = \alpha \\ y(b) = \beta \end{matrix}$$

Αριθμητικές μέθοδοι:

- Βολής(ή σκόπευσης)
- Πεπερασμένων διαφορών



Η παρακάτω πρόταση εξασφαλίζει την ύπαρξη και το μονοσήμαντο της λύσης ενός Π.Α.Τ.

Πρόταση 3.1. *Έστω η διαφορική εξίσωση* $\frac{dy}{dt} = f(t,y)$, $y(t_0) = y_0$ *και* $B = \{|t - t_0| < \alpha,\ |y - y_0| < \beta\}$ *ένας τόπος, όπου*

α) $f(t,y)\Big|_B$ *συνεχής και φραγμένη*

β) $f(t,y)\Big|_B$ *πληροί τη τη συνθήκη του Lipschitz με σταθερά* $L > 0$*, δηλαδή*

$$|f(t,y_1) - f(t,y_2)| \le L|y_1 - y_2| \quad \text{για } (t,y_1),(t,y_2) \in B$$

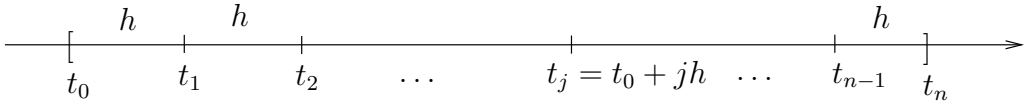
Τότε υπάρχει ακριβώς μια λύση $y = y(t)$ *ορισμένη στο διάστημα* $|t - t_0| < \alpha$*.*

3.1 Μέθοδος Euler

Έστω $[t_0, t_n]$ ένα διάστημα στο οποίο αναζητούμε τη λύση του προβλήματος αρχικών τιμών

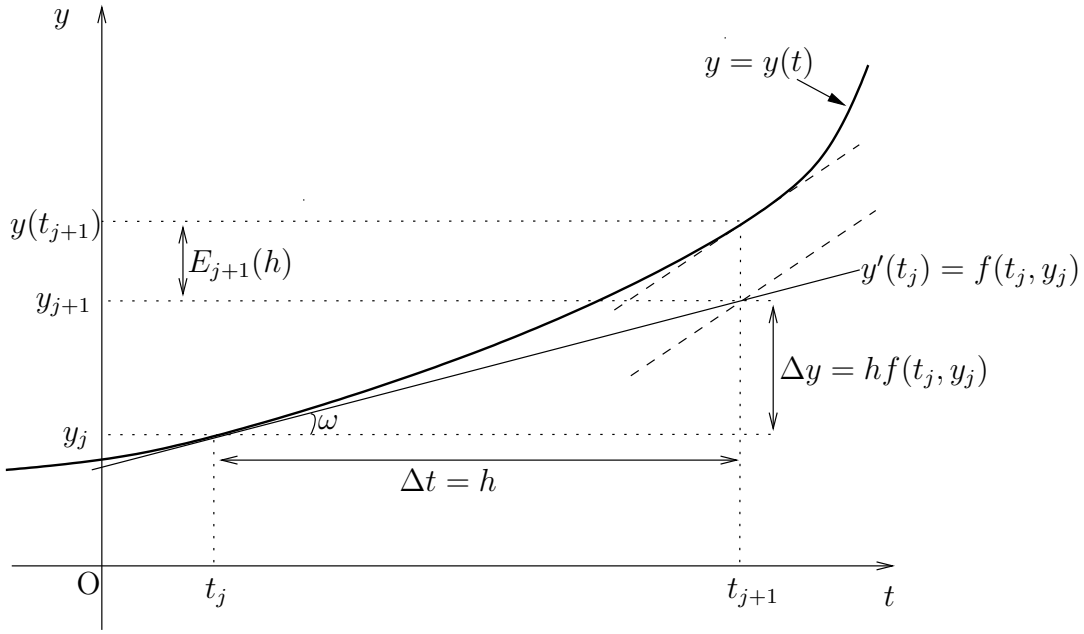
$$\text{(Π.Α.Τ.)} \qquad \frac{dy}{dt} = f(t, y), \quad y(t_0) = y_0$$

Αν υποδιαιρέσουμε το $[t_0, t_n]$ σε n διαστήματα πλάτους $h = \dfrac{t_n - t_0}{n}$ παίρνουμε τα ισαπέχοντα σημεία $t_1, t_2, \ldots, t_{n-1}$ όπως φαίνεται στο σχήμα



Μια αριθμητική μέθοδος για τη επίλυση του Π.Α.Τ. ξεκινά με $y_0 = y(t_0)$ και στη συνέχεια δημιουργεί τις τιμές $y_1, y_2, \ldots, y_n$ έτσι ώστε η τιμή y_j να προσεγγίζει την ακριβή τιμή της λύσης $y(t_j)$ για $j = 1(1)n$.

Για μικρό h, η καμπύλη της λύσης $y = y(t)$ προσεγγίζεται στο διάστημα $[t_j, t_{j+1}]$ με την εφαπτομένη ευθεία στο σημείο $(t_j, y(t_j))$, όπως φαίνεται στο σχήμα $\dfrac{\Delta y}{\Delta t} = \tan\omega = f(t_j, y_j)$



Ο αναγωγικός τύπος που βασίζεται σε αυτή τη στρατηγική είναι η μέθοδος του Euler:

$$y_{j+1} = y_j + hf(t_j, y_j) \quad , \quad j = 0(1)n - 1$$

Το ολικό σφάλμα αποκοπής στη θέση $j + 1$ είναι $E_{j+1}(h) = y(t_{j+1}) - y_{j+1}$.

Παράδειγμα 3.1. Χρησιμοποιήστε τη μέθοδο του Euler για την επίλυση του Π.Α.Τ.

$$\frac{dy}{dt} = -ty^2, \quad y(2) = 1$$

στο $[2, 3]$ με βήμα $h = 0.1$ και $h = 0.05$. Εξετάστε την επιτυγχανόμενη ακρίβεια, δεδομένου ότι η ακριβής λύση είναι $y(t) = \dfrac{2}{t^2 - 2}$.

Επειδή $f_j = -t_j y_j^2$ ο τύπος του Euler για αυτό το πρόβλημα είναι

$$y_{j+1} = y_j + h(-t_j y_j^2) \quad \cong \quad y(t_{j+1})$$

όπου $\quad t_j = 2 + jh.$

Ξεκινώντας με $t_0 = 2$ και $y_0 = 1$ και παίρνοντας $h = 0.1$ με αριθμητική τεσσάρων σημαντικών ψηφίων είναι:

$$\begin{aligned}
j = 0 \quad &: \quad y_1 = y_0 - h(t_0 y_0^2) = 0.8000 \cong y(2.1) \\
j = 1 \quad &: \quad y_2 = y_1 - h(t_1 y_1^2) = 0.6656 \cong y(2.2) \\
j = 2 \quad &: \quad y_3 = y_2 - h(t_2 y_2^2) = 0.5681 \cong y(2.3) \\
j = 3 \quad &: \quad y_4 = y_3 - h(t_3 y_3^2) = 0.4939 \cong y(2.4)
\end{aligned}$$

Παρόμοια για $h = 0.05$ έχουμε

$$\begin{aligned}
j = 0 \quad &: \quad y_1 = y_0 - h(t_0 y_0^2) = 0.9000 \cong y(2.05) \\
j = 1 \quad &: \quad y_2 = y_1 - h(t_1 y_1^2) = 0.8170 \cong y(2.10) \\
j = 2 \quad &: \quad y_3 = y_2 - h(t_2 y_2^2) = 0.7469 \cong y(2.15) \\
j = 3 \quad &: \quad y_4 = y_3 - h(t_3 y_3^2) = 0.6869 \cong y(2.20)
\end{aligned}$$

Τα αποτελέσματα πινακοποιούνται για $t = 2, 2.1, 2.2, \ldots, 3$ και $h = 0.1, 0.05$ όπου οι ακριβείς τιμές $y(t_j)$ προκύπτουν από την ακριβή λύση $y(t) = \frac{2}{t^2-2}$.

Μια εξέταση των σφαλμάτων $E(0.1)$ και $E(0.05)$ από τον πίνακα αποτελεσμάτων δείχνει ότι υποδιπλασιάζοντας το h υποδιπλασιάζεται προσεγγιστικά το σφάλμα $E(h)$.

Αυτό φανερώνει ότι η μέθοδος Euler έχει σφάλμα τάξης $O(h)$. Υποθέτοντας ότι αυτό αληθεύει, μπορούμε να εφαρμόσουμε το βελτιωτικό τύπο του Richardson με $F(h) = y_j(h)$, $h = 0.05,\ n = 1$, και $q = \frac{0.1}{0.05} = 2$ για να πάρουμε βελτιωμένες προσεγγίσεις:

$$y_j(0.05)_{\text{βελτ.}} = \frac{2y_j(0.05) - y_j(0.1)}{2 - 1} = 2y_j(0.05) - y_j(0.1)$$

Οι τιμές που προκύπτουν είναι πιο ακριβείς από τις αντίστοιχες $y_j(0.05)$ τουλάχιστον κατά ένα δεκαδικό ψηφίο.

3.2 Η τάξη μιας αριθμητικής μεθόδου

Για να εξετάσουμε την ακρίβεια μιας αριθμητικής μεθόδου διακρίνουμε δυο ειδών σφάλματα αποκοπής στο t_{j+1}:

$$e_{j+1}(h) = y(t_{j+1}) - y_{j+1}, \qquad \text{αν } y_j = y(t_j) \qquad \text{τοπικό (ή ανά βήμα) σφάλμα}$$
$$\text{και } E_{j+1}(h) = y(t_{j+1}) - y_{j+1}, \qquad \text{αν } y_j \cong y(t_j) \qquad \text{συσσωρευμένο (ή ολικό) σφάλμα}$$

Για τη μέθοδο του Euler το $e_{j+1}(h)$ είναι απλά το υπόλοιπο του αναπτύγματος Taylor πρώτου βαθμού της y, δηλαδή

$$y(t_j + h) \cong y(t_j) + hy'(t_j) \qquad \text{όπου} \qquad y(t_j) = f(t_j, y(t_j))$$

'ρα το $e_{j+1}(h)$ είναι τάξης $O(h^2)$, ενώ το $E_{j+1}(h)$ είναι $O(h)$ καθώς $n = \frac{t_n - t_0}{h}$ και

$$E_{j+1}(h) = \sum_{i=1}^{n} e_{j+1}(h) = n \cdot O(h^2) = \frac{t_n - t_0}{h} \cdot O(h^2) = O(h)$$

Στην επόμενη παράγραφο θα αναπτύξουμε δυο υψηλότερης τάξης μεθόδους, τις μεθόδους Taylor και Runge-Kutta.

3.3 Μέθοδος Taylor

Η μέθοδος Euler μπορεί να προκύψει από το ανάπτυγμα κατά Taylor τάξης $O(h^2)$

$$y(t_{j+1}) = y(t_j + h) \cong y(t_j) + hy'(t_j) \tag{3.1}$$

αντικαθιστώντας τις άγνωστες ακριβείς τιμές $y(t_j)$ και $y'(t_j)$ με τις τρέχουσες προσεγγίσεις $y(t_j) \cong y_j$ και $y'(t_j) \cong f_j = f(t_j, y_j)$. Έτσι ο τύπος της (πρώτης τάξης) μεθόδου Euler γράφεται:

$$y_{j+1} = y_j + h\Phi_{T,1} \qquad \text{όπου} \qquad \Phi_{T,1} = f_j \tag{3.2}$$

Η φυσική επέκταση της (3.2) για μια $n-$τάξης μέθοδο είναι να ξεκινήσουμε με το ανάπτυγμα Taylor τάξης $O(h^{n+1})$

$$y(t_{j+1}) \cong y(t_j) + hy'(t_j) + \frac{h^2}{2!}y''(t_j) + \cdots + \frac{h^n}{n!}y^{(n)}(t_j) \tag{3.3}$$

και αντικαθιστώντας τις άγνωστες ακριβείς τιμές $y(t_j),\ y'(t_j),\ y''(t_j), \ldots,\ y^{(n)}(t_j)$ με τις υπολογίσιμες προσεγγίσεις $y_j,\ y'_j,\ y''_j, \ldots,\ y_j^{(n)}$ προκύπτει ο τύπος της μεθόδου Taylor $n-$τάξης:

$$y_{j+1} = y_j + h\Phi_{T,n} \qquad \text{όπου} \qquad \Phi_{T,n} = f_j + \frac{h}{2!}y''_j + \cdots + \frac{h^{n-1}}{n!}y_j^{(n)} \tag{3.4}$$

Για να πάρουμε τις επιθυμητές απαιτούμενες προσεγγίσεις των παραγώγων της y εφαρμόζουμε τον κανόνα της αλυσίδας* για συναρτήσεις δυο μεταβλητών, ώστε να διαφορίσουμε την $y' = f(t,y)$ όπου $y = y(t)$ ως προς t. Μια διαφόριση δίνει

$$y'' = f_t + f_y y' = f_t + f_y \cdot f \tag{3.5}$$

όπου $f_t = \dfrac{\partial f(t,y)}{\partial t}$, $f_y = \dfrac{\partial f(t,y)}{\partial y}$ και $f = f(t,y)$.

Παρόμοια

$$y''' = \frac{d}{dt}(f_t) + \frac{d}{dt}(f_y y') = (f_{tt} + f_{ty}y') + \Big(f_y y'' + (f_{yt} + f_{yy}y')y'\Big)$$

Με την υπόθεση ότι οι μερικές παράγωγοι είναι συνεχείς προκύπτει $f_{ty} = f_{yt}$. Έτσι έχουμε

$$y''' = f_{tt} + 2f_{ty}y' + f_{yy}(y')^2 + f_y y'' \tag{3.6}$$

Από (3.5) και αφού $y' = f(t,y)$ η (3.6) δίνει

$$y''' = f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_y(f_t + f_y f)$$

Τελικά κάθε παράγωγος $y^{(k)}$ μπορεί να εκφραστεί μόνο συναρτήσει της f και των μερικών της παραγώγων. Αυτές οι εκφράσεις μπορούν να χρησιμοποιηθούν για να προσεγγίσουμε τις $y^{(k)}(t_j)$ στην (3.4).

Για παράδειγμα, από την (3.5) η $y''(t_j)$ μπορεί να προσεγγιστεί από τον τύπο

$$y''_j = [f_t + f_y f]_j = f_t(t_j, y_j) + f_y(t_j, y_j) \cdot f(t_j, y_j)$$

θέτοντας αυτήν την προσέγγιση στην (3.4) και για $n = 2$ προκύπτει ο τύπος της μεθόδου Taylor δεύτερης τάξης:

$$y_{j+1} = y_j + h\Phi_{T,2} \qquad \text{όπου} \qquad \Phi_{T,2} = f_j + \frac{h}{2}[f_t + f_y f]_j \tag{3.7}$$

Παρόμοια μπορεί να προκύψει ο τύπος της μεθόδου Taylor τρίτης τάξης

$$y_{j+1} = y_j + h\Phi_{T,3} \qquad \text{όπου} \qquad \Phi_{T,3} = \Phi_{T,2} + \frac{h^2}{6}y'''_j$$

όπου το y'''_j υπολογίζεται από τον τύπο (3.6) στο σημείο (t_j, y_j).

---

*__Κανόνας αλυσίδας:__ Αν $f(x_1, x_2, \ldots, x_n)$ διαφορίσιμη συνάρτηση ως προς κάθε $x_i(u_1, u_2, \ldots, u_m)$ τότε για κάθε $j = 1, 2, \ldots, m$ ισχύει

$$\frac{\partial f}{\partial u_j} = \frac{\partial f}{\partial x_1} \cdot \frac{\partial x_1}{\partial u_j} + \frac{\partial f}{\partial x_2} \cdot \frac{\partial x_2}{\partial u_j} + \cdots + \frac{\partial f}{\partial x_n} \cdot \frac{\partial x_n}{\partial u_j}$$

όπου οι μερικές παράγωγοι των x_i ως προς u_j υπάρχουν για ένα σταθερό $\boldsymbol{v} = (v_1, v_2, \ldots, v_m)$ και οι μερικές παράγωγοι της f ως προς x_j υπάρχουν στο αντίστοιχο $x(\boldsymbol{v}) = (x_1(\boldsymbol{v}), x_2(\boldsymbol{v}), \ldots, x_n(\boldsymbol{v}))$.

3.4 Μέθοδος Runge-Kutta δεύτερης τάξης

Όπως είδαμε ο τύπος της μεθόδου Taylor δεύτερης τάξης είναι

$$y_{j+1} = y_j + h\Phi_{T,2} \qquad \text{όπου} \qquad \Phi_{T,2} = f_j + \frac{h}{2}\Big(f_t(t_j, y_j) + f_j \cdot f_y(t_j, y_j)\Big) \tag{3.8}$$

Πρώτος ο γερμανός μαθηματικός Runge παρατήρησε ότι η έκφραση $\Phi_{T,2}$ μοιάζει με την προσέγγιση $O(h^2)$ του αναπτύγματος Taylor:

$$f(t_j + ph,\ \ y_j + qhf_j) \cong f(t_j) + df(t_j, y_j) = f_j + phf_t(t_j, y_j) + qhf_j f_y(t_j, y_j) \tag{3.9}$$

όπου $df(t, y)$ το ολικό διαφορικό[†] της f. Θεωρώντας $p = q = \frac{1}{2}$ στην (3.9) και συγκρίνοντας με την έκφραση του $\Phi_{T,2}$ προκύπτει

$$\Phi_{T,2} \cong f(t_j + ph,\quad y_j + qhf_j), \quad \text{με σφάλμα } O(h^2) \tag{3.10}$$

Αν αντικαταστήσουμε την (3.10) στη (3.7) προκύπτει ο τύπος της τροποποιημένης μεθόδου του Euler:

$$y_{j+1} = y_j + hf\left(t_j + \frac{1}{2}h,\ \ y_j + \frac{1}{2}hf_j\right), \quad j = 0, 1, 2, \cdots n-1 \tag{3.11}$$

Παρατηρούμε ότι το σφάλμα της μεθόδου αυτής σε κάθε βήμα (τοπικό σφάλμα) είναι τάξης $O(h^3)$ δηλαδή η μέθοδος αυτή είναι δεύτερης τάξης. Σε κάθε βήμα χρειάζονται δυο υπολογισμοί της $f\cdot$ πρώτα στο σημείο (t_j, y_j) κι έπειτα στο $(t_j + \frac{1}{2}h, y_j + \frac{1}{2}hf_j)$.

Ακόμα, είναι δυνατό να προκύψουν άλλες μέθοδοι δεύτερης τάξης προσεγγίζοντας το $\Phi_{T,2}$ με ένα άθροισμα με συντελεστές βάρους στις κλίσεις, δηλαδή

$$\Phi_{T,2} \cong \alpha_1 f(t_j,\ y_j) + \alpha_2 f(t_j + ph,\ \ y_j + qhf_j) \tag{3.12}$$

[†]Αν $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ και υπάρχουν οι μερικές παράγωγοι $f_{x_1}(\mathbf{x}),\ f_{x_2}(\mathbf{x}), \ldots,\ f_{x_n}(\mathbf{x})$ τότε το ολικό διαφορικό της f στο $\mathbf{x}$ ορίζεται ως εξής:

$$df(\mathbf{x}) = f_{x_1}(\mathbf{x})dx_1 + f_{x_2}(\mathbf{x})dx_2 + \cdots + f_{x_n}(\mathbf{x})dx_n$$

όπου τα $dx_1, dx_2, \ldots, dx_n$ μπορούν να θεωρηθούν ως μεταβολές των $x_1, x_2, \ldots, x_n$ αντίστοιχα.

Γνωρίζουμε ότι μπορεί να χρησιμοποιηθεί η γραμμική προσέγγιση

$$f(\mathbf{x} + d\mathbf{x}) = f(x_1 + dx_1, x_2 + dx_2, \ldots, x_n + dx_n) \cong f(\mathbf{x}) + df(\mathbf{x})$$

όταν η μεταβολή $d\mathbf{x} = (dx_1, dx_2, \ldots, dx_n)$ είναι αρκούντως μικρή, υπό την έννοια ότι η νόρμα $\|d\mathbf{x}\|_\infty \cong 0$, όπου $\|d\mathbf{x}\|_\infty = \max\{|dx_1|, |dx_2|, \ldots, |dx_n|\}$.

όπου τα βάρη $\alpha_1,\ \alpha_2$ και οι παράγοντες $p,\ q$ προσδιορίζονται έτσι ώστε η προσέγγιση να είναι τάξης $O(h^2)$. Οι αλγόριθμοι που βασίζονται σε αυτή τη στρατηγική λέγονται μέθοδοι Runge-Kutta.

Αν αντικαταστήσουμε την (3.9) στην (3.8) και εξισώσουμε τους συντελεστές των $f_t(t_j, y_j)$, $f_y(t_j, y_j)$ με εκείνους της (3.7) προκύπτει

$$f_j+\frac{h}{2}\Big(f_t(t_j,y_j)+f(t_j,y_j)\cdot f_y(t_j,y_j)\Big) = \alpha_1 f(t_j,y_j)+\alpha_2 f(t_j,y_j)+\alpha_2 p h f_t(t_j,y_j)+\alpha_2 q h f(t_j,y_j)f_y(t_j,y_j)$$

δηλαδή

$$\alpha_1+\alpha_2=1 \qquad \text{και} \qquad \alpha_2 p=\alpha_2 q=\frac{1}{2}$$

Επιλέγοντας αυθαίρετα το α_2 έχουμε

$$\alpha_1=1-\alpha_2 \qquad \text{και} \qquad p=q=\frac{1}{2\alpha_2}$$

όπου $\alpha_2 \neq 0$.

- Για $\alpha_2 = 1$ προκύπτει η τροποποιημένη μέθοδος Euler
- Για $\alpha_2 = \frac{1}{2}$, δηλαδή $\alpha_1 = \frac{1}{2}$ και $p = q = 1$, προκύπτει ο τύπος της μεθόδου Huen:

$$y_{j+1} = y_j + \frac{h}{2}\Big[f_j + f(t_j + h,\ \ y_j + h f_j)\Big], \quad j = 0(1)n-1 \tag{3.13}$$

3.5 Μέθοδοι Runge-Kutta ανώτερης τάξης

Για να πάρουμε τον τύπο της μεθόδου Runge-Kutta τέταρτης τάξης ξεκινούμε με τον τύπο της μεθόδου Taylor τέταρτης τάξης

$$y_{j+1} = y_j + h\Phi_{T,4} \qquad \text{όπου} \qquad \Phi_{T,4} = f_j + \frac{h}{2}y_j'' + \frac{h^2}{6}y_j''' + \frac{h^3}{24}y_j'''' \tag{3.14}$$

και έπειτα προσεγγίζουμε το $\Phi_{T,4}$ με ένα άθροισμα με συντελεστές βάρους

$$\Phi_{T,4} \cong w_1 m_1 + w_2 m_2 + w_3 m_3 + w_4 m_4$$

όπου οι χρησιμοποιούμενες κλίσεις m_1, m_2, m_3, m_4 ορίζονται αναγωγικά ως εξής:

$$\begin{aligned} m_1 &= f(t_j,\ y_j) = f_j \\ m_2 &= f(t_j + p_2 h,\ \ y_j + h q_{21} m_1) \\ m_3 &= f\Big(t_j + p_3 h,\ \ y_j + h(q_{31}m_1 + q_{32}m_2)\Big) \\ m_4 &= f\Big(t_j + h,\ \ y_j + h(q_{41}m_1 + q_{42}m_2 + q_{43}m_3)\Big) \end{aligned} \tag{3.15}$$

Τα βάρη w_i και οι παράγοντες(scale factors) p και q προσδιορίζονται με αντικατάσταση των εκφράσεων (3.15) στον τύπο τον αντίστοιχο του (3.9) για το $\Phi_{T,4}$ της προσέγγισης Taylor τάξης $O(h^4)$.

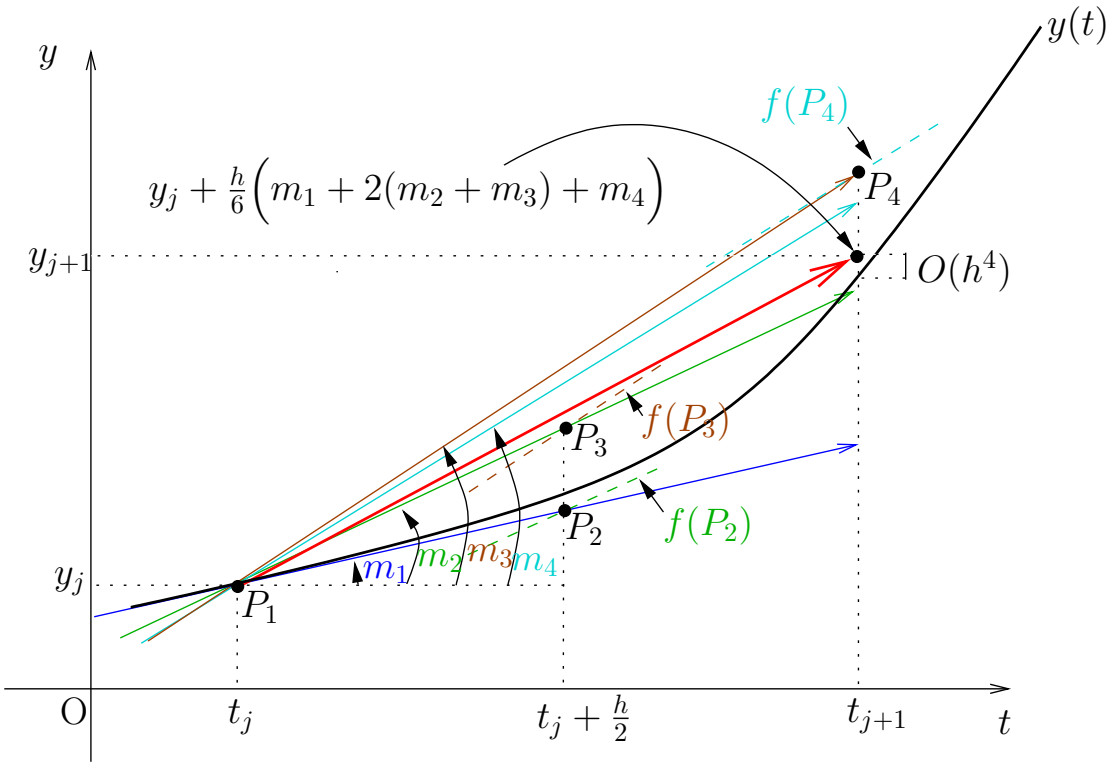
Η πιο συχνά χρησιμοποιούμενη μέθοδος είναι η μέθοδος Runge-Kutta τέταρτης τάξης η οποία προκύπτει παίρνοντας $p_2 = p_3 = \frac{1}{2}$ (οπότε $q_{21} = \frac{1}{2},\ q_{31} = 0,\ q_{32} = \frac{1}{2},\ q_{41} = q_{42} = 0,\ q_{43} = 1$), δηλαδή

$$y_{j+1} = y_j + \frac{h}{6}\Big(m_1 + 2(m_2 + m_3) + m_4\Big) \tag{3.16}$$

όπου

$$\begin{aligned} m_1 &= f(t_j,\ y_j) \\ m_2 &= f\Big(t_j + \frac{1}{2}h,\ \ y_j + \frac{1}{2}hm_1\Big) \\ m_3 &= f\Big(t_j + \frac{1}{2}h,\ \ y_j + \frac{1}{2}hm_2\Big) \\ m_4 &= f\Big(t_j + h,\ \ y_j + hm_3\Big) \end{aligned}$$

Η μέθοδος αυτή θα συμβολίζεται σύντομα με RK4. Οι κλίσεις m_1, m_2, m_3, m_4 είναι οι τιμές της f στα σημεία P_1, P_2, P_3, P_4, όπως αυτά φαίνονται στο παρακάτω σχήμα:



Στο σχήμα θεωρούμε ότι η y_j είναι η ακριβής τιμή $y(t_j)$. Οι κλίσεις που υπολογίζονται είναι

$$\begin{aligned}
P_1 &= (t_j, \ \ y_j) && \Rightarrow f(P_1) = f(t_j, \ \ y_j) = y'(t_j) = m_1 \\
P_2 &= (t_j + \frac{h}{2}, \ \ y_j + \frac{h}{2}m_1) && \Rightarrow f(P_2) = f(t_j + \frac{h}{2}, \ \ y_j + \frac{h}{2}m_1) = m_2 \\
P_3 &= (t_j + \frac{h}{2}, \ \ y_j + \frac{h}{2}m_2) && \Rightarrow f(P_3) = f(t_j + \frac{h}{2}, \ \ y_j + \frac{h}{2}m_2) = m_3 \\
P_4 &= (t_j + h, \ \ y_j + hm_3) && \Rightarrow f(P_4) = f(t_j + h, \ \ y_j + hm_3) = m_4
\end{aligned}$$

Σημειώνεται ότι αν η $f(t, y)$ εξαρτάται μόνο από το t, δηλαδή $f(t, y) = g(t)$ τότε χρησιμοποιούμε τον κανόνα του Simpson για να ολοκληρώσουμε την $y'(t) = g(t)$ από t_j μέχρι t_{j+1}.

Παράδειγμα 3.2. Χρησιμοποιήστε την RK4 με $h = 0.1$ για την επίλυση του Π.Α.Τ.

$$\frac{dy}{dt} = -ty^2, \quad y(2) = 1$$

και συγκρίνετε τα αποτελέσματα με την ακριβή λύση, δεδομένου ότι αυτή είναι $y(t) = \dfrac{2}{t^2 - 2}$.

Ξεκινώντας με $t_0 = 2, \ y_0 = 1$ η (3.16) δίνει

$$\begin{aligned}
m_1 &= f(2, \ 1) = -2 \cdot 1^2 = -2 \\
m_2 &= f(2.05, \ \ 1 + 0.05 \cdot (-2)) = -(2.05) \cdot (0.9)^2 = -1.6605 \\
m_3 &= f(2.05, \ \ 1 + 0.05 \cdot (-1.6605)) = -(2.05) \cdot (0.916975)^2 = -1.72373 \\
m_4 &= f(2.1, \ \ 1 + 0.1 \cdot (-1.72373)) = -(2.1) \cdot (0.82763)^2 = -1.43843
\end{aligned}$$

$$y_1 = y_0 + \frac{0.1}{6}\Big(-2 + 2(-1.6605 - 1.72373) - 1.43843\Big) = 0.829885$$

Η ακριβής τιμή είναι $y(t_1) = 0.829876$.

Παρατηρούμε ότι η RK4 δίνει εξαιρετικά ακριβείς τιμές του $y(t_{j+1})$. Αυτό διότι η $f(t, y)$ είναι πολυώνυμο βαθμού μικρότερου από τέσσερα ως προς t και y, οπότε το ανάπτυγμα Taylor τετάρτου βαθμού τάξης $O(h^5)$ είναι ακριβές.

Υπάρχουν βέβαια ορισμένες ατέλειες στην RK4. Μια από αυτές είναι το γεγονός ότι σε κάθε βήμα πρέπει να υπολογίζονται τέσσερις τιμές κλίσεων. Αυτή η ατέλεια μπορεί να είναι σοβαρή όταν η συνάρτηση-κλίση είναι πολύπλοκη ή ο διαθέσιμος χρόνος για τον υπολογισμό της λύσης είναι περιορισμένος. Μια αντιμετώπιση αυτού είναι να σταματά κάθε μερικά βήματα και να επαναλαμβάνει την αριθμητική λύση από το t_j στο t_{j+1} χρησιμοποιώντας την RK4 υποδιπλασιάζοντας το μέγεθος βήματος(πλάτος) h.

Η διαφορά μεταξύ των δυο προσεγγίσεων του $y(t_{j+1})$ μπορεί να χρησιμοποιηθεί για να εκτιμήσει το σφάλμα και αν είναι απαραίτητο να μεταβάλει το h. Οπωσδήποτε για να γίνει αυτό χρειάζονται 8 επιπλέον υπολογισμοί ανά έλεγχο.

Μια εναλλακτική αντιμετώπιση είναι να χρησιμοποιήσουμε περισσότερες από τέσσερις (αλλά όχι λιγότερες από 8) τιμές $f(t, y)$ ανά βήμα και με αυτήν την επιπλέον πληροφορία να πάρουμε δυο εκτιμήσεις του $y(t_{j+1})$. Για να είναι μια μέθοδος τέταρτης τάξης πρέπει η πρώτη εκτίμηση για το y_{j+1} να έχει τοπικό σφάλμα $e_{j+1}(h) = O(h^5)$. Στη δεύτερη εκτίμηση, ένας τύπος υψηλότερης τάξης, συνήθως $O(h^6)$ μας επιτρέπει να πάρουμε μια υπολογίσιμη εκτίμηση για το τοπικό σφάλμα αποκοπής στη θέση $j + 1$.

Αυτή η εκτίμηση του $e_{j+1}(h)$ μπορεί να χρησιμοποιηθεί για να αναθέτει το h για το επόμενο βήμα: Ο στόχος είναι να παίρνουμε μεγαλύτερο βήμα όταν η $y'(t)$ μεταβάλλεται αργά και μικρότερο βήμα όταν η $y'(t)$ μεταβάλλεται γρήγορα (π.χ. λόγω ταλάντωσης ή μιας κατακόρυφης ασύμπτωτης στο $t \cong t_j$).

Μια από τις πιο αποτελεσματικές μεθόδους που χρησιμοποιούν αυτή τη στρατηγική είναι η μέθοδος Runge-Kutta-Fehlberg τέταρτης τάξης (RKF4), η οποία χρησιμοποιεί έξι επαναλήψεις ανά βήμα.

3.6 Μέθοδοι πολλαπλού βήματος (στρατηγικές Πρόβλεψης–

Στις προηγούμενες μεθόδους (Taylor και Runge-Kutta) όλες οι απαιτούμενες εκτιμήσεις της $f(t, y)$ τον υπολογισμό του y_{j+1} γίνονται αφού προηγουμένως υπολογιστεί το y_j.

Οι μέθοδοι με την ιδιότητα αυτή λέγονται *αυτο-εκκινούμενες*(self-starting) μέθοδοι διότι μπορούν να εφαρμοσθούν ξεκινώντας με την αρχική τιμή y_0. Δυστυχώς αυτές οι μέθοδοι δεν χρησιμοποιούν τις προηγούμενες επιτευχθείσες τιμές, δηλαδή τις

$$\begin{aligned} &\ldots,\ y_{j-4},\ y_{j-3},\ y_{j-2},\ y_{j-1} \quad \text{όπου} \quad y_j \cong y(t_j) \\ &\ldots,\ f_{j-4},\ f_{j-3},\ f_{j-2},\ f_{j-1} \quad \text{όπου} \quad f_j \cong f(t_j, y_j) \cong y'(t_j) \end{aligned}$$

Οι μέθοδοι που χρησιμοποιούν τις πληροφορίες των τιμών $y(t),\ y'(t)$ με $t < t_j$ για τον υπολογισμό της προσεγγιστικής τιμής y_{j+1} λέγονται *μέθοδοι πολλαπλού βήματος*(multistep).

Οι τύποι για όλες τις μεθόδους (self-starting και multistep) μπορούν να προκύψουν εφαρμόζοντας το Θεμελιώδες Θεώρημα του Απειροστικού Λογισμού. Πράγματι, προκύπτει ο

ακόλουθος τύπος ολοκλήρωσης για να πάρουμε την τιμή $y(t_{j+1})$ από την $y(t_j)$:

$$\int_{t_j}^{t_{j+1}} y'(t)dt = \Big[y(t)\Big]_{t_j}^{t_{j+1}} \iff y(t_{j+1}) = y(t_j) + \int_{t_j}^{t_{j+1}} f(t, y(t))dt \tag{3.17}$$

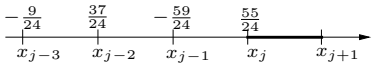
Επειδή γενικά δεν είναι γνωστή η $y(t)$ δε μπορεί να εφαρμοσθεί άμεσα ο τύπος (3.17). Εφόσον όμως γνωρίζουμε τις τιμές $y_j \cong y(t_j)$ και $f_j \cong y'(t_j)$ μπορούμε να χρησιμοποιήσουμε αυτές σε έναν τύπο αριθμητικής ολοκλήρωσης για το ολοκλήρωμα του (3.17), για να προκύψει η τιμή $y_{j+1} \cong y(t_{j+1})$. Το αποτέλεσμα θα είναι ένας τύπος της μορφής

$$y_{j+1} = y_j + h\Phi(h, t_j, y_j, f_j, \dots) \tag{3.18}$$

όπου η Φ μπορεί να είναι συνάρτηση και άλλων τιμών της y και της $f(t, y)$ και πολλαπλασιασμένη με το h εκφράζει έναν τύπο ολοκλήρωσης της $f(t, y)$ στο διάστημα $[t_j, t_{j+1}]$ τάξης $O(h^{n+1})$.

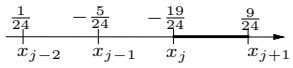
Ενδιαφέρον παρουσιάζουν οι τύποι αριθμητικής ολοκλήρωσης του Adams:

- Adams $O(h^5)$ Predictor:



$$\begin{aligned} \int_{x_j}^{x_{j+1}} f(x)dx &\cong \frac{h}{24}\Big(-9f(x_{j-3}) + 37f(x_{j-2}) - 59f(x_{j-1}) + 55f(x_j)\Big) \\ \tau(h) &= \frac{251}{720}h^5 f^{(4)}(\xi) \quad \text{όπου} \quad x_{j-3} \le \xi \le x_{j+1} \end{aligned} \tag{3.19}$$

- Adams $O(h^5)$ Corrector:



$$\begin{aligned} \int_{x_j}^{x_{j+1}} f(x)dx &\cong \frac{h}{24}\Big(f(x_{j-2}) - 5f(x_{j-1}) + 19f(x_j) + 9f(x_{j+1})\Big) \\ \tau(h) &= -\frac{19}{720}h^5 f^{(4)}(\xi) \quad \text{όπου} \quad x_{j-2} \le \xi \le x_{j+1} \end{aligned} \tag{3.20}$$

Η χρήση ενός $(n+1)-$τάξης τύπου αριθμητικής ολοκλήρωσης στη θέση του $h\Phi$ οδηγεί σε μια $n-$τάξης μέθοδο.

3.7 Μέθοδος Πρόβλεψης–Διόρθωσης του Adams

Αν στον τύπο (3.18) πάρουμε το $h\Phi$ να είναι ο τύπος αριθμητικής ολοκλήρωσης του Adams τάξης $O(h^5)$ τότε προκύπτουν:

- Τύπος πρόβλεψης του Adams (Adams predictor)

$$\begin{aligned} p_{j+1} &= y_j + \frac{h}{24}\Big(-9f_{j-3} + 37f_{j-2} - 59f_{j-1} + 55f_j \Big) \\ e_{j+1}(h)_p &= y(t_{j+1}) - p_{j+1} = \frac{251}{720} y^{(5)}(\xi_p) \cdot h^5 \quad \text{όπου} \quad t_{j-3} \le \xi_p \le t_{j+1} \end{aligned} \tag{3.21}$$

- Τύπος διόρθωσης του Adams (Adams corrector)

$$\begin{aligned} c_{j+1} &= y_j + \frac{h}{24}\Big(f_{j-2} - 5f_{j-1} + 19f_j + 9f_{j+1} \Big) \\ e_{j+1}(h)_c &= y(t_{j+1}) - c_{j+1} = -\frac{19}{720} y^{(5)}(\xi_c) \cdot h^5 \quad \text{όπου} \quad t_{j-2} \le \xi_c \le t_{j+1} \end{aligned} \tag{3.22}$$

Ο τύπος (3.21) μπορεί να χρησιμοποιηθεί για να παράγει μια προβλεπόμενη εκτίμηση p_{j+1}. Η τιμή $f(t_{j+1}, p_{j+1})$ μπορεί να χρησιμοποιηθεί ως f_{j+1} στον τύπο (3.22) για να πετύχουμε τη διορθωμένη εκτίμηση c_{j+1}. Οι μέθοδοι που ακολουθούν αυτή τη στρατηγική λέγονται μέθοδοι *Πρόβλεψης-Διόρθωσης*(Predictor-Corrector).

Αν οι τύποι για τα p_{j+1} και c_{j+1} είναι της ίδιας τάξης, τότε οι τιμές τους μπορούν να χρησιμοποιηθούν για να εκτιμήσουν το τοπικό σφάλμα του c_{j+1}. Για παράδειγμα, αν το h στις (3.21),(3.22) είναι αρκετά μικρό, έτσι ώστε η $y^{(5)}$ να είναι περίπου σταθερή στο $[t_j, t_{j+1}]$, τότε διαιρώντας κατά μέλη τα σφάλματα στις (3.21),(3.22) προκύπτει

$$19(y(t_{j+1}) - p_{j+1}) + 251(y(t_{j+1}) - c_{j+1}) \cong 0$$

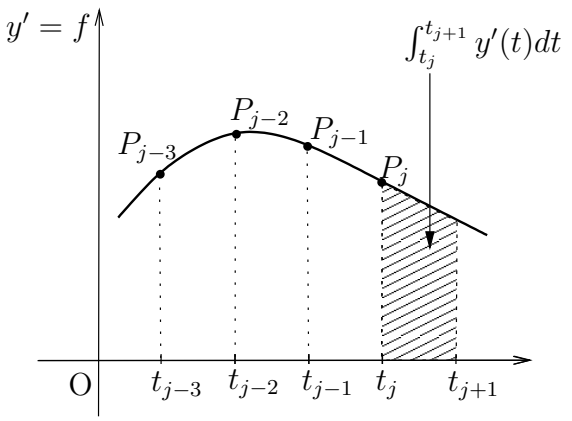
και λύνοντας ως προς $y(t_{j+1})$ βρίσκουμε

$$y_{j+1} \cong \frac{251}{270} c_{j+1} + \frac{19}{270} p_{j+1} = c_{j+1} - \frac{19}{270}(c_{j+1} - p_{j+1}) \tag{3.23}$$

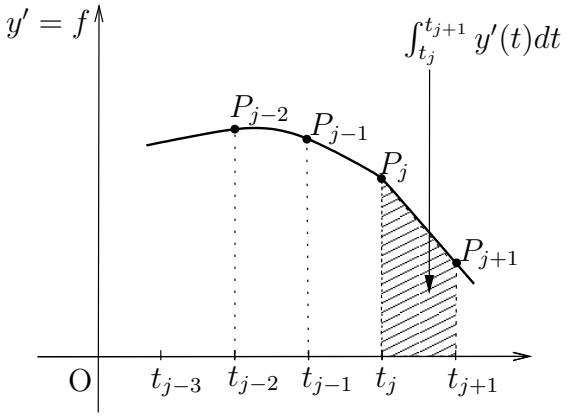
Συγκρίνοντας το σφάλμα στην (3.22) και την (3.23) παρατηρούμε ότι η ποσότητα

$$\delta_{j+1} = -\frac{19}{720}(c_{j+1} - p_{j+1})$$

είναι μια υπολογίσιμη εκτίμηση του $e_{j+1}(h)$. Αν το δ_{j+1} δείχνει ότι η τιμή c_{j+1} δεν είναι αρκετά ακριβής τότε υπολογίζουμε ξανά το c_{j+1} χρησιμοποιώντας την τιμή $f(t_{j+1}, c_{j+1})$ στον τύπο (3.22) σαν μια ακριβέστερη εκτίμηση του f_{j+1}. Η μέθοδος πρόβλεψης-διόρθωσης που δίνεται από τους τύπους (3.21) και (3.22) λέγεται μέθοδος Adams τέταρτης τάξης (ή Adams-Bashforth ή Adams-Moulton), συντομογραφικά APC4. Η γεωμετρική ερμηνεία της μεθόδου αυτής φαίνεται στο παρακάτω σχήμα



(α) Adams Predictor



(β) Adams Corrector

Η καμπύλη που φαίνεται στο (α) είναι το πολυώνυμο παρεμβολής στα $P_{j-3}, P_{j-2}, P_{j-1}, P_j$ ενώ αυτή που έχει σχεδιαστεί στο (β) είναι το πολυώνυμο παρεμβολής στα $P_{j-2}, P_{j-1}, P_j, P_{j+1}$, όπου $P_i = (t_i, f_i)$.

Παράδειγμα 3.3. Χρησιμοποιήστε την APC4 με $h = 0.1$ για την επίλυση του Π.Α.Τ.

$$\frac{dy}{dt} = -ty^2, \quad y(2) = 1$$

στο $[2, 3]$ με ακρίβεια τεσσάρων σημαντικών ψηφίων. Η ακριβής λύση είναι $y(t) = \dfrac{2}{t^2 - 2}$.

Για να ξεκινήσει η μέθοδος χρησιμοποιούμε τις ακόλουθες ακριβείς τιμές (με 7 σημαντικά ψηφία):

$$\begin{aligned}
t_0 = 2.0 \;&:\; y_0 = y(2.0) = 1.0000000, \quad f_0 = -t_0 y_0^2 = -2.0000000 \\
t_1 = 2.1 \;&:\; y_1 = y(2.1) = 0.8298755, \quad f_1 = -t_1 y_1^2 = -1.4462560 \\
t_2 = 2.2 \;&:\; y_2 = y(2.2) = 1.7042254, \quad f_2 = -t_2 y_2^2 = -1.0910530 \\
t_3 = 2.3 \;&:\; y_3 = y(2.3) = 1.6079027, \quad f_3 = -t_3 y_3^2 = -2.8499552
\end{aligned}$$

- Για $j = 3$ είναι

$$\begin{aligned}
p_4 &= y_3 + \frac{h}{24}\Big(-9f_0 + 37f_1 - 59f_2 + 55f_3 \Big) = 0.5333741 \\
y_4 &= y_3 + \frac{h}{24}\Big(f_1 - 5f_2 + 19f_3 + 9(-t_4 p_4^2) \Big) = 0.5317149 \quad (= c_4) \\
\delta_4 &= -\frac{19}{720}(y_4 - p_4) = 0.0001144
\end{aligned}$$

Επειδή το δ_4 δείχνει πιθανή ανακρίβεια στο τέταρτο δεκαδικό ψηφίο του y_4, παίρνουμε το c_4 σαν βελτιωμένη τιμή του p_4 για να προκύψει η βελτιωμένη τιμή του y_4:

$$y_4 = y_3 + \frac{h}{24}\Big(f_1 - 5f_2 + 19f_3 + 9(-t_4(0.5317149)^2) \Big) = 0.5318739$$

Η εκτίμηση του τοπικού σφάλματος αποκοπής στην τιμή y_4 είναι

$$\delta_4 = -\frac{19}{720}(0.5318739 - 0.5317149) = 0.0000112$$

που δείχνει ότι έχουμε πετύχει την επιθυμητή ακρίβεια 4 ψηφίων.

- Για $j = 4$ είναι

$$\begin{aligned} f_5 &= f(t_4,\ y_4) = -(2.4)\cdot(0.5318739)^2 = 0.6789358 \\ p_5 &= y_4 + \frac{h}{24}\Big(-9f_1 + 37f_2 - 59f_3 + 55f_4\Big) = 0.4712642 \\ y_5 &= y_4 + \frac{h}{24}\Big(f_2 - 5f_3 + 19f_4 + 9(-t_5 p_5^2)\Big) = 0.4704654 \quad (= c_5) \\ \delta_5 &= -\frac{19}{720}(y_5 - p_5) = 0.0000562 \end{aligned}$$

Όπως προηγουμένως το δ_5 δείχνει πιθανή ανακρίβεια στο πέμπτο δεκαδικό ψηφίο του y_5, άρα επιτυγχάνεται ακρίβεια τεσσάρων ψηφίων. Αν παρόλα αυτά βελτιώσουμε το y_5 ως εξής:

$$y_5 = y_4 + \frac{h}{24}\Big(f_2 - 5f_3 + 19f_4 + 9(-t_5(0.4704654)^2)\Big) = 0.4705358$$

Η εκτίμηση του τοπικού σφάλματος αποκοπής στην τιμή y_5 είναι

$$\delta_5 = -\frac{19}{720}(0.4705358 - 0.4704654) = 0.0000050$$

και δείχνει ότι η νέα τιμή έχει ακρίβεια περίπου 5 ψηφίων.

Παρατήρηση 3.1. Έχει αποδειχθεί ότι οι τιμές που προκύπτουν εκτελώντας το πολύ μια διόρθωση του y_{j+1} είναι πιθανό να είναι της ίδιας ακρίβειας με εκείνες που προκύπτουν από τη χρήση μιας γενικής στρατηγικής με επανάληψη του τύπου διόρθωσης. Αν χρειάζονται περισσότερες από μια διόρθωση τότε ελαττώνουμε το πλάτος h.

Παρατήρηση 3.2. Όπως και στη μέθοδο RK4 έτσι και στην APC4 μπορούμε να χρησιμοποιήσουμε το δ_{j+1} ως κριτήριο για να μεταβάλλουμε το h. Όταν το $\frac{|\delta_{i+1}|}{h}$ είναι πολύ μεγάλο ελαττώνουμε το h, ενώ όταν είναι πολύ μικρό αυξάνουμε το πλάτος h.

3.8 Σύγκριση των μεθόδων RK και PC

Αν η μοναδικότητα της $y(t)$ είναι δυνατή στο $[t_0, t_n]$ τότε συνίσταται μια μέθοδος με κατάλληλο έλεγχο του βήματος, όπως η RKF4. Αν η $f(t, y)$ είναι δαπανηρή στο να εκτιμηθεί, τότε πρέπει να προτιμηθεί η APC4 (2-3 εκτιμήσεις ανά βήμα).

Τέλος σε μια εφαρμογή πραγματικού χρόνου (real time), όπου μια ποσότητα y_{j+1} πρέπει να υπολογισθεί σε κάποιο σταθερό σύντομο χρονικό διάστημα, πρέπει να προτιμηθεί η μέθοδος του Euler και κατά προτίμηση η τροποποιημένη μέθοδος Euler (εφόσον ο χρόνος το επιτρέπει).

Σχετικά με τις μεθόδους PC αξίζει να αναφερθεί ότι επειδή το δ_{j+1} είναι μια εκτίμηση του $e_{j+1}(h)_c$, μπορεί να προστεθεί στο c_{j+1} για να δώσει μια βελτιωμένη διόρθωση $c_{j+1} + \delta_{j+1}$. Όμως αυτή η στρατηγική είναι παραπλανητική. Εμπειρικά τεστ αποδεικνύουν ότι οι μέθοδοι που την ακολουθούν τείνουν να είναι ασταθείς και επομένως γενικά δεν προτιμούνται.

Ο παρακάτω πίνακας συνοψίζει τα κύρια χαρακτηριστικά κάθε μεθόδου:

Μέθοδος	Self Starting	Τοπικό σφάλμα	Ολικό σφάλμα	Υπολογισμοί της f ανά βήμα	Έλεγχος μεγέθους βήματος
Euler	Ναι	$O(h^2)$	$O(h)$	1	δυνατός
Τροπ. Euler	Ναι	$O(h^3)$	$O(h^2)$	2	δυνατός
Huen	Ναι	$O(h^3)$	$O(h^2)$	2	δυνατός
RK4	Ναι	$O(h^5)$	$O(h^4)$	4	δυνατός
RKF4	Ναι	$O(h^5)$	$O(h^4)$	6	εύκολος
APC4	Όχι	$O(h^5)$	$O(h^4)$	2-3	εύκολος

3.9 Συστήματα διαφορικών εξισώσεων και Π.Α.Τ. $n-$τάξης

Γενικά τα φυσικά συστήματα χρειάζονται μερικές μεταβλητές για να περιγράψουν την κατάστασή τους σε κάθε χρονική στιγμή t. Για παράδειγμα, η κατάσταση μεταβλητών ενός θερμοδυναμικού συστήματος περιλαμβάνει θερμοκρασία, πίεση, όγκο και εντροπία. Επίσης η κατάσταση μεταβλητών ενός μηχανικού συστήματος περιλαμβάνει τις μετατοπίσεις ορισμένων σημείων, ή η κατάσταση των μεταβλητών ενός ηλεκτρικού κυκλώματος περιλαμβάνει τις τάσεις και τις εντάσεις που προσδιορίζουν την ενέργεια που κατανέμεται σε διάφορα ηλεκτρικά όργανα.

3.9.1 Συμβολισμός και ορολογία

Η κατάσταση μεταβλητών ενός συστήματος συμβολίζεται με $y_1, y_2, \ldots, y_n$ όπου καθεμία μεταβάλλεται ως προς μια απλή μεταβλητή t (συνήθως χρόνος) σύμφωνα με κάποιο φυσικό νόμο. Συνήθως αυτοί οι νόμοι παίρνουν τη μορφή συστήματος n το πλήθος Π.Α.Τ. πρώτης

τάξης της μορφής:

$$\begin{aligned} y_1' = \frac{dy_1}{dt} &= f_1(t, y_1, y_2, \ldots, y_n), \quad y_1(t_0) = y_{01} \\ y_2' = \frac{dy_2}{dt} &= f_2(t, y_1, y_2, \ldots, y_n), \quad y_2(t_0) = y_{02} \\ &\vdots \\ y_n' = \frac{dy_n}{dt} &= f_n(t, y_1, y_2, \ldots, y_n), \quad y_n(t_0) = y_{0n} \end{aligned} \tag{3.24}$$

Δηλαδή, γνωρίζουμε την αρχική κατάσταση σε κάποιο t_0 και γνωρίζουμε το ρυθμό με τον οποίο μεταβάλλεται κάθε μεταβλητή όταν το σύστημα βρίσκεται σε μια συγκεκριμένη κατάσταση.

Παράδειγμα 3.4. Έστω

$$\begin{aligned} y_1' &= 3ty_2 + y_1y_2, y_1(t_0) = -1 \\ y_2' &= 2\ln(1+y_2^2) + \frac{1}{y_3}, y_2(t_0) = 1 \\ y_3' &= y_3^2 - y_1e^{y_2} + 1, y_3(t_0) = 1 \end{aligned}$$

Παρατηρούμε ότι οι ρυθμοί μεταβολής των y_1, y_2 και y_3 εξαρτώνται και από τις τρεις μεταβλητές ως προς t. Για αυτό το λόγο συχνά αναφέρεται ως ένα σύστημα ζεύξης πρώτης τάξης εξισώσεων.

Μια λύση του Π.Α.Τ. στην (3.24) αποτελείται από n συναρτήσεις

$$y_1 = y_1(t) \quad , \quad y_2 = y_2(t) \quad , \quad \ldots \quad , \quad y_n = y_n(t)$$

που ικανοποιούν το Π.Α.Τ., δηλαδή για $i = 1, 2, \ldots, n$ ισχύουν

$$y_i'(t) = f_i\Big(t, y_1(t), y_2(t), \ldots, y_n(t)\Big) \quad \text{και} \quad y_i(t_0) = y_{0i}$$

3.9.2 Τύποι υπό διανυσματική μορφή των αριθμητικών μεθόδων επίλυσης του Π.Α.Τ. $n-$τάξης

Αν θεωρήσουμε το διάνυσμα κατάστασης των μεταβλητών $\mathbf{y} = \mathbf{y}(t) = (y_1, y_2, \ldots, y_n)^T$ τότε το Π.Α.Τ. στην (3.24) γράφεται υπό την διανυσματική μορφή

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \tag{3.25}$$

όπου

$$\mathbf{y}' = \begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix} \quad , \quad \mathbf{f}(t, \mathbf{y}) = \begin{bmatrix} f_1(t, \mathbf{y}) \\ f_2(t, \mathbf{y}) \\ \vdots \\ f_n(t, \mathbf{y}) \end{bmatrix} \quad \text{και} \quad \mathbf{y}_0 = \begin{bmatrix} y_{01} \\ y_{02} \\ \vdots \\ y_{0n} \end{bmatrix}$$

Μια αριθμητική λύση του (3.25) στο διάστημα $[t_0, t_n]$ είναι μια ακολουθία διανυσμάτων $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_n$ όπου το $\mathbf{y}_j$ προσεγγίζει την ακριβή λύση $\mathbf{y}(t_j)$ για $j = 0(1)n$.

Μπορούμε εύκολα να παρατηρήσουμε ότι από ένα Π.Α.Τ. πρώτης τάξης προκύπτει ένα Π.Α.Τ. $n-$τάξης της μορφής (3.25) αν αντικαταστήσουμε τα $y, y', f(t,y)$ και y_0 με τα διανύσματα $\mathbf{y}, \mathbf{y}', \mathbf{f}(t,\mathbf{y})$ και $\mathbf{y}_0$. Αν εκτελέσουμε αυτές τις αντικαταστάσεις στον τύπο μιας λύσης ενός Π.Α.Τ. προκύπτει ο τύπος λύσης του Π.Α.Τ. στην (3.24). Ειδικά από την RK4 μέθοδο ενός Π.Α.Τ. προκύπτει η αντίστοιχη μέθοδος :

$$\mathbf{y}_{j+1} = \mathbf{y}_j + \frac{h}{6}\Big(m_1 + 2(m_2 + m_3) + m_4\Big)$$

όπου m_1, m_2, m_3, m_4 τα διανύσματα κλίσεων που προκύπτουν αναγωγικά από τους τύπους

$$\begin{aligned}
\mathbf{m}_1 &= \mathbf{f}(t_j, \mathbf{y}_j) \\
\mathbf{m}_2 &= \mathbf{f}\Big(t_j + \frac{h}{2}, \mathbf{y}_j + \frac{h}{2}\mathbf{m}_1\Big) \\
\mathbf{m}_3 &= \mathbf{f}\Big(t_j + \frac{h}{2}, \mathbf{y}_j + \frac{h}{2}\mathbf{m}_2\Big) \\
\mathbf{m}_4 &= \mathbf{f}\Big(t_j + h, y_j + h\mathbf{m}_3\Big)
\end{aligned}$$

3.9.3 Επίλυση ενός $n-$τάξης Π.Α.Τ.

Το $n-$τάξης Π.Α.Τ. έχει τη μορφή

$$y^{(n)} = f(t, y, y', y'', \ldots, y^{(n-1)}) \tag{3.26}$$

κάτω από n αρχικές συνθήκες $y(t_0) = y_0,\ y'(t_0) = y'_0, \ldots, y^{(n-1)}(t_0) = y_0^{(n-1)}$.

Η γενική μέθοδος για την επίλυση του $n-$τάξης Π.Α.Τ. είναι να το μετατρέψουμε σε ένα ισοδύναμο σύστημα από $n-$ τάξης εξισώσεις. Για να το πετύχουμε αυτό θεωρούμε

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} \quad \text{όπου} \quad \begin{aligned} y_1 &= y \\ y_2 &= y' \\ &\vdots \\ y_{n-1} &= y^{(n-2)} \\ y_n &= y^{(n-1)} \end{aligned} \tag{3.27}$$

Αν αντικαταστήσουμε τις (3.27) στην (3.26) και συμβολίζοντας με $y' = \dfrac{dy^{(n-1)}}{dt} = y^{(n)}$ προ-

κύπτει το ισοδύναμο σύστημα από n ΠΑΤ πρώτης τάξης:

$$\begin{array}{ll} y_1' = y_2 & y_1(t_0) = y_0 \\ y_2' = y_3 & y_2(t_0) = y_0' \\ \vdots \qquad \text{όπου} & \vdots \\ y_{n-1}' = y_n & y_{n-1}(t_0) = y_0^{(n-2)} \\ y_n' = f(t, y_1, y_2, \ldots, y_n) & y_n(t_0) = y_0^{(n-1)} \end{array}$$

Αυτό τώρα είναι της μορφής (3.25) και μπορεί να λυθεί με μια από τις προηγούμενα αναφερθείσες μεθόδους.

3.10 Προβλήματα Συνοριακών Τιμών (Π.Σ.Τ.)

Θα εξετάσουμε το πρόβλημα της λύσης της $n-$τάξης διαφορικής εξίσωσης

$$y^{(n)} = f(t, y, y', y'', \ldots, y^{(n-1)}) \tag{3.28}$$

με τους n περιορισμούς της y και/ή των παραγώγων της μορφής

$$y^{(k_1)}(t_1) = \beta_1,\ y^{(k_2)}(t_2) = \beta_2,\ \ldots\ ,\ y^{(k_n)}(t_n) = \beta_n$$

- Αν $t_1 = t_2 = \cdots = t_n = t_0$ τότε ανάγεται σε ένα $n-$τάξης Π.Α.Τ.
- Αν οι συνοριακές συνθήκες περιέχουν m διακεκριμένα t_i, όπου $m > 1$ τότε το (3.28) είναι ένα $n-$τάξης Π.Σ.Τ. $m-$σημείων.

Θα μελετήσουμε το δεύτερης τάξης Π.Σ.Τ. δυο σημείων:

$$y'' = f(t, y, y'), \quad y(a) = \alpha,\ \ y(b) = \beta \tag{3.29}$$

3.10.1 Μέθοδος της βολής (ή σκόπευσης)(shooting)

Θεωρούμε το ΠΑΤ δεύτερης τάξης

$$(\Pi\text{ΑΤ})_x\ :\quad y'' = f(t, y, y'), \quad y(a) = \alpha,\ \ y'(a) = x$$

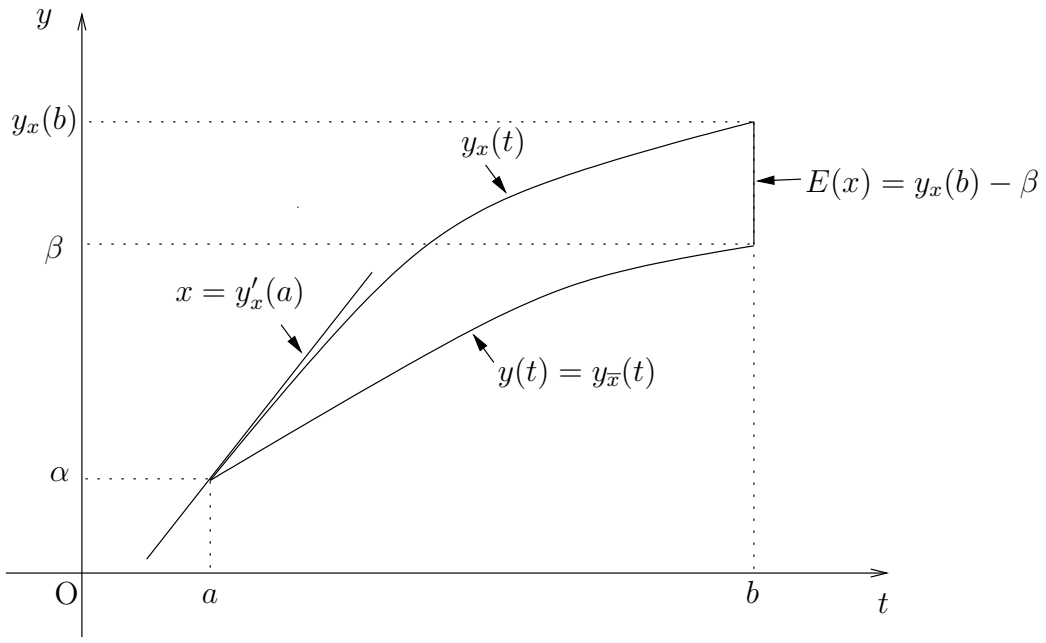
Έστω $y_x(t)$ η λύση του. Τότε, η τιμή x είναι η αρχική κλίση (στο $t = a$) της $y_x(t)$. Ζητείται να βρεθεί τιμή $\overline{x}$ για την οποία η $y_{\overline{x}}(b) = \beta$ είναι η συνοριακή συνθήκη στο $t = b$.

Αν η καμπύλη λύση $y_{\overline{x}}(t)$ είναι μια τροχιά βολής όπου η συνθήκη $y_x(b) = \beta$ αντιστοιχεί στην τροχιά βολής της εφαπτομένης τιμής β στο $t = b$, τότε η $y_{\overline{x}}(t)$ ικανοποιεί το Π.Σ.Τ. (3.29) και επομένως είναι η επιθυμητή $y(t)$.

Για να υλοποιήσουμε τη μέθοδο βολής εισάγουμε τη συνάρτηση σφάλματος

$$E(x) = y_x(b) - \beta$$

που είναι η ποσότητα κατά την οποία η τιμή $y_x(b)$ διαφέρει από την εφαπτόμενη τιμή β στο $t = b$.



Έτσι το πρόβλημα της λύσης του Π.Σ.Τ. (3.29) ανάγεται στην εύρεση μιας ρίζας $\overline{x}$ της $E(x)$, δηλαδή στη λύση της $E(x) = 0$. Εφόσον κάθε υπολογισμός της $E(x)$ απαιτεί σημαντική υπολογιστική εργασία, δηλαδή ολοκλήρωση του (Π.Α.Τ.)$_x$ στο $[t_0, t_n] = [a, b]$ είναι ανάγκη να χρησιμοποιηθεί μια μέθοδος που να συγκλίνει γρήγορα στην επιθυμητή ρίζα της $E(x)$.

Η μέθοδος Newton-Raphson (N-R) είναι η ταχύτερη σε σύγκλιση αλλά δεν είναι κατάλληλη για αυτήν την εφαρμογή, διότι γενικά η $E(x)$ δεν έχει αναλυτική έκφραση. Η πλέον κατάλληλη μέθοδος σε αυτήν την περίπτωση είναι η μέθοδος της Τέμνουσας:

$$x_{k+1} = x_k - \frac{E(x_k) \cdot (x_k - x_{k-1})}{E(x_k) - E(x_{k-1})}, \quad k = 0, 1, 2, \cdots$$

Παράδειγμα 3.5. Χρησιμοποιήστε τη μέθοδο βολής (shooting) για την επίλυση του μη γραμμικού δεύτερης τάξης Π.Σ.Τ. δυο σημείων

$$y'' = y' \left(\frac{1}{t} + \frac{2y'}{y} \right) \quad , y(1) = 4, \ y(2) = 8,$$

με ακρίβεια 5 σημαντικών ψηφίων.

Αναζητούμε μια ρίζα $\overline{x}$ της $E(x) = y_x(b) - \beta = y_x(2) - 8$ όπου $y_x(t)$ είναι η λύση του αντίστοιχου Π.Α.Τ.

$$(\Pi\text{ΑΤ})_x \ : \quad y'' = y'\left(\frac{1}{t} + \frac{2y'}{y}\right), \quad y(1) = 4, \quad y'(1) = x$$

Μια φυσική υποψήφια τιμή για μια αρχική τιμή του $y'(1)$ είναι η κλίση της ευθείας που συνδέει τα σημεία (a, α) και (b, β):

$$x = \frac{\beta - \alpha}{b - a} = \frac{8 - 4}{2 - 1} = 4$$

Μια προσπάθεια για τη λύση του $(\Pi\text{ΑΤ})_x$ με $x = 4$ δείχνει ότι η $y_x(t)$ έχει μια κανονικότητα μεταξύ $a = 1$ και $b = 2$. Συνεπώς σκοπεύουμε χαμηλότερα, δηλαδή παίρνουμε το μισό της κλίσης $(x = 2)$ σαν αρχική τιμή. Τότε

$$x = 2 \implies y_x(2) = 16 \implies E(x) = y_x(2) - 8 = 8$$

Επειδή η μέθοδος της τέμνουσας απαιτεί δυο αρχικές τιμές και $E(2) > 0$ παίρνουμε μια μικρότερη κλίση, τη $x = 1$. Χρησιμοποιώντας έναν ακριβή λύτη $n-$τάξεως για τη λύση του $(\Pi.\text{Α.Τ.})_x$ με $x = 1$ προκύπτει ότι

$$x = 1 \implies y_x(2) = 6.4 \implies E(x) = y_x(2) - 8 = -1.6$$

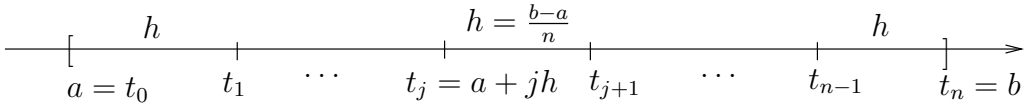
Μπορούμε τώρα να εφαρμόσουμε την επαναληπτική μέθοδο της Τέμνουσας παίρνοντας ως αρχικές τιμές $x_{-1} = 2,\ x_0 = 1$ και

$$x_{k+1} = x_k + \Delta x_k \qquad \text{όπου} \qquad \Delta x_k = \frac{E(x_k) \cdot (x_k - x_{k-1})}{E(x_k) - E(x_{k-1})}, \qquad k = 0, 1, 2, \ldots$$

k	$x_k \equiv x$	$y_x(2)$	$E(x) = y_x(2) - 8$	Δx_k	$x_{k+1} = x_k + \Delta x_k$
-1	2	16	8		
0	1	6.4	-1.6	0.1666667	1.1666667
1	1.6666667	7.1111111	-0.8888889	0.2083333	1.3750000
2	1.3750000	8.2580645	0.2580645	-0.0468750	1.3281250
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

3.10.2 Μέθοδος των Πεπερασμένων Διαφορών

Μια άλλη στρατηγική για την αριθμητική επίλυση του Π.Σ.Τ. στην (3.29) είναι να διαμερίσουμε το $[a,b]$ σε n υποδιαστήματα με $n+1$ ισαπέχοντα σημεία:



και να αντικαταστήσουμε τις τιμές $y'(t_j),\ y''(t_j)$ με τις προσεγγιστικές τιμές τους που προκύπτουν από τους προηγούμενους τύπους κεντρικών διαφορών τάξης $O(h^2)$:

$$y''(t_j) \cong \frac{y_{j+1}-2y_j+y_{j-1}}{h^2} \qquad \text{και} \qquad y'(t_j) = \frac{y_{j+1}-y_{j-1}}{2h}$$

για $j = 1, 2, \ldots, n-1$. Αυτή η στρατηγική λέγεται μέθοδος *των πεπερασμένων διαφορών*.

Έτσι το αναλυτικό πρόβλημα της λύσης του Π.Σ.Τ. στην (3.29) ανάγεται στη λύση ενός προσεγγιστικού αλγεβρικού προβλήματος $n-1$ εξισώσεων με $n-1$ αγνώστους $y_1, y_2, \ldots, y_{n-1}$ όπου $y_j \cong y(t_j),\ j = 1, 2, \ldots, n-1$ και τα $y_0 = y(a) = \alpha$ και $y_n = y(b) = \beta$ είναι γνωστά.

Παράδειγμα 3.6. Χρησιμοποιήστε τη μέθοδο των πεπερασμένων διαφορών για τη λύση του μη γραμμικού Π.Σ.Τ. δεύτερης τάξης

$$y'' = y'\left(\frac{1}{t} + \frac{2y'}{y}\right), \quad y(1) = 4, \ \ y(2) = 8$$

Αν αντικαταστήσουμε τις y'' και y' με τις αντίστοιχες προσεγγίσεις κεντρικών διαφορών τάξης $O(h^2)$ στα $t_j,\ j = 1, 2, \ldots, n-1$ προκύπτει:

$$\frac{y_{j+1}-2y_j+y_{j-1}}{h^2} = \frac{y_{j+1}-y_{j-1}}{2h}\left(\frac{1}{t_j} + \frac{2(y_{j+1}-y_{j-1})}{2hy_j}\right), \quad y_0 = 4, \ \ y_n = 8$$

Πολλαπλασιάζοντας επί $2h^2t_j$ και μεταφέροντας τους γραμμικούς όρους των y_{j-1}, y_j, y_{j+1} στο αριστερό μέλος προκύπτει

$$(2t_j+h)y_{j-1} - 4t_jy_j + (2t_j-h)y_{j+1} = d_j, \qquad y_0 = 4, \ \ y_n = 8$$

και

$$d_j = \frac{t_j}{y_j}(y_{j+1}-y_{j-1})^2, \qquad j = 1, 2, \ldots, n-1$$

Η υπεροχή του συντελεστή $-4t_j$ εξασφαλίζει ότι μπορεί να λυθεί η $j-$ εξίσωση με

$$y_j^{(\text{new})} = \frac{1}{4t_j}\Big[(2t_j+h)y_{j-1} + (2t_j-h)y_{j+1} - d_j\Big], \qquad j = 1, 2, \ldots, n-1$$

και έπειτα χρησιμοποιούμε την επαναληπτική μέθοδο Gauss-Seidel.

$$y_j^{(k+1)} = \frac{1}{4t_j}\Big[(2t_j + h)y_{j-1}^{(k+1)} + (2t_j - h)y_{j+1}^{(k)} - d_j\Big], \qquad j = 1, 2, \ldots, n-1$$

όπου παίρνουμε τα $y_j^{(0)}$ να είναι επί της ευθείας της τροχιάς βολής από το σημείο (a, α) στο (b, β), δηλαδή της $y - \alpha = \dfrac{\beta - \alpha}{b - a}(t - \alpha)$, άρα είναι

$$y_j^{(0)} = \alpha + \frac{\beta - \alpha}{b - a}(t_j - \alpha), \qquad j = 0, 1, \ldots, n$$

Η προηγούμενη μέθοδος λέγεται διαδικασία Liebmann.

3.10.3 Μέθοδος των πεπερασμένων διαφορών, όταν το Π.Σ.Τ. είναι γραμμικό

Το γραμμικό ΠΣΤ δεύτερης τάξης έχει τη γενική μορφή:

$$y'' = \varphi_1(t)y + \varphi_2(t)y' + F(t), \qquad y(a) = \alpha, \quad y(b) = \beta \tag{3.30}$$

Αν αντικαταστήσουμε τις προσεγγίσεις τάξης $O(h^2)$ των $y''(t_j)$ και $y'(t_j)$ προκύπτει

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = \varphi_1(t_j)y_j + \varphi_2(t_j)\frac{y_{j+1} - y_{j-1}}{2h} + F(t_j)$$

Αν πολλαπλασιάσουμε κατά μέλη επί $2h^2$ και μεταφέρουμε τους όρους των y_{j-1}, y_j, y_{j+1} στο αριστερό μέλος για $j = 1, 2, \ldots, n-1$ τότε προκύπτει

$$\Big(\underbrace{2 + h\varphi_2(t_j)}_{a_j}\Big)y_{j-1} - \Big(\underbrace{4 + 2h^2\varphi_1(t_j)}_{b_j}\Big)y_j + \Big(\underbrace{2 - h\varphi_2(t_j)}_{c_j}\Big)y_{j+1} = \underbrace{2h^2F(t_j)}_{d_j}$$

εφόσον $y_0 = y(t_0) = y(a) = \alpha$ και $y_n = y(t_n) = y(b) = \beta$ έχουμε το τριδιαγώνιο σύστημα

$$\begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-2} & b_{n-2} & c_{n-2} \\ & & & a_{n-1} & b_{n-1} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} d_1 - a_1\alpha \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} - c_{n-1}\beta \end{bmatrix}$$

όπου $y_j \simeq y(t_j) = y(a + jh), \quad j = 0, 1, 2, \cdots n-1.$

3.10.4 Σύγκριση της μεθόδου βολής και της μεθόδου των πεπερασμένων διαφορών

Συνέπεια της γραμμικότητας:

- Αν το Π.Σ.Τ. δεύτερης τάξης είναι γραμμικό, τότε η μέθοδος βολής απαιτεί μόνο δυο ολοκληρώσεις του (Π.Α.Τ.)$_x$ στο $[a, b)$, ενώ η μέθοδος πεπερασμένων διαφορών δίνει ένα γραμμικό σύστημα με μια διαγώνια υπεροχή και ΄πίνακα συντελεστών μορφής δέσμης που μπορεί να λυθεί σύντομα και ακριβώς με τη μέθοδο απαλοιφής του Gauss για αρκετά μεγάλο n.

- Αν το Π.Σ.Τ. δεν είναι γραμμικό, η μέθοδος βολής οδηγεί σε μια επαναληπτική μέθοδο (δηλαδή την επαναληπτική μέθοδο της τέμνουσας), ενώ η μέθοδος των πεπερασμένων διαφορών δίνει ένα μη γραμμικό σύστημα που μπορεί να λυθεί με την επαναληπτική μέθοδο Gauss-Seidel(διαδικασία Liebmann).

Αίτια σφάλματος:

- Όταν το Π.Σ.Τ.είναι γραμμικό ή μη γραμμικό, η λύση που προκύπτει με τη μέθοδο πεπερασμένων διαφορών περιέχει το σφάλμα αποκοπής των προσεγγίσεων πεπερασμένων διαφορών των y'' και y'. Η ακρίβεια εκτιμάται με σύγκριση των υπολογισθεισών τιμών y_j με εκείνες που προκύπτουν με διπλάσιο n.

- Η ακρίβεια που επιτυγχάνεται με τη μέθοδο της βολής περιορίζεται στην ακρίβεια των υπολογισθεισών τιμών $y_x(b)$.