

Улсын бүртгэлийн
дугаар

Нууцлалын зэрэглэл:
Энгийн

Аравтын бүрэн
ангиллын код

Төсөл гүйцэтгэх гэрээний
дугаар: SST_18/2018

**МОНГОЛ УЛСЫН ШИНЖЛЭХ УХААНЫ АКАДЕМИ
МАТЕМАТИК, ТООН ТЕХНОЛОГИЙН ХҮРЭЭЛЭН**

**СПЛАЙН ФУНКЦ БОЛОН ИТЕРАЦИЙН
АРГЫН ОНОЛ, ХЭРЭГЛЭЭ**

**Суурь судалгааны төслийн тайлан
2018-2020**

Төслийн удирдагч :

**Рэнчин-Очирын Мижиддорж
Монгол Улсын Боловсролын Их
Сургууль;
Математик, Тоон Технологийн
Хүрээлэн**

Санхүүжүүлэгч байгууллага:

Шинжлэх Ухаан, Технологийн Сан

Захиалагч байгууллага:

**Боловсрол, Шинжлэх Ухааны
Яам**

Тайлан өмчлөгч:

**Шинжлэх ухааны академийн
нэгдсэн II байр, Энхтайваны
өргөн чөлөө 54Б, Баянзүрх
Дүүрэг, Улаанбаатар 13330,
Монгол Улс**

Улаанбаатар хот

2021 он

Монгол Улсын Шинжлэх Ухааны Академи
Математик, Тоон Технологийн Хүрээлэн

Слайн функц болон итерацийн аргын онол, хэрэглээ (2018-2020)

Улаанбаатар хот 2021 он

РЕФЕРАТ

Сэдэвт ажлыг

- Шугаман бус тэгшитгэл, түүний системийг бодох Ньютоны төрлийн аргуудын нийлэлт
- Интегро сплайн байгуулах, түүнийг хэрэглэх
- Шредингерийн тэгшитгэлийн шийдийн тоон ба чанарын судалгааны

чиглэлүүдийн хүрээнд гүйцэтгэв. Дурдсан чиглэлүүдэд шийдвэрлэгдээгүй зарим асуудлыг шийдэх, шинээр дэвшүүлэх, шинэ арга, алгоритмууд байгуулах, тооцоо хийж онолын үр дүн, дүгнэлтүүдийг баталгаажуулах, олон улсын түвшинд судалгаа явуулж дорвитой үр дүнд хүрэх зорилттой байсан. Мөн өмнөх жилүүдэд хийсэн судалгааны зарим үр дүнг цаашид сайжруулах, өргөтгөх шаардлага байсаар байна.

Дээрх чиглэлүүдээр олон улсын нэр хүнд бүхий мэргэжлийн сэтгүүлүүдэд эрдэмтдийн судалгааны ажлууд тогтмол хэвлэгдэж өрсөлдөөн ид явагдаж байгаа ба манайд ч судалгааны алхмууд хийгдсэн, цаашид тодорхой үр дүнд хүрэх найдлагатай, гүнзгийрүүлэн судлах зайлшгүй шаардлагатай, өрсөлдөөнд оролцож тухайн чиглэлүүдэд өөрсдийн хувь нэмрээ оруулах боломжтой гэж үзсэнийг дурдах нь зүйтэй. Сэдвийн дагуу хийсэн судалгааны ажлын үр дүн нь математикийн шинжлэх ухаанд шинэ мэдээлэл, онол, арга дүгнэлт болох бөгөөд мэргэжлийн сэтгүүлүүдэд хэвлэгдэж нийтийн хүртээл болгох нь ажлын онолын болон практик ач холбогдлыг харуулж байна. Сэдвийн хүрээнд эрдэм шинжилгээний өгүүлэл 14 (үүнээс 8 нь Web of Science-н импакт фактор индекстэй сэтгүүлд хэвлэгдсэн), илтгэл 12 (10 нь олон улсын хуралд илтгэгдсэн), монограф нэг хэвлэгдсэн.

Хэвлүүлсэн бүтээлүүдийн гол үр дүн:

1. Дэвшүүлсэн схем, алгоритм	3
2. Баталсан лемм, өгүүлбэр	1
3. Баталсан теорем	27
4. Докторын зэрэг хамгаалсан диссертац	1

Төслийн эцсийн үр дүнг “шугаман бус тэгшитгэл, түүний системийг бодох”, “интегро сплайн байгуулах, түүнийг хэрэглэх”, “Шредингерийн тэгшитгэлийн шийдийн тоон ба чанарын судалгаа” гэсэн бүтцээр тайлагнаж байна.

Түлхүүр үг: Итерацийн арга, шугаман бус тэгшитгэлийг ойролцоо бодох аргууд, Шредингерийн тэгшитгэлийн тоон шийд, сплайн дөхөлт.

Гүйцэтгэгчид:

- | | | | |
|---------------------|-------------|----------------|--|
| 1. Р. Мижиддорж, | удирдагч, | доктор (Ph.D.) | Монгол Улсын
Боловсролын Их Сургууль;
Математик, Тоон
Технологийн Хүрээлэн; |
| 2. Т. Жанлав, | гүйцэтгэгч, | доктор (Sc.D.) | Математик, Тоон
Технологийн Хүрээлэн |
| 3. О. Чулуунбаатар, | гүйцэтгэгч, | доктор (Sc.D.) | Дубна хот дахь Цөмийн
Шинжилгээний Нэгдсэн Институт;
Математик, Тоон
Технологийн Хүрээлэн |
| 4. Х. Отгондорж, | гүйцэтгэгч, | доктор (Ph.D.) | Монгол Улсын Шинжлэх Ухаан
Технологийн Их Сургууль;
Математик, Тоон
Технологийн Хүрээлэн |

Гарчиг

РЕФЕРАТ	3
Үр дүнгийн тойм	6
Удиртгал	6
1 Шугаман бус тэгшитгэл, түүний системийг бодох	7
1.1 Уламжлалгүй, хоёр алхамт итерацийн арга	7
1.2 Уламжлалгүй, гурван алхамт итерацийн арга	9
1.3 Уламжлалгүй оновчтой аргуудын харьцуулалт	10
1.4 Уламжлалгүй аргын динамик төлөвийн судалгаа	12
1.5 Уламжлалгүй наймдугаар эрэмбийн оновчтой аргуудыг байгуулах	13
1.6 Хоёр алхамт итерацийн арга дахь параметрийн оновчтой сонголт	16
1.7 Оновчтой зургаа болон долоодугаар эрэмбийн арга	16
1.8 Санах ойтой долоодугаар эрэмбийн арга	17
1.9 Шугаман биш тэгшитгэлийн систем бодох дээд эрэмбийн арга	18
1.10 Хоёр алхамт арга	18
1.11 Гурван алхамт арга	20
2 Интегро сплайн байгуулах, түүнийг хэрэглэх	22
2.1 Локаль интегро сплайны чанар болон хэрэглээ	22
2.1.1 Локаль интегро куб сплайны байгуулалт	25
2.1.2 Алдааны шинжилгээ болон хэлбэр хадгалах чанар	27
2.2 Интегро сплайны харьцуулалт, бусад чанар	32
3 Шредингерийн тэгшитгэлийн шийдийн тоон ба чанарын судалгаа	34
Хэвлүүлсэн өгүүллүүд	38
Ашигласан материалын жагсаалт	41
Хавсралтууд	43

Үр дүнгийн тойм

Сэдэвт ажлын хүрээнд гарсан үр дүнг товч тоймловол.

Судалгааны ажлын үндэслэл:

Шугаман биш тэгшитгэл ба шугаман биш тэгшитгэлийн системийг бодох өндөр эрэмбийн итерацийн аргуудыг байгуулах нь тооцон бодох математик болон шинжлэх ухаан, инженерийн хэрэглээнд чухал ач холбогдолтой. Шугаман биш тэгшитгэл, түүний системийн шийд нь аналитик байдлаар олдох нь маш ховор учраас шийдийг ямарваа аргаар ойролцоогоор олох асуудал тавигддаг. Өөрөөр хэлбэл шугаман биш тэгшитгэлийн шийдийг итерацийн аргаар өндөр нарийвчлалттай ойролцоогоор бодох шаардлага урган гардаг. Энэ зорилгоор хамгийн өргөн хэрэглэдэг аргууд нь Ньютон ба Стефенсоны аргууд бөгөөд эдгээр арга нь квадрат хурдтайгаар локаль нийлдэг онцлогтой. Энэхүү сэдэвт ажлын хүрээнд Ньютон ба Стефенсон төрлийн аргуудыг судалж, нийлэлтийн эрэмбийг сайжруулж өндөр эрэмбийн нийлэлттэй аргууд байгуулахыг зорьсон.

Судалгааны шинэлэг тал:

Сэдэвт ажлын шинэлэг талыг дараах байдлаар тодорхойлж байна. Үүнд:

- Шугаман биш тэгшитгэлийг бодох уламжлалгүй хоёр ба гурван алхамт аргуудын нийлэх нөхцөлүүдийг байгуулж, холбогдох теоремуудыг баталсан ба нийлэлтийн хурдыг тогтоосон.
- Уламжлалгүй өндөр эрэмбийн аргуудын динамик төлөвийн судалгаа хийж, график харьцуулалт хийсэн.
- Хоёр алхамт аргууд дахь итерацийн параметруудийн оновчтой сонголтуудыг символ тооцоолол ашиглахгүй аналитик аргаар гаргаж авсан.
- Хоёр ба гурван алхамт аргуудын нийлэх зайлшгүй ба хүрэлцээтэй нөхцөлүүдийг шугаман биш тэгшитгэлийн систем рүү өргөтгөж холбогдох теоремуудыг баталсан.
- Шугаман биш тэгшитгэлийн систем бодох өндөр эрэмбийн аргуудыг байгуулж, дэвшүүлсэн аргуудын итерац тутамд гүйцэтгэх нийт үйлдлийн тоо хамгийн бага байх параметрийн сонголтыг байгуулсан.
- Жигд бус тор дээр интегро куб сплайн байгуулсан.
- Байгуулсан интегро куб сплайны хувьд алдааны болон гүдгэр чанарын шинжилгээ хийсэн.

- Сплайныг локалиар байгуулсан тул роботын байрлалыг бодит цаг хугацаанд тооцоолох боломжтой болсон.

Судалгааны ажлын ач холбогдол:

Төслийн хүрээнд шугаман биш тэгшитгэлийг бодох уламжлалгүй өндөр эрэмбийн аргуудыг байгуулсан. Тэгшитгэлийн шийдийн орчинд функцийн уламжлал оршихгүй эсвэл уламжлалыг хэрэглэхэд хүндрэлтэй үед энэ арга нь шийдийг өндөр нарийвчлалтай олох боломж олгодгоороо практик ач холбогдолтой болсон.

Скаляр тохиолдолд баталсан хоёр ба гурван алхамт итерацийн аргын нийлэлтийн нөхцөлүүдийг шугаман биш тэгшитгэлийн системийн тохиолдолд өргөтгөсөн. Эдгээр нөхцөлүүдийг ашиглаж хоёр ба гурван алхамт итерацийн арга өндөр эрэмбийн нийлэлттэй байх параметрийн утгуудыг байгуулсан.

Практикаас урган гарсан олон бодлого дээр параметрийн янз бүрийн утгад тоон туршилт хийж улмаар бусад ижил эрэмбийн аргуудтай харьцуулалт хийн, шийдийг өндөр нарийвчлалтай ба бодолтын хугацаа бага зарцуулах аргуудыг дэвшүүлсэн бөгөөд ийм арга алгоритмыг шугаман биш тэгшитгэлийн системийг бодоход үр ашигтайгаар хэрэглэж болохыг үзүүлсэнд судалгааны ажлын практик ач холбогдол оршино.

Бүлэг 1

Шугаман бус тэгшитгэл, түүний системийг бодох Ньютоны төрлийн аргуудын нийлэлт

Ньютоны төрлийн аргуудын нийлэлт, нийлэлтийн төлөвийг удирдах параметр оновчтой байх утгын мужийг байгуулах судалгааны чиглэлд нийт 9 өгүүллийг Англи, Орос хэл дээр бичиж мэргэжлийн сэтгүүлд хэвлүүлэн, эдгээр судалгааны үр дүнг гадаад дотоодын эрдэм шинжилгээний хуралд 6 удаа илтгэн танилцуулсан. Мөн энэ чиглэлд хийгдсэн өөр хоёр чухал үр дүн нь төслийн үндсэн гүйцэтгэгч академич Т. Жанлавын зохион бичсэн “New Development of Newton-Type Iterations for Solving Nonlinear Problems” сэдэвт 200 хуудас бүхийн нэг сэдэвт ном, төслийн үндсэн гүйцэтгэгч доктор Х. Отгондоржийн “Шугаман биш тэгшитгэлийн системийг бодох өндөр эрэмбийн нийлэлттэй итерацийн аргууд байгуулах” сэдэвт докторын зэрэг горилсон диссертацийн ажлууд юм.

1.1 Уламжлалгүй, хоёр алхамт итерацийн арга

Уламжлалгүй хоёр алхамт дараах итерацийг авч үзье.

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \quad (1.1a)$$

$$x_{k+1} = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi(x_k)}, \quad (1.1b)$$

энд

$$f'(x) \approx \phi(x) = \frac{f(x + \gamma f(x)) - f(x)}{\gamma f(x)}, \quad \gamma \in R, \quad (1.2)$$

γ нь тэгээс ялгаатай тогтмол ба $\bar{\tau}_k$ нь итерацийн параметр. $\phi(x) \equiv \phi(x, \gamma)$ функц нь x -с гадна γ -с хамаарах ба уламжлалын тодорхойлолтоор

$$f'(x) = \phi(x, \gamma), \quad \gamma \rightarrow 0, \quad (1.3)$$

гэж үзэж болно. $f(x) \in C^3(D)$ ба D нь $f(x) = 0$ тэгшитгэлийн шийдийг агуулах завсар байг.

Теорем 1.1. $f(x) \in C^3(D)$ ба x_0 анхны дөхөлт $f(x)$ -ийн эгэл шийд $x^* \in D$ -д хангалттай ойрхон бол (1.1) итерацийн нийлэлтийн эрэмбэ дөрөв байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь $\bar{\tau}_k$ нь дараах нөхцөлийг хангах явдал

$$\bar{\tau}_k = \frac{1}{1 - \hat{d}_k \theta_k} + O(f(x_k)^2) = 1 + \hat{d}_k \theta_k + O(f(x_k)^2). \quad (1.4)$$

(1.1) итерацийн арга нь давталт бүрдээ $f(x_k), f(y_k)$ болон $\phi(x_k)$ утгуудыг ашиглаж байгаа тул Кунг-Траубын [17] таамаглалаар оновчтой арга болно. (1.1)-ийн хоёрдугаар алхмыг доорх байдлаар бичиж болно

$$x_{k+1} = x_k - \tau_k \frac{f(x_k)}{\phi(x_k)}. \quad (1.5)$$

Энд

$$\tau_k = 1 + \bar{\tau}_k \theta_k = 1 + \theta_k + \hat{d}_k \theta_k^2 + O(f(x_k)^3). \quad (1.6)$$

Хэрэв $\gamma \rightarrow 0$ үед $\phi(x_k, \gamma) = f'(x_k)$ болох ба (1.4), (1.6) нь

$$\bar{\tau}_k = 1 + 2\theta_k + O(f(x_k)^2),$$

ба

$$\tau_k = 1 + \theta_k + 2\theta_k^2 + O(f(x_k)^3)$$

болно. Тэгвэл (1.1) итерац

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= x_k - \tau_k \frac{f(x_k)}{f'(x_k)}, \end{aligned} \quad (1.7)$$

болох ба хоёр алхамт оновчтой дөрөвдүгээр эрэмбийн арга болно [3]. Итерацийн параметрийг үүсгэгч функцийг [4] тусламжтай байгуулах нь итерацийн аргуудын шинэ бүлийг гаргах боломж олгодог.

$$H(0) = 1, \quad H'(0) = \hat{d}_k, \quad (1.8)$$

нөхцөлийг хангах $\bar{\tau}_k = H(\theta_k)$ үүсгэгч функцийг янз бүрээр авч үзэх боломжтой. Тухайлбал нэг хялбар хувилбар нь

$$H(x) = \frac{c + (H'(0)c + d)x + \omega x^2}{c + dx + bx^2}, \quad c + d + b \neq 0, \quad c, d, b, \omega \in R. \quad (1.9)$$

Энэ үүсгэгч функцийг параметруудийг дараах тохиолдлуудад авч үзье.

1. (1.9)-д $c = 1, d = \beta - 2, b = \omega = 0$ бол бид

$$H(x) = \frac{1 + (\beta - \frac{\gamma \phi_k}{1 + \gamma \phi_k})x}{1 + (\beta - 2)x}$$

үүсгэгч функцтэй болно. $\bar{\tau}_k = H(\theta_k)$ бүхий (1.1) итерац нь

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi(x_k)}, \\ x_{k+1} &= y_k - \frac{1 + (\beta - \frac{\gamma\phi_k}{1+\gamma\phi_k})\theta_k}{1 + (\beta - 2)\theta_k} \cdot \frac{f(y_k)}{\phi(x_k)}. \end{aligned} \quad (1.10)$$

$\gamma \rightarrow 0$ үед (1.10) итерац нь Кингийн арга болно. (1.10) итерацийг төгсгөлөг ялгаварт уламжлалгүй Кингийн аргын хувилбар гэж нэрлэе.

2. (1.9)-д $c = b = 1$, $d = -2$, $\omega = 0$ бол бид

$$H(x) = \frac{1 - \frac{\gamma\phi_k}{1+\gamma\phi_k}x}{(1-x)^2}$$

үүсгэгч функцтэй болно. $\bar{\tau}_k = H(\theta_k)$ бүхий (1.1) итерац нь

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi(x_k)}, \\ x_{k+1} &= y_k - \frac{1 - \frac{\gamma\phi_k}{1+\gamma\phi_k}\theta_k}{(1-\theta_k)^2} \frac{f(y_k)}{\phi(x_k)}. \end{aligned} \quad (1.11)$$

$\gamma \rightarrow 0$ үед (1.11) итерац Кунг-Траубын арга болно.

3. (1.9)-д $c = 1$, $\omega = d = -1$, $b = 0$ бол бид

$$H(x) = \frac{1 + \frac{1}{1+\gamma\phi_k}x - x^2}{1-x}.$$

үүсгэгч функцтэй болно. $\bar{\tau}_k = H(\theta_k)$ бүхий (1.1) итерац нь

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi(x_k)}, \\ x_{k+1} &= y_k - \frac{1 + \frac{1}{1+\gamma\phi_k}\theta_k - \theta_k^2}{1-\theta_k} \frac{f(y_k)}{\phi(x_k)}. \end{aligned} \quad (1.12)$$

$\gamma \rightarrow 0$ үед (1.12) итерац нь Махешварын арга болно.

1.2 Уламжлалгүй, гурван алхамт итерацийн арга

Дараах гурван алхамт аргыг авч үзье.

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi(x_k)}, \\ z_k &= y_k - \bar{\tau}_k \frac{f(y_k)}{\phi(x_k)}, \\ x_{k+1} &= z_k - \alpha_k \frac{f(z_k)}{\phi(x_k)}, \end{aligned} \quad (1.13)$$

(1.13) итерацийг 8 дугаар эрэмбэтэй байхаар α_k -г олох шаардлагатай.

Теорем 1.2. Теорем 1.1-ийн нөхцөлүүд биелж байг. Уламжлалгүй гурван алхамт (1.13) итерацийн арга 8 дугаар эрэмбийн нийлэлттэй байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь итерацийн параметр $\bar{\tau}_k$ ба α_k нь

$$\bar{\tau}_k = 1 + \hat{d}_k \theta_k + \tilde{\beta}_k \theta_k^2 + \tilde{\gamma}_k \theta_k^3 + \dots, \quad (1.14)$$

болон

$$\alpha_k = 1 + 2\theta_k + (\tilde{\beta} + 1)\theta_k^2 + (\tilde{\gamma} + 2\tilde{\beta} - 4)\theta_k^3 + (1 + 4\theta_k)v_k + O(f(x_k)^4), \quad (1.15)$$

нөхцөлүүдийг хангах явдал.

Хэрэв (1.13) итерацид $\bar{\tau}_k = H(\theta_k)$ -г (1.8), (1.9)-р тооцоолбол

$$\bar{\tau}_k = H(\theta_k) = \frac{c + (\hat{d}_k c + d)\theta_k + \omega\theta_k^2}{c + d\theta_k + b\theta_k^2}, \quad (1.16)$$

$c + d + b \neq 0, c, d, b, \omega \in R,$

$$\alpha_k = H(\theta_k) + \frac{1}{1 + \gamma\phi_k}\theta_k^2 + \hat{d}_k \left(\tilde{\beta} - \frac{2}{1 + \gamma\phi_k} \right) \theta_k^3 + (1 + 2\hat{d}_k \theta_k)v_k, \quad (1.17)$$

болох ба бид гурван алхамт уламжлалгүй бүл итерацийн аргатай болно. Үнэхээр $\bar{\tau}_k$ ба α_k нь (1.16) ба (1.17)-р өгөгдөхөд (1.14), (1.15) нөхцөлүүд

$$\tilde{\beta} = \frac{\omega - b}{c} - \frac{d}{c} \left(\frac{d}{c} + \hat{d}_k \right); \quad \tilde{\gamma} = -\frac{(b + \omega)d}{c^2} + \frac{d^2 - bc}{c^2} \hat{d}_k$$

тогтмолтойгоор биелнэ. Иймд [4] үүсгэгч функц нь уламжлалгүй гурван алхамт итерацийн бүл аргыг өгч байна.

1.3 Уламжлалгүй оновчтой аргуудын харьцуулалт

Уламжлалгүй гурван алхамт олон тооны аргууд боловсруулагдсан байдаг [2, 5–7, 9, 11–14, ZOC2]. Эдгээр нь $\bar{\tau}_k$ болон сүүлийн алхмыг тооцоолох $f'(z_k)$ -ээр өөр хоорондоо ялгагддаг. $f'(z_k)$ тооцоолоход гурван төрлийн арга ихэнхдээ хэрэглэгддэг. Эхнийх нь [6],[11–13],[9] ажлуудад авч үзсэн

$$f'(z_k) \approx N'_3(z_k), \quad (1.18)$$

дөхөлт. Энд $N_3(z)$ x_k, w_k, y_k ба z_k цэгүүд дээрх Ньютоны 3 зэргийн интерполяци. Хоёр дахь арга [7] нь ажилд дэвшүүлсэн дөхөлт

$$f'(z_k) \approx \nu_1 f(x_k) + \nu_2 f(w_k) + \nu_3 f(y_k) + \nu_4 f(z_k). \quad (1.19)$$

ν_1, ν_2, ν_3 ба ν_4 тогтмол ба $f(x) = 1, x, x^2, x^3$ хувьд (1.19) нь адилтгал болохоор тодорхойлогдоно.

[ZOC2] ажилд дараах дөхөлтийг авч үзсэн.

$$\begin{aligned} f'(z_k) &\approx af(x_k) + bf(y_k) + cf(z_k) + d\phi(x_k), \\ \phi(x_k) &= \frac{f(w_k) - f(x_k)}{w_k - x_k} = f[x_k, w_k]. \end{aligned} \quad (1.20)$$

(1.20)-д a, b, c ба d нь (1.20) тэнцэтгэл $O(f(x_k)^4)$ эрэмбийн нарийвчлалтай байх тогтмолууд.

Одоо бид (1.18), (1.19) болон (1.20) дөхөлтүүд дээр үндэслэсэн зарим аргуудыг авч үзье. Zheng нарын арга (Z8) [9] нь (1.18) дээр үндэслэсэн бөгөөд дараах хэлбэртэй

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[x_k, w_k]}, \quad w_k = x_k + \gamma f(x_k), \quad \gamma \in R \setminus \{0\} \\ z_k &= y_k - \frac{f(y_k)}{f[x_k, y_k] + f[y_k, w_k] - f[x_k, w_k]}, \\ x_{k+1} &= z_k - \frac{f(z_k)}{f[z_k, y_k] + (z_k - y_k)f[z_k, y_k, x_k] + (z_k - y_k)(z_k - x_k)f[z_k, y_k, x_k, w_k]}. \end{aligned} \quad (1.21)$$

Khatrri нарын арга (KS8) [7] нь (1.19) дээр үндэслэсэн бөгөөд дараах хэлбэртэй

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[x_k, w_k]}, \\ z_k &= y_k - \frac{f(y_k)}{\frac{x_k - y_k + \gamma f(x_k)}{(x_k - y_k)\gamma} - \frac{(x_k - y_k)f(w_k)}{(w_k - y_k)\gamma f(x_k)} - \frac{(2x_k - 2y_k + \gamma f(x_k))f(y_k)}{(x_k - y_k)(w_k - y_k)}}, \\ x_{k+1} &= z_k - \frac{f(z_k)}{H_1 + H_2 + H_3 - H_4}. \end{aligned} \quad (1.22)$$

Энд

$$\begin{aligned} H_1 &= -\frac{(y_k - z_k)(w_k - z_k)}{(x_k - z_k)\gamma(x_k - y_k)}, \\ H_2 &= \frac{(y_k - z_k)(x_k - z_k)f(w_k)}{(w_k - z_k)(w_k - y_k)\gamma f(x_k)}, \\ H_3 &= \frac{(x_k - z_k)(w_k - z_k)f(y_k)}{(y_k - z_k)(w_k - y_k)(x_k - y_k)}, \\ H_4 &= \frac{\gamma(x_k - 2z_k + y_k)f(x_k) + x_k^2 + (-4z_k + 2y_k)x_k + 3z_k^2 - 2y_k z_k}{(y_k - z_k)(x_k - z_k)(w_k - z_k)} f(z_k). \end{aligned} \quad (1.23)$$

(1.22), (1.23) арга нь [4, 11–13] ажилд дэвшүүлсэн аргуудтай төстэй гэж [7]-д тэмдэглэсэн боловч [9] дэх аргаас ялгаатай. (1.21) болон (1.22)-аас харахад (1.22) дахь хоёр болон гуравдугаар алхам нь (1.21)-өөс илүү төвөгтэй. Томьёоны хувьд олон тооны математикийн үйлдлүүд хийдэг нь тооцооллын болон тогтворжилтын үүднээс нэн тохиромжгүй. Иймд (1.22) томьёог хялбарчлах шаардлагатай.

[ZOC2]-д дэвшүүлсэн аргуудын бүл нь (1.20) дээр үндэслэсэн ба дараах хэлбэртэй

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[x_k, w_k]}, \\ z_k &= y_k - \bar{\tau}_k \frac{f(y_k)}{f[x_k, w_k]}, \\ x_{k+1} &= z_k - \alpha_k \frac{f(z_k)}{f[x_k, w_k]}, \end{aligned} \quad (1.24)$$

ҮҮНД

$$\bar{\tau}_k = \frac{c + (\hat{d}_k c + d)\theta_k + \omega\theta_k^2}{c + d\theta_k + b\theta_k^2}, \quad c + d + b \neq 0, \quad c, d, b, \omega \in R, \quad (1.25)$$

ба

$$\alpha_k = \frac{1}{\left(1 + a_k w_k \left(\frac{f[z_k, x_k]}{f[x_k, w_k]} - 1\right) + b_k \gamma_k \left(\frac{f[z_k, y_k]}{f[x_k, w_k]} - 1\right)\right)}, \quad (1.26)$$

$$a_k w_k = (1 - \tau_k) \frac{2\tau_k + \gamma\phi_k + (\tau_k + \gamma\phi_k)^2}{(\tau_k + \gamma\phi_k)(1 + \gamma\phi_k)},$$

$$b_k \gamma_k = \frac{\tau_k(\tau_k + \gamma\phi_k)}{1 + \gamma\phi_k}, \quad \phi_k = f[x_k, w_k], \quad (1.27)$$

$$\tau_k = 1 + \bar{\tau}_k \theta_k, \quad \theta_k = \frac{f(y_k)}{f(x_k)}.$$

[ZO4] ажилд (1.21), (1.22) аргууд эн чацуу болохыг үзүүлсэн.

(1.24) итерацийг илүү эвтэйхэн бичвэл

$$y_k = x_k - \frac{f(x_k)}{f[x_k, w_k]},$$

$$z_k = y_k - \bar{\tau}_k \frac{f(y_k)}{f[x_k, w_k]}, \quad (1.28)$$

$$x_{k+1} = z_k - \frac{f(z_k)}{f[z_k, y_k] + (z_k - y_k)f[z_k, y_k, x_k] + (z_k - y_k)(z_k - x_k)f[z_k, y_k, x_k, w_k]}.$$

$\bar{\tau}_k$ -н тодорхой сонголтуудад (1.28) нь хүснэгт 1.1-д дурдсан аргуудыг өгч байгаа учраас түүнийг эдгээр аргуудын өргөтгөл гэж үзэж болно.

1.4 Уламжлалгүй аргын динамик төлөвийн судалгаа

M1 болон бусад аргуудын динамик төлөвийн харьцуулалтуудыг хийе. Хүснэгт 1.1 дэх сүүлийн баганын товчилсон нэрийг цаашид ашиглана. Өндөр эрэмбийн нийлэлттэй аргууд нь олон алхмаас бүрддэг учраас функцийн илүү олон утгуудыг шаарддаг. Ийм учраас олон алхамт аргууд нь хар цэгүүдтэй байдаг (илүүдэл шийд). Энэ цэгүүдийг олохын тулд бид гурван алхамт аргуудыг доорх байдлаар бичнэ [16]:

$$x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k]} H_f(x_k).$$

Энд $H_f = 1 + \theta_k(\bar{\tau}_k + \alpha_k v_k)$. Харьцуулалт хийх үүднээс хар цэгүүдийг олъё. Хялбарчлах үүднээс $z = \pm 1$ язгууртай $p(z) = z^2 - 1$ олон гишүүнтийг авъя. Хүснэгт 1.2-д Z8, KS8, M1, L8, K8, S8, CH8 аргуудын хар цэгүүдийг оруулав.

Авч үзэж буй итерациудын хооронд харьцуулалт хийх өөр нэг арга нь таталцлын мужийг (basin of attraction) байгуулах. Бидний дэвшүүлсэн (1.28)-г $p(z) = z^k - 1$, $k = 3$ олон гишүүнтийг ашиглан бусад аргуудтай харьцуулалт хийе. Итерациудын динамик төлөвийг [16] судлахдаа бид $[-3, 3] \times [-3, 3]$ комплекс хавтгайг 600×600 хэмжээтэй торуудад хувааж энэ мужид орших бүх язгууруудыг олсон. Авч үзэж буй 12 аргын таталцлын мужийг Зураг 1.1-д дүрслэв. Зураг 1.1 ба Хүснэгт 1.2-с харахад M1 болон Z8 аргууд бусад аргаас илүү тогтвортой төлөвтэй болох нь харагдаж байна.

Хүснэгт 1.1: Параметрийн сонголт

c	d	b	w	$\bar{\tau}_k$	Арга
1	$-\hat{d}_k$	0	0	$\frac{1}{1-\hat{d}_k\theta_k}$	(KS8), (Z8) [7, 9]
1	$-\frac{1}{1+\gamma\phi_k}$	0	$\frac{a\hat{d}_k}{2}$	$\frac{1+\theta_k+a\hat{d}_k\frac{\theta_k^2}{2}}{1-\frac{\theta_k}{1+\gamma\phi_k}}$	Лотфи (L8) [11]
1	$\beta - 1 - \hat{d}_k$	$\frac{2-\beta}{1+\gamma\phi_k}$	β	$\frac{1+(\beta-1)\theta_k+\beta\theta_k^2}{1+(\beta-2-\frac{1}{1+\gamma\phi_k})\theta_k+\frac{\beta-2}{1+\gamma\phi_k}\theta_k^2}$	Кинг (K8) [12]
1	$-\frac{1}{1+\gamma\phi_k}$	0	0	$\frac{1+\theta_k}{1-\frac{\theta_k}{1+\gamma\phi_k}}$	Шарма (S8)[13]
1	$-2\alpha - \frac{1}{1+\gamma\phi_k}$	$\frac{2\alpha}{1+\gamma\phi_k}$	$H(\theta_k)$	$\frac{1}{1-2\alpha\theta_k} \frac{H(\theta_k)}{(1-\frac{\theta_k}{1+\gamma\phi_k})}$	Чебышев-Халей (CH8)[6]
1	$-\hat{d}_k$	$\frac{\hat{d}_k^2}{4}$	0	$\frac{1}{(1-\frac{\hat{d}_k}{2}\theta_k)^2}$	[16]
1	$-\hat{d}_k$	$\frac{1}{1+\gamma\phi_k}$	0	$\frac{1}{1-\hat{d}_k\theta_k+\frac{1}{1+\gamma\phi_k}\theta_k^2}$	Тукрал (T8)[8] Кунг-Трауб (KT8)[17]
1	$-\hat{d}_k$	$\frac{1}{1-\phi_k}$	0	$\frac{1}{(1-\frac{\theta_k}{1-\phi_k})(1-\theta_k)}$	Солеймани (SS8) [18]
1	-1	0	$\frac{1}{(1+\gamma\phi_k)^2}$	$(1 + \frac{\theta_k}{(1+\gamma\phi_k)} + \frac{\theta_k^2}{(1+\gamma\phi_k)^2}) \frac{1}{1-\theta_k}$	Солеймани (SV8) [15]
1	$-\hat{d}_k$	$-\frac{1}{1+\gamma\phi_k}$	0	$\frac{1}{1-\hat{d}_k\theta_k-\frac{\theta_k^2}{1+\gamma\phi_k}}$	M1

1.5 Уламжлалгүй наймдугаар эрэмбийн оновчтой аргуудыг байгуулах

Уламжлалгүй дараах итерацийн аргыг авч үзье.

$$\begin{aligned}
 y_k &= \psi_2(x_k, y_k) = x_k - \frac{f(x_k)}{\phi_k}, \\
 z_k &= \psi_4(x_k, y_k) = y_k - \tilde{\tau}_k \frac{f(y_k)}{\phi_k}, \\
 x_{k+1} &= z_k - \tilde{\alpha}_k \frac{f(z_k)}{\phi_k},
 \end{aligned} \tag{1.29}$$

ҮҮНД

$$w_k = x_k + \gamma f(x_k), \quad \phi_k = \frac{1}{\gamma} \left(\frac{f(w_k)}{f(x_k)} - 1 \right) \approx f'(x_k), \quad \gamma \in R. \tag{1.30}$$

Энд $\psi_2(x_k, y_k)$ хоёрдугаар эрэмбийн нийлэлттэй итерац.

$$\theta_k = \frac{f(y_k)}{f(x_k)}, \quad \sigma_k = \frac{f(y_k)}{f(w_k)}, \quad \text{ба} \quad v_k = \frac{f(z_k)}{f(y_k)},$$

нь (1.29) дэх үндсэн хэмжигдэхүүнүүд. Тэгвэл $x_k \rightarrow x^*$ үед $\theta_k = O(f(x_k))$, $\sigma_k = O(f(x_k))$ ба x^* нь $f(x)$ -ийн эгэл шийд. Хэрэв $\psi_4(x_k, y_k)$ нь дөрөвдүгээр эрэмбийн нийлэлттэй оновчтой итерац бол $f(z_k) = O(f(x_k)^4)$. Иймд $v_k = O(f(x)^2)$. Хэрэв бид

$$\tilde{c}_k = \frac{1}{1 + \gamma\phi_k}, \quad \tilde{d}_k = 1 + \tilde{c}_k,$$

Хүснэгт 1.2: Илүүдэл шийдүүд

Арга	Хар цэгүүд ξ	ξ тоо
Z8	$-0.555220397255420 \pm 1.15928646739103i$ $-0.460115602837211 \pm 0.456390703516719i$ $-0.450000501793328 \pm 0.129063966758804i$ $1.89303155290658 \pm 0.233570409469479i$ 1.79931236664623, 2.67863086464586	10
KS8	$-0.555220397255420 \pm 1.15928646739103i$ $-0.460115602837211 \pm 0.456390703516719i$ $-0.450000501793328 \pm 0.129063966758804i$ $1.89303155290658 \pm 0.233570409469479i$ 1.79931236664623, 2.67863086464586	10
L8 $a = 0$	$-0.667945591214872 \pm 0.100425568541848i$ $-0.664810542206285 \pm 0.236676931399034i$ $-0.256812572074558 \pm 0.169283171474748i$ $-0.235816064609120 \pm 0.101845050899904i$ $1.785182636 \pm .1865090131i$ $2.10884950227391 \pm 1.16590411063961i$ $2.176232698 \pm .5457413141i$ 2.15196391166621, 6.87682348522886 2.10230777448310, 1.71672219502143	20
M1	$-0.676558832763406 \pm 1.36018262584118i$ $-0.624888463964184 \pm 0.20890104128772i$ $-0.493766364512498 \pm 0.607060501953625i$ $-0.461962845726289 \pm 0.221119195986523i$ $-0.204327487662501 \pm 0.86651046669376i$ $1.932083323 \pm 0.1163156841i$ $2.004864313 \pm 0.7365790432i$ $2.083325978 \pm 0.4554281653i$	16

ТЭМДЭГЛЭЛ АШИГЛАВАЛ

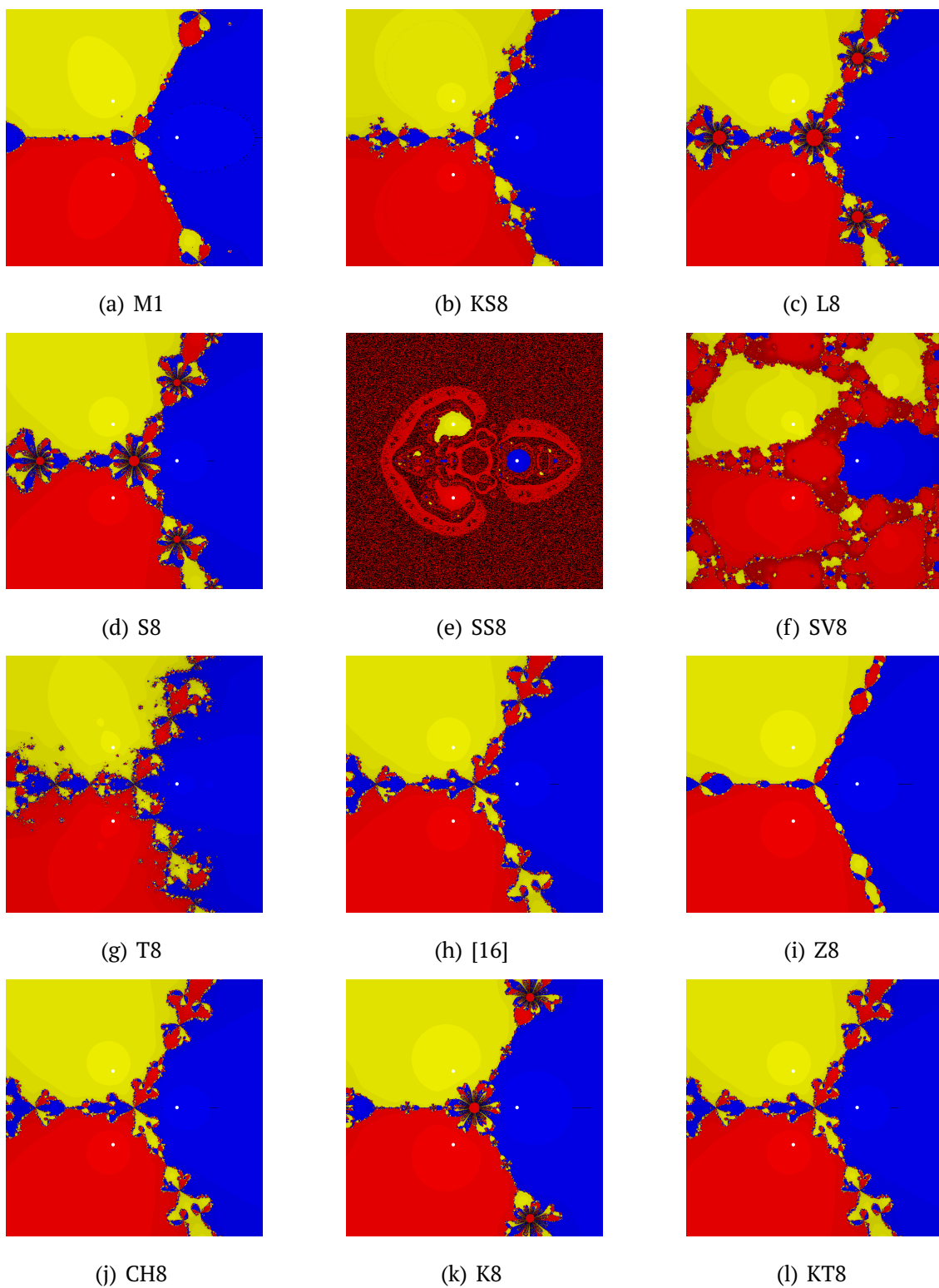
$$\sigma_k = \tilde{c}_k \theta_k, \quad \theta_k + \sigma_k = \tilde{d}_k \theta_k. \tag{1.31}$$

Тэгвэл дараах үр дүн хүчинтэй [ZOC2].

Теорем 1.3. Теорем 1.1-н нөхцөлүүд биелж байг. Тэгвэл (1.29) итерацийн нийлэлтийн эрэмбэ 8 байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь (1.29) дэх $\tilde{\tau}_k, \tilde{\alpha}_k$ нь дараах нөхцөлийг хангах явдал

$$\tilde{\tau}_k = 1 + \tilde{d}_k \theta_k + \tilde{\beta} \theta_k^2 + \tilde{\gamma} \theta_k^3 + \dots, \tag{1.32}$$

$$\begin{aligned} \tilde{\alpha}_k = & 1 + \tilde{d}_k \theta_k + (\tilde{\beta} + \tilde{c}_k) \theta_k^2 + \left(\tilde{\gamma} + \tilde{d}_k (\tilde{\beta} - 1 - \tilde{c}_k^2) \right) \theta_k^3 \\ & + (1 + 2\tilde{d}_k \theta_k) \nu_k + O(f(x_k)^4). \end{aligned} \tag{1.33}$$



Зураг 1.1: $z^3 - 1$ хувьд уламжлалгүй, гурван алхамт аргуудын таталцлын муж

(1.29) итерац нь Кунг-Траубын таамаглал ёсоор гурван алхамт уламжлалгүй оновчтой арга болно. Учир нь (1.29) арга нь давталт бүрд функцийн дөрвөн утгыг ашиглаж байгаа өөрөөр хэлбэл $m = 4$. [4, 10] дахь санааг ашиглавал $\tilde{\tau}_k$ болон $\tilde{\alpha}_k$ -г дараах байдлаар илүү ерөнхий сонгож болно:

θ_k, σ_k, v_k хангалттай гөлгөр функцийн хувьд $\tilde{\tau}_k = h(\theta_k, \sigma_k)$, $\tilde{\alpha}_k = g(\theta_k, \sigma_k, v_k)$ гэж то-

дорхойлно. $f(z_k) = O(f(x_k)^4)$ байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөлийг $h_{00} = h_{10} = h_{01} = 1$ гэж шалгаж болно. Энд $h_{ij} = h^{(i,j)}(0,0)$, $(i \geq 0, j \geq 0)$. Иймд $h_{11} = h_{02} = h_{21} = h_{12} = h_{03} = 1$ байхад (1.29) нь оновчтой итерац байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь

$$\begin{aligned} g_{000} &= 1, \\ g_{100} &= g_{010} = g_{001} = 1, \\ g_{101} &= g_{011} = 2, \\ g_{200} &= h_{20}, \quad g_{110} = 1, \quad g_{020} = 0, \\ g_{300} &= h_{20} + h_{30} - 1, \quad g_{210} = h_{20} - 1, \quad g_{120} = g_{030} = -1. \end{aligned}$$

Эдгээрийг (1.31) ашиглан шалгаж болно. $v_k = O(f(x)^2)$ учраас оновчтой томъёоны хувьд (1.33) дахь үлдэгдэл гишүүд $O(f(x_k)^4)$. Энэ утгаар бид (1.29) нь оновчтой байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь $\tilde{\tau}_k, \tilde{\alpha}_k$ -г (1.32) болон (1.33) гэж бичих явдал юм.

1.6 Хоёр алхамт итерацийн арга дахь параметрийн оновчтой сонголт

Дараах итерацийг авч үзье.

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi_k + \lambda f(w_k)}, \quad \lambda \in R, \\ x_{k+1} &= y_k - \bar{\tau}_k \frac{f(y_k)}{\phi_k + \lambda f(w_k)}, \end{aligned} \tag{1.34}$$

үүнд $w_k = x_k + \gamma f(x_k)$, $\gamma \in R$, $\phi_k = f[x_k, w_k] = \frac{f(w_k) - f(x_k)}{\gamma f_k}$. Бидний зорилго (1.34) аргыг дөрөвдүгээр эрэмбийн нийлэлттэй байхаар $\bar{\tau}_k$ параметрийг олох.

Теорем 1.4. $f : D \subset R \rightarrow R$ хүрэлцээтэй удаа тасралтгүй дифференциалчлагдах бөгөөд $x^* \in D$ эгэл шийд байг. Анхны дөхөлт x_0 -г x^* -д хүрэлцээтэй ойр бөгөөд $\bar{\tau}_k$ параметрийг

$$\bar{\tau}_k = 1 + \hat{d}_k \theta_k - \frac{\lambda f(x_k)}{\phi_k} + O(f(x_k)^2) \tag{1.35}$$

байхаар сонгосон байг. Тэгвэл (1.34) арга дөрөвдүгээр эрэмбийн нийлэлттэй.

1.7 Оновчтой зургаа болон долоодугаар эрэмбийн арга

(1.34)-д $\gamma = 0$ байг. Тэгвэл (1.34) нь

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k) + \lambda f(x_k)}, \\ x_{k+1} &= y_k - \bar{\tau}_k \frac{f(y_k)}{f'(x_k) + \lambda f(x_k)} \end{aligned} \tag{1.36}$$

болно. Теорем 1.4-р $\bar{\tau}_k$ -г

$$\bar{\tau}_k = 1 + 2\theta_k - \frac{\lambda f(x_k)}{f'(x_k)} + O(f(x_k)^2) \quad (1.37)$$

сонгосон үед (1.36) итерац нь дөрөвдүгээр эрэмбийн нийлэлттэй байна. (1.36) арга итерацийн алхам бүрд $f(x_k)$, $f(y_k)$, $f'(x_k)$ гэсэн функцийг гурван утгыг бодно. Иймд үр ашгийн индексийг тооцвол $YAI = \sqrt[3]{4} \approx 1.587$.

Теорем 1.5. $f : D \subset R \rightarrow R$ хүрэлцээтэй удаа дифференциалчлагдах бөгөөд $x^* \in D$ эгэл шийд байг. Анхны дөхөлт x_0 -г x^* -д хүрэлцээтэй ойр бөгөөд λ_k ба $\bar{\tau}_k$ параметруудийг

$$\lambda = \lambda_k = -\frac{f_k''}{2f_k'}, \quad (1.38)$$

$$\bar{\tau}_k = 1 + \frac{a_k}{2} + \frac{a_k^2}{4} + 3\theta_k + O(f(x_k)^3) \quad (1.39)$$

байхаар сонгосон байг. Тэгвэл (1.36) арга нь зургаадугаар эрэмбэтэй нийлнэ.

1.8 Санах ойтой долоодугаар эрэмбийн арга

Санах ойтой аргын хувьд дараах теорем хүчинтэй.

Теорем 1.6. $f : D \subset R \rightarrow R$ хүрэлцээтэй удаа дифференциалчлагдах бөгөөд $x^* \in D$ эгэл шийд байг. Анхны дөхөлт x_0 -г x^* -д хүрэлцээтэй ойр бөгөөд (1.34) дахь γ нь

$$\gamma = \gamma_k = -\frac{1}{f_k'}, \quad (1.40)$$

λ нь (1.38), болон $\bar{\tau}_k$ -г (1.35) ба

$$\bar{\tau}_k = 1 - \frac{a_k}{2} + \frac{3}{4}a_k^2 + 2(1 + \gamma_k\phi_k) + O(f(x_k)^3), \quad (1.41)$$

байхаар сонгосон байг. Тэгвэл (1.34) арга нь долоодугаар эрэмбийн нийлэлттэй.

Эндээс параметруудийн оновчтой сонголт нь нийлэлтийн эрэмбийг 4-с 7 хүртэл өсгөх боломж олгож байна.

(1.40) болон (1.38) сонголтууд дээр үндэслэн хоёр алхамт уламжлалгүй санах ойтой долоодугаар эрэмбийн нийлэлттэй арга байгуулагдана.

$$\begin{aligned} x_0, \lambda_0, \gamma_0 & \text{ өгөгдсөн, бол } w_0 = x_0 + \gamma_0 f(x_0), \\ \gamma_k & = -\frac{1}{N_3'(x_k)}, \quad w_k = x_k + \gamma_k f(x_k), \quad \lambda_k = -\frac{N_4''(x_k)}{2N_4'(x_k)}, \quad k = 1, 2, \dots, \\ y_k & = x_k - \frac{f(x_k)}{\phi_k + \lambda_k f(w_k)}, \\ x_{k+1} & = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi_k + \lambda_k f(w_k)}, \quad k = 0, 1, \dots, \end{aligned} \quad (1.42)$$

үүнд $\bar{\tau}_k$ нь (1.35) нөхцөлийг хангана. Энд $N_3(t, x_k, y_{k-1}, x_{k-1}, w_{k-1})$,

$N_4(t, w_k, x_k, w_{k-1}, y_{k-1}, x_{k-1})$ нь $(x_k, x_{k-1}, y_{k-1}, w_{k-1})$ болон

$(x_k, w_k, x_{k-1}, y_{k-1}, w_{k-1})$ цэгүүд дээрх гурав болон дөрвөн зэргийн Ньютоны интерполяцийн олон гишүүнт.

1.9 Шугаман биш тэгшитгэлийн систем бодох дээд эрэмбийн арга

Өгөгдсөн шугаман биш $F(x) : D \subset R^n \rightarrow R^n$ системийн хувьд $F(x) = 0$ нөхцөл хангах $(x_{(1)}^*, x_{(2)}^*, \dots, x_{(n)}^*)^T$ векторыг олох бодлого авч үзье. Энэ төрлийн бодлого нь тоон анализ, инженерчлэлд ихэвчлэн таардаг. Энэ бодлогыг бодох хамгийн өргөн хэрэглэгддэг арга нь квадрат нийлэлттэй Ньютоны арга юм

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \dots \quad (1.43)$$

Энд x_0 нь анхны дөхөлт, $F'(x)^{-1}$ нь $F(x)$ -ийн Фреше уламжлал $F'(x)$ -ийн урвуу.

1.10 Хоёр алхамт арга

Хоёр алхамт дараах аргыг авч үзье [ZO3]

$$y_k = x_k - F'(x_k)^{-1}F(x_k), \quad (1.44a)$$

$$x_{k+1} = y_k - \bar{\tau}_k F'(x_k)^{-1}F(y_k), \quad (1.44b)$$

үүнд $\bar{\tau}_k$ нь $n \times n$ хэмжээст матриц. y_k -ийн орчинд $F(x_{k+1})$ -ийн Тейлорын задаргааг ашиглавал

$$F(x_{k+1}) = \left(I - F'(y_k)\bar{\tau}_k F'(x_k)^{-1} \right) F(y_k) + O(\|F(y_k)\|^2), \quad (1.45)$$

үүнд I нь нэгж матриц. (1.45)-с $\bar{\tau}_k$ -г

$$I - F'(y_k)\bar{\tau}_k F'(x_k)^{-1} = 0 \quad \text{ог} \quad \bar{\tau}_k = F'(y_k)^{-1}F'(x_k) \quad (1.46)$$

байхаар сонговол $F(x_{k+1}) = O(\|F(x_k)\|^4)$. (1.46)-г (1.44b)-д орлуулбал

$$x_{k+1} = y_k - F'(y_k)^{-1}F(y_k). \quad (1.47)$$

Бидний зорилго $F'(y_k)^{-1}$ -г $O(\|F(x_k)\|^2)$ нарийвчлалтайгаар олох юм.

x_k орчин дахь $F'(y_k)$ -ийн Тейлорын задаргааг ашиглавал

$$\begin{aligned} F'(y_k) &= F'(x_k) - F''(x_k)F'(x_k)^{-1}F(x_k) + O(h^2) \\ &= F'(x_k)(I - P_k) + O(h^2), \end{aligned} \quad (1.48)$$

үүнд

$$P_k = F'(x_k)^{-1}F''(x_k)F'(x_k)^{-1}F(x_k), \quad (1.49)$$

ба $h = \|F(x_k)\|$, $O(\|F_k\|) = O(h)$. x_k нь x^* -д хангалттай ойрхон үед

$$\|P_k\| \leq 1, \quad (1.50)$$

гэж үзэж чадна.

$$(I - P_k)^{-1} = \sum_{j=0}^{\infty} P_k^j = I + P_k + O(h^2). \quad (1.51)$$

учир урвуугийн тухай Банахын лемм ёсоор (1.48)-с

$$F'(y_k)^{-1} = (I - P_k)^{-1} F'(x_k)^{-1} + O(h^2) = (I + P_k) F'(x_k)^{-1} + O(h^2). \quad (1.52)$$

(1.52) ойролцоо томъёоноос

$$F'(y_k)^{-1} \approx (I + P_k) F'(x_k)^{-1}, \quad (1.53)$$

үүнийг (1.47)-д орлуулбал

$$x_{k+1} = y_k - (I + P_k) F'(x_k)^{-1} F(y_k). \quad (1.54)$$

(1.46)-д (1.53)-г ашиглавал

$$\bar{\tau}_k = I + P_k + O(h^2) = I + 2\Theta_k + O(h^2). \quad (1.55)$$

Энд

$$\Theta_k = \frac{P_k}{2} = \frac{1}{2} F'(x_k)^{-1} F''(x_k) F'(x_k)^{-1} F(x_k). \quad (1.56)$$

Иймд

$$F(y_k) = \frac{F''(x_k)}{2} \left(F'(x_k)^{-1} F(x_k) \right)^2 + O(h^3), \quad (1.57)$$

$$F'(x_k)^{-1} F(y_k) = \frac{F'(x_k)^{-1} F''(x_k)}{2} \left(F'(x_k)^{-1} F(x_k) \right)^2 + O(h^3). \quad (1.58)$$

(1.56)-г (1.58)-д орлуулбал

$$F'(x_k)^{-1} F(y_k) = \Theta_k F'(x_k)^{-1} F(x_k) + O(h^3). \quad (1.59)$$

Нөгөө талаас (1.44а) ба (1.59)-г ашиглавал (1.54)-г

$$x_{k+1} = x_k - \left(I + (I + P_k) \frac{P_k}{2} \right) F'(x_k)^{-1} F(x_k),$$

эсвэл

$$x_{k+1} = x_k - \tau_k F'(x_k)^{-1} F(x_k), \quad (1.60)$$

гэж бичиж болно. Энд

$$\tau_k = I + (I + P_k) \frac{P_k}{2} = I + \Theta_k + 2\Theta_k^2 + O(h^3). \quad (1.61)$$

(1.55) ба (1.61)-с τ_k ба $\bar{\tau}_k$ хамаарал

$$\bar{\tau}_k = (\tau_k - I) \Theta_k^{-1} = I + 2\Theta_k + O(h^2) \quad (1.62)$$

гэж гарна. Иймд [3]-д өгөгдсөн Теорем 1 шугаман биш тэгшитгэлүүдийн системийн хувьд өргөтгөгдөнө.

Теорем 1.7. $F(x) : D \subset R^n \rightarrow R^n$ ба $F(x)$ -н шийд x^* -г агуулах $D \subset R^n$ гүдгэр задгай олонлогт хүрэлцээтэй удаа Фреше уламжлалтай байг. Мөн $F'(x)$ нь тасралтгүй ба x^* орчинд урвуутай байг. Тэгвэл x^* -д хүрэлцээтэй ойр анхны дөхөлтийн хувьд (1.60)-р олдох $\{x_k\}_{k \geq 0}$, $x_0 \in D$ дараалал дөрөвдүгээр эрэмбийн нийлэлттэй байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь τ_k матриц (1.61) нөхцөлийг хангах явдал юм.

Теорем 1.7-с хоёр алхамт (1.44) итерац дөрөвдүгээр эрэмбийн нийлэлттэй байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь $\bar{\tau}_k$ -г (1.62)-р сонгох гэж мөрдөнө. Θ_k матрицын утгыг тооцоолох нь практик талаасаа хүндрэлтэй байдаг. Энэ хүндрэлийг давахын тулд бид Θ_k нарийвчлал сайтай, хялбараар тооцоолох шаардлагатай. y_k -г

$$y_k = x_k - aF'(x_k)^{-1}F(x_k), \quad a \neq 0 \quad (1.63)$$

гэж тодорхойлсон байг. Θ_k -г тооцоолохын тулд $F(x)$ -н дараах ялгаварт харьцааг авч үзье [1]

$$[x + h, x; F] = \int_0^1 F'(x + th)dt = F'(x) + \frac{1}{2}F''(x)h + \frac{1}{6}F'''(x)h^2 + O(h^3), \quad (1.64)$$

үүнд $h^i = (h, h, \dots, h)$, $h \in R^n$. Эдгээрийг ашиглавал дараах дөрвөн хялбар томьёо гарна.

$$\Theta_k = \frac{1}{a}F'(x_k)^{-1} \left([y_k, x_k; F] - F'(y_k) \right) + O(h^3), \quad (1.65)$$

$$\Theta_k = \frac{1}{a}F'(x_k)^{-1} \left(F'(x_k) - [y_k, x_k; F] \right) + O(h^3), \quad (1.66)$$

$$\Theta_k = \frac{1}{2a} \left(I - F'(x_k)^{-1}F'(y_k) \right) + O(h^3), \quad (1.67)$$

$$\Theta_k = \frac{1}{2a} \left(-I + F'(y_k)^{-1}F'(x_k) \right) + O(h^2). \quad (1.68)$$

F -н ялгаварт харьцаа $[y, x; F]$ нь $n \times n$ хэмжээс матриц ([1])

$$\begin{aligned} [y, x; F]_{i,j} &= \\ &= \frac{F_i(y_{(1)}, \dots, y_{(j)}, x_{(j+1)}, \dots, x_{(n)}) - F_i(y_{(1)}, \dots, y_{(j-1)}, x_{(j)}, \dots, x_{(n)})}{y_{(j)} - x_{(j)}}, \end{aligned} \quad (1.69)$$

үүнд $1 \leq i, j \leq n$.

1.11 Гурван алхамт арга

Дараах гурван алхамт аргыг авч үзье:

$$y_k = x_k - F'(x_k)^{-1}F(x_k), \quad (1.70a)$$

$$z_k = y_k - \bar{\tau}_k F'(x_k)^{-1}F(y_k), \quad (1.70b)$$

$$x_{k+1} = z_k - \alpha_k F'(x_k)^{-1}F(z_k). \quad (1.70c)$$

(1.70) итерацийн нийлэлт дараах теоремоор өгөгдөнө:

Теорем 1.8. Теорем 1-ийн нөхцөлүүд биелж байг. Тэгвэл (1.70) итерац p эрэмбийн нийлэлттэй байх зайлшгүй бөгөөд хүрэлцээтэй нь итерацийн τ_k ба α_k параметр нь Хүснэгт 1.3 дахь нөхцөлийг хангах явдал юм.

Хүснэгт 1.3: Параметрийн сонголт

p	α_k	$\bar{\tau}_k$
5	$I + O(\Theta_k)$	$I + 2\Theta_k + \beta\Theta_k^2$
	$I + 2\Theta_k + O(\Theta_k^2)$	$I + O(\Theta_k)$
6	$I + 2\Theta_k + O(\Theta_k^2)$	$I + 2\Theta_k + O(\Theta_k^2)$
7	$I + 2\Theta_k + 6\Theta_k^2 + 3d_k$	$I + 2\Theta_k + O(\Theta_k^2)$

Энэхүү сэдэвт ажилд шугаман биш тэгшитгэл ба тэгшитгэлүүдийн системийг бодох өндөр эрэмбийн аргуудыг боловсруулж, нийлэлтийн шинжилгээ хийсэн болно.

Гол үр дүнг тоймловол:

- Дөрөв болон наймдугаар эрэмбийн нийлэлттэй оновчтой аргуудыг дэвшүүлсэн. Хоёр болон гурван алхамт уламжлалгүй аргуудын хувьд нийлэх зайлшгүй бөгөөд хүрэлцээтэй нөхцөлийг гаргаж авсан. Тэдгээр нөхцөлүүд нь нийлэлтийн эрэмбийг тогтоохоос гадна шинэ аргуудыг өгч байна. Мөн үүсгэгч функцийг тусламжтайгаар бид оновчтой, уламжлалгүй аргуудын өргөн ангийг дэвшүүлсэн. Уламжлалгүй аргуудын динамик төлөв байдлын судалгаа болон бусад аргуудтай харьцуулалт хийсэн болно. Дэвшүүлсэн бүл аргуудаас аль нь сайн болохыг бид динамик төлөвийн судалгаагаар тодорхойлсон.
- Чөлөөт параметрийг агуулах хоёр алхамт аргуудын шинэ бүлийг дэвшүүлсэн. Хоёр алхамт арга дахь параметруудийн оновчтой утгуудын хувь дахь аналитик томъёог олсон. Тэдгээр сонголтууд нь хоёр алхамт аргуудын нийлэлтийн эрэмбийг ихэсгэх боломжийг олгож байна. Иймд ямар нэг нэмэлт тооцоололгүйгээр нийлэлтийн эрэмбэ дөрвөөс долоо болж өссөн.
- Шугаман биш системийг бодох дөрвөөс долоодугаар эрэмбийн нийлэлт бүхий аргуудын бүлийг боловсруулсан. Энэ бүл арга нь өмнө боловсруулагдсан аргуудыг тухайн тохиолдол болгон өөртөө агуулж байна. Ийм учраас дэвшүүлсэн итерацийн аргын бүл нь өмнө судлаачдын боловсруулсан аргуудын өргөтгөл гэж үзэж болно. Эдгээр аргуудын хувьд нийлэх зайлшгүй бөгөөд хүрэлцээтэй нөхцөлийг томъёолсон. Мөн хоёр болон гуравдугаар эрэмбийн Фрешегийн уламжлалыг тооцоололгүйгээр параметрийн утгыг бодож байгаа нь тооцооны хүндрэлийг арилгасан. Өөрөөр хэлбэл тооцоолол дахь үйлдлийн тоо цөөрнө. Онолын үр дүнг жишээ туршилтуудаар баталгаажуулсан болно.

Бүлэг 2

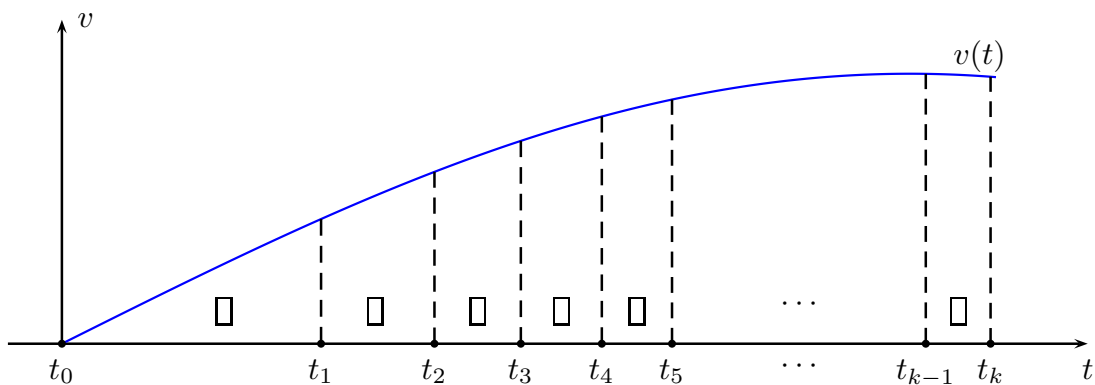
Интегро сплайн байгуулах, түүнийг хэрэглэх

“Сплайн дөхөлт, локаль интегро-сплайн байгуулах” сэдвийн хүрээнд 2018-2020 онд бид олон улсын сэтгүүлд 4 өгүүлэл хэвлүүлснээс хоёр нь Web of Science-н импакт фактор индекстэй сэтгүүлд хэвлэгдсэн. Мөн дотоодын болон олон улсын эрдэм шинжилгээний хуралд 3 удаа илтгэл хэлэлцүүлсэн.

2.1 Локаль интегро сплайны чанар болон хэрэглээ

Монгол Улсын Их Сургуулийн багш М. Баярпүрэв нарын судлаачид энкодер төхөөрөмжөөр тоноглогдсон робот машины байршлыг тодорхойлох асуудлыг судалж байна. Уг машины дугуйны хурдыг тодорхойлж өгөх даалгаврыг “Үйлдвэртэй хамтарсан математикийн семинар 2019”-д танилцуулж шийдэж өгөхийг хүссэн юм. Энэ асуудлыг шийдвэрлэх онолын нэг шийдэл нь энэхүү судалгааны ажил. Робот машины ард эсвэл урд талын хоёр дугуйд хугацаа бүртгэх энкодер төхөөрөмж суурилуулсан. Энкодер нь тогтмол талбайтай ($A_{\text{area}} = \square$) зай яваад хугацааг (Зураг 2.1-г харна уу) бүртгэнэ. Энэхүү хугацааны дарааллыг ашиглан тухайн дугуйны хурд $v(t)$ -г хэрхэн тооцоолох вэ? Энэ тохиолдолд $v(t)$ үл мэдэгдэх функцийг (ойролцоогоор) тодорхойлох нь гисто-сплайн эсвэл интегро сплайн байгуулах бодлого болно. Интегро сплайн байгуулах асуудлыг олон судлаачид авч үзсэн байдаг [19–22, 28, 33, 34]. Эдгээр судалгаа нь жигд тор дээр сплайн байгуулахтай холбоотой. Wu болон Zhang нар [27] ажилд жигд бус тор дээр интегро квадрат сплайн, Kirsiaed нар [25] ажилд интегро куб сплайн байгуулах судалгааг хийж дөхөлтийн чанарыг судалсан байдаг. Гэхдээ эдгээр ажил нь дээрх асуудлыг шийдэхэд хараахан тохиромжгүй юм. Учир нь эдгээр байгуулалт нь хугацааны $[t_0, t_k]$ завсар дахь бүх мэдээллийг ашигладаг нь бодит хугацаанд өндөр хурдаар тооцоолол хийх боломжгүй. Бид жигд бус торон дээр локаль интегро куб сплайныг хэрхэн байгуулсныг энд танилцуулъя.

$\mathcal{T}_k := \{t_0 < t_1 < \dots < t_k\}$ нь $[t_0, t_k]$ завсар дээрх жигд бус тор ба жижиг торын урт $h_{i+1} = t_{i+1} - t_i$ байг. $v(t)$ үл мэдэгдэх функцийн утгын талаарх мэдээлэл байхгүй харин



Зураг 2.1: Энкодер төхөөрөмжөөр бүртгэсэн хугацааны цуваа.

$[t_i, t_{i+1}]$ завсрын дээрх уг функцийн талбайн мэдээлэл \square өгөгдсөн.

Энэхүү $v(t)$ функцийн төлөөлөгч болох сплайн $S(t)$ -г дараах нөхцөл хангаж байхаар байгуулъя:

(i) $[t_i, t_{i+1}]$ дэд завсар бүр дээр $S(t)$ нь куб зэргийн олон гишүүнт,

$$(ii) \frac{1}{h_i} \int_{t_{i-1}}^{t_i} S(t)dt = \frac{1}{h_i} \int_{t_{i-1}}^{t_i} v(t)dt = I_i, \quad i = 1, 2, \dots, k.$$

Зураг 2.1 болон (ii) нөхцөлөөс $\square = h_i I_i$ болох нь илэрхий. $S_3(\mathcal{T}_k)$ -р \mathcal{T}_k хуваалт дээрх олон гишүүнтийн огторгуйг тэмдэглэе. Өөрөөр хэлбэл

$$S_3(\mathcal{T}_k) = \{p(x) | p(x) \in C^2[t_0, t_k]\}.$$

Энд $p(x)$ нь \mathcal{T}_k дээрх куб хүртэлх зэргийн олон гишүүнт. [29] ажилд авч үзсэнээр $S \in S_3(\mathcal{T}_k)$ -г дараах хэлбэртэй дүрсэлж болно:

$$S(t) = (1 - \xi)^2(1 + 2\xi)S_{i-1} + \xi^2(3 - 2\xi)S_i + h_i\xi(1 - \xi)\{(1 - \xi)S'_{i-1} - \xi S'_i\}, \quad (2.1)$$

эсвэл

$$S(t) = (1 - \xi)S_{i-1} + \xi S_i - \frac{h_i^2}{6}\xi(1 - \xi)[(2 - \xi)S''_{i-1} + (1 + \xi)S''_i], \quad (2.2)$$

$$t \in [t_{i-1}, t_i], \quad \xi = \frac{t - t_{i-1}}{h_i}, \quad \xi \in [0, 1].$$

Энд $S_i = S(t_i)$, $S'_i = S'(t_i)$ болон $S''_i = S''(t_i)$ тэмдэглэл хийв. (2.1) болон (2.2) -г (ii)-д орлуулбал

$$S_{i-1} + S_i = 2I_i - \frac{h_i}{6}(S'_{i-1} - S'_i), \quad i = 1, 2, \dots, k, \quad (2.3a)$$

$$S_{i-1} + S_i = 2I_i + \frac{h_i^2}{12}(S''_{i-1} + S''_i), \quad i = 1, 2, \dots, k. \quad (2.3b)$$

(2.3b)-с (2.3a)-г хасвал

$$S''_{i-1} + S''_i = \frac{2}{h_i}(S'_i - S'_{i-1}), \quad i = 1, 2, \dots, k, \quad (2.4)$$

гарах ба (2.4) тэгшитгэлээс

$$h_i^2 S''_{i-1} + (h_i^2 - h_{i+1}^2) S''_i - h_{i+1}^2 S''_{i+1} = 2(-h_i S'_{i-1} + (h_i + h_{i+1}) S'_i - h_{i+1} S'_{i+1}), \quad (2.5)$$

бүр цаашилбал

$$\mu_i S''_{i-1} + S''_i + \lambda_i S''_{i+1} = \frac{2}{h_i + h_{i+1}} (S'_{i+1} - S'_{i-1}), \quad i = 1, 2, \dots, k-1. \quad (2.6)$$

Энд

$$\mu_i = \frac{h_i}{h_i + h_{i+1}}, \quad \lambda_i = 1 - \mu_i.$$

(2.3b)-д i -г $i+1$ -р солиод (2.3a)-с нэмж эсвэл хасвал

$$I_i + I_{i+1} = \frac{1}{2}(S_{i-1} + 2S_i + S_{i+1}) - \frac{1}{24}(h_i^2 S''_{i-1} + (h_i^2 + h_{i+1}^2) S''_i + h_{i+1}^2 S''_{i+1}), \quad (2.7)$$

$$I_{i+1} - I_i = \frac{1}{2}(S_{i+1} - S_{i-1}) + \frac{1}{24}(h_i^2 S''_{i-1} + (h_i^2 - h_{i+1}^2) S''_i - h_{i+1}^2 S''_{i+1}). \quad (2.8)$$

(2.2)-с

$$S'''(t_i - 0) = \frac{S''_i - S''_{i-1}}{h_i}, \quad S'''(t_i + 0) = \frac{S''_{i+1} - S''_i}{h_{i+1}}. \quad (2.9)$$

(2.9) болон S_{i-1}, S_{i+1} -ийн Тейлорын задаргааг (2.7) болон (2.8)-д хэрэглэвэл

$$\begin{aligned} I_i + I_{i+1} &= 2S_i + \frac{h_{i+1} - h_i}{2} S'_i + \frac{h_i^2 + h_{i+1}^2}{6} S''_i + \frac{1}{24}(h_{i+1}^3 S'''(t_i + 0) - h_i^3 S'''(t_i - 0)), \\ I_{i+1} - I_i &= \frac{h_i + h_{i+1}}{2} S'_i + \frac{h_{i+1}^2 - h_i^2}{6} S''_i + \frac{1}{24}(h_i^3 S'''(t_i - 0) + h_{i+1}^3 S'''(t_i + 0)), \\ &i = 1, 2, \dots, k-1. \end{aligned} \quad (2.10)$$

(2.9)-г (2.10)-д орлуулбал

$$S'_i = \frac{2}{h_i + h_{i+1}} (I_{i+1} - I_i) + \frac{1}{12}(h_i \mu_i S''_{i-1} + 3(h_i - h_{i+1}) S''_i - h_{i+1} \lambda_i S''_{i+1}). \quad (2.11)$$

(2.5)-г (2.11)-д хэрэглэвэл

$$\begin{aligned} \mu_i S'_{i-1} + 5S'_i + \lambda_i S'_{i+1} &= \frac{12}{h_i + h_{i+1}} (I_{i+1} - I_i) + (h_i - h_{i+1}) S''_i, \\ &i = 1, 2, \dots, k-1. \end{aligned} \quad (2.12)$$

Түүнээс гадна (2.9)-г (2.6)-д орлуулбал

$$S''_i \approx \frac{S'_{i+1} - S'_{i-1}}{h_i + h_{i+1}} \quad (2.13)$$

нь $O(h_i - h_{i+1})$ нарийвчлалын хувьд үнэн. (2.12) дахь S'' гишүүнийг (2.13)-р соливол

$$(2 - 3\lambda_i) S'_{i-1} + 5S'_i + (3\lambda_i - 1) S'_{i+1} = \frac{12}{h_i + h_{i+1}} (I_{i+1} - I_i), \quad i = 1, 2, \dots, k-1. \quad (2.14)$$

(2.1)-н тасралтгүй байх чанар болон (2.3а), (2.13)-с мөн (2.14)-г бид гаргаж чадна. Жигд торны хувьд (2.14) тэгшитгэл нь сплайны тасралтгүй чанарыг илтгэх харьцаа [19] болох ба жигд бус торын хувьд ойролцоо утгатай тасралтгүйн харьцаа юм. (2.14) тэгшитгэлүүдэд S'_0, S'_k захын нөхцөл өгөгдвөл энэ нь битүү систем болно. Энэхүү систем шийдтэй эсэхийг судалъя. (2.14)-н матрицын гишүүдийг a_{ij} гэвэл

$$r_i = a_{ii} - \sum_{j \neq i} |a_{ij}|.$$

Тэгвэл дараах гурван тохиолдол гарна.

1. Хэрэв $0 \leq \lambda_i \leq \frac{1}{3}$ бол $3\lambda_i - 1 \leq 0, 2 - 3\lambda_i \geq 0$ ба $r_i = 2(1 + 3\lambda_i) \geq 2$.
2. Хэрэв $\frac{1}{3} \leq \lambda_i \leq \frac{2}{3}$ бол $3\lambda_i - 1 \geq 0, 2 - 3\lambda_i \geq 0$ болон $r_i = 4$.
3. Хэрэв $\frac{2}{3} \leq \lambda_i \leq 1$ бол $3\lambda_i - 1 \geq 0, 2 - 3\lambda_i \leq 0$ болон $r_i = 8 - 6\lambda_i \geq 2$.

Иймээс (2.14) нь диагоналийн давамгайлалтай болж цор ганц шийдтэй. Интерполяцын куб сплайнаас ялгаатай нь S'_0 ба S'_k нөхцөлүүдээс гадна S_0 эсвэл S_k өгөгдөх хэрэгтэй. S_i ба S'_i -ийг $i = 0, 1, \dots, k$ хувьд (2.3а) болон (2.14)-с олно. Гэхдээ S'_0, S'_k болон S_0 (эсвэл S_k) мэдэгдэж байх хэрэгтэй. Ингээд (2.1) хэлбэртэй дүрслэгдэх $S(t)$ олон гишүүнтийн сплайн байгуулагдлаа.

Интегро сплайн байгуулахад гуравдугаар эрэмбийн уламжлал захын цэгүүд дээр тасралтгүй байх (өөрөөр хэлбэл not-a-knot захын нөхцөл)

$$S'''(t_i - 0) = S'''(t_i + 0), \quad i = 1, 2, k - 2, k - 1$$

нөхцөл тохиромжтой байдаг. Энэ нөхцөлийг S''_i -ийн хэл дээр бичвэл

$$\lambda_i S''_{i-1} - S''_i + \mu_i S''_{i+1} = 0, \quad i = 1, 2, k - 2, k - 1. \quad (2.15)$$

(2.2)-ийн тасралтгүй байх чанар

$$\frac{6}{h_i h_{i+1}} (\lambda_i S_{i-1} - S_i + \mu_i S_{i+1}) = \mu_i S''_{i-1} + 2S''_i + \lambda_i S''_{i+1}, \quad i = 1, 2, \dots, k - 1, \quad (2.16)$$

болон (2.3b), (2.15) (эндээс гурван тэгшитгэл сонгоход хангалттай) хамтдаа S_0, S_1, \dots, S_k болон $S''_0, S''_1, \dots, S''_k$ үл мэдэгдэх бүхийн $2k + 2$ тэгшитгэлийн систем болно. Үүний бодоход мөн (2.2) гэсэн интегро сплайн байгуулагдана.

2.1.1 Локаль интегро куб сплайны байгуулалт

Тасралтгүйн чанар (2.16) болон (2.3b)-с

$$S_i = \lambda_i I_i + \mu_i I_{i+1} - \frac{h_i h_{i+1}}{24} (\mu_i S''_{i-1} + 3S''_i + \lambda_i S''_{i+1}), \quad i = 1, 2, \dots, k - 1. \quad (2.17)$$

Хэрвээ $S''_0, S''_1, \dots, S''_k$ өгөгдвөл (2.12) болон (2.17)-г ашиглаад интегро сплайныг хялбархан байгуулах боломжтой. Цаашид тооцоололд хэрэг болох тул $h_{i+1} - h_i = O(\bar{h}^2)$, $i =$

$1, 2, \dots, k-1$ нөхцөл биелдэг хуваалт \mathcal{T}_k -г бараг жигд тор гэж нэрлэнэ. Энд $\bar{h} = \max_{1 \leq i \leq k} \{h_i\}$. Тэгэхээр бараг жигд торны хувьд (2.13) харьцаа нь $O(\bar{h}^2)$ нарийвчлалын хүрээнд хүчинтэй. (2.11)-ийн баруун гар талын хоёр дахь гишүүн $O(\bar{h}^2)$ -тэй тэнцүү. Энэ бага хэмжигдэхүүнийг орхивол

$$S'_i \approx \frac{2}{h_i + h_{i+1}}(I_{i+1} - I_i) \quad (2.18)$$

ойролцоо харьцаа гарна. Гэвч энэ нь хангалттай дөхөлт болж чадахгүй тул сайжруулалт шаардлагатай. (2.11)-г (2.13)-д хэрэглэвэл

$$S''_i = \frac{2}{h_i + h_{i+1}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) + \frac{1}{12(h_i + h_{i+1})}(A_{i+1} - A_{i-1}). \quad (2.19)$$

Энд

$$A_i = h_i \mu_i S''_{i-1} + 3(h_i - h_{i+1})S''_i - h_{i+1} \lambda_i S''_{i+1}.$$

(2.9) ёсоор

$$\begin{aligned} A_{i+1} - A_{i-1} &= 4(h_{i+1} + h_i - h_{i-1} - h_{i+2})S''_i \\ &+ [4h_{i+1}(h_{i+1} - h_{i+2})S'''(t_i + 0) \\ &- 4h_i(h_{i-1} - h_i)S'''(t_i - 0) + D_{i-1} - D_{i+1}]. \end{aligned} \quad (2.20)$$

Энд

$$D_i = \frac{h_i^3 S'''(t_i - 0) + h_{i+1}^3 S'''(t_i + 0)}{h_i + h_{i+1}}.$$

$$D_{i+1} - D_{i-1} = O(\bar{h}^3), \quad h_{i+1} - h_i = O(\bar{h}^2)$$

гэж үзвэл (2.20) дахь дөрвөлжин хаалт доторх $O(\bar{h}^3)$ утгатай. (2.20)-г (2.19)-д орлуулж бага хэмжигдэхүүн $O(\bar{h}^2)$ -г орхивол

$$\tilde{S}''_i = \frac{6}{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right), \quad (2.21)$$

$$i = 2, 3, \dots, k-2.$$

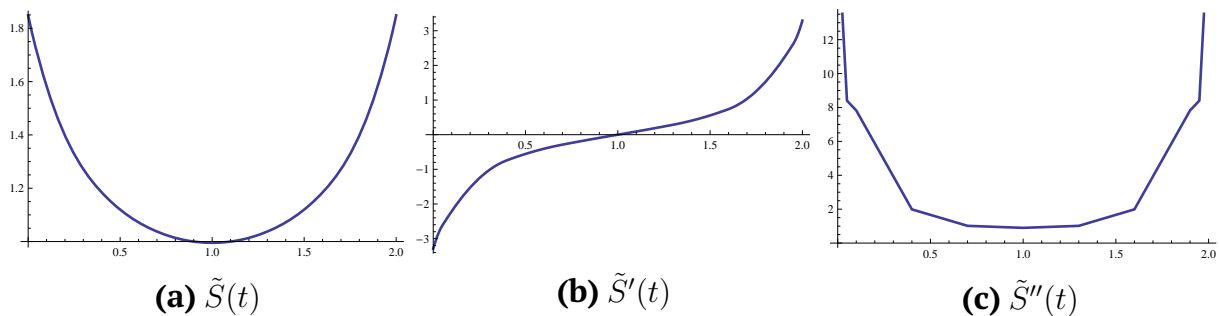
Үлдсэн \tilde{S}''_i , $i = 1, 2, k-2, k-1$ утгуудыг (2.15) тэгшитгэлийг ашиглан олно. (2.21)-г (2.11) болон (2.17)-д орлуулбал \tilde{S}'_i болон \tilde{S}_i гэсэн ойролцоо утгууд олдоно. (2.3b)-д $i = 1, k$ гэж орлуулан \tilde{S}_i -г харин (2.12)-д $i = 1, k-1$ байхад \tilde{S}'_i -г тус тус олно. Жигд торны хувьд (2.11), (2.17) болон (2.21) нь [ZM7]-д олсон ил томьёонууд болж хувирна. \tilde{S}_i , \tilde{S}'_i болон \tilde{S}''_i гэсэн ойролцоо утгуудыг ашиглан $C^2[t_0, t_k]$ интегро сплайн байгуулахдаа сплайны B -бичлэг:

$$\tilde{S}(t) = \sum_{i=-1}^{k+1} \tilde{\alpha}_i B_i(t) \quad (2.22)$$

ашиглана [29]. (2.22) дахь коэффициентууд нь

$$\begin{aligned} \tilde{\alpha}_{-1} &= \tilde{S}_0, \\ \tilde{\alpha}_i &= \tilde{S}_i + \frac{h_{i+1} - h_i}{3} \tilde{S}'_i - \frac{h_i h_{i+1}}{6} \tilde{S}''_i, \quad i = 0, 1, \dots, k, \\ \tilde{\alpha}_{k+1} &= \tilde{S}_k, \end{aligned} \quad (2.23)$$

байдгийг [30] ажилд тогтоосон. Энд $t_{-3} = t_{-2} = t_{-1} = t_0$ ба $t_k = t_{k+1} = t_{k+2} = t_{k+3}$. Ойролцоо утгууд $\tilde{S}_i, \tilde{S}'_i$ болон \tilde{S}''_i -г (2.23)-д орлуулбал $\tilde{\alpha}_i$ гэсэн ойролцоо утгууд мөн олдоно. Иймд бид локаль интегро куб сплайныг B -бичлэг дүрслэлээр байгууллаа. Байгуулсан сплайны тасралтгүй байх $\tilde{S}(t) \in C^2[t_0, t_k]$ чанарыг жишээ зургаар үзүүлбэл Зураг 2.2. Жигд торны хувьд $\tilde{\alpha}_i$ коэффициентууд нь [31]-д байгуулсан сплайны коэффициентуудтай адилхан юм.



Зураг 2.2: $v(t) = 2 - \sqrt{t(2-t)}$ -н хувьд локаль интегро куб сплайн $\tilde{S}(t) \in C^2[t_0, t_k]$.

2.1.2 Алдааны шинжилгээ болон хэлбэр хадгалах чанар

$v(t) \in C^5[t_0, t_k]$ функцийн интеграл утга I_i өгөгдсөн байг. $v(t)$ функцийн Тейлорын задаргааг (ii)-д орлуулбал

$$I_{i+1} = v_i + \frac{h_{i+1}}{2}v'_i + \frac{h_{i+1}^2}{6}v''_i + \frac{h_{i+1}^3}{24}v'''_i + O(h_{i+1}^4), \quad (2.24)$$

$$I_{i+1} = v_{i+1} - \frac{h_{i+1}}{2}v'_{i+1} + \frac{h_{i+1}^2}{6}v''_{i+1} - \frac{h_{i+1}^3}{24}v'''_{i+1} + O(h_{i+1}^4),$$

$$I_i = v_i - \frac{h_i}{2}v'_i + \frac{h_i^2}{6}v''_i - \frac{h_i^3}{24}v'''_i + O(h_i^4). \quad (2.25)$$

Энд $v_i = v(t_i)$, $v'_i = v'(t_i)$, $v''_i = v''(t_i)$ ба $v'''_i = v'''(t_i)$. Тейлорын задаргаа болон (2.24)-д i -г $i + 1$ -р орлуулан тооцоо хийвэл

$$\begin{aligned} I_{i+2} = v_i + \frac{2h_{i+1} + h_{i+2}}{2}v'_i + \frac{3h_{i+1}^2 + 3h_{i+1}h_{i+2} + h_{i+2}^2}{6}v''_i \\ + \frac{4h_{i+1}^3 + 6h_{i+1}^2h_{i+2} + 4h_{i+1}h_{i+2}^2 + h_{i+2}^3}{24}v'''_i + O(\bar{h}^4). \end{aligned} \quad (2.26)$$

Адилханаар (2.25)-д i -г $i - 1$ -р орлуулбал

$$\begin{aligned} I_{i-1} = v_i - \frac{2h_i + h_{i-1}}{2}v'_i + \frac{3h_i^2 + 3h_{i-1}h_i + h_{i-1}^2}{6}v''_i \\ - \frac{4h_i^3 + 6h_i^2h_{i-1} + 4h_{i-1}^2h_i + h_{i-1}^3}{24}v'''_i + O(\bar{h}^4). \end{aligned}$$

(2.26)-с (2.24)-г хасвал

$$\frac{2}{h_{i+1} + h_{i+2}}(I_{i+2} - I_{i+1}) = v'_i + \frac{2h_{i+1} + h_{i+2}}{3}v''_i + \frac{3h_{i+1}^2 + 3h_{i+1}h_{i+2} + h_{i+2}^2}{12}v'''_i + O(\bar{h}^3). \quad (2.27)$$

Мөн төсөөтэйгээр тооцоолол хийвэл

$$\begin{aligned} \frac{2}{h_i + h_{i+1}}(I_{i+1} - I_i) &= v'_i + \frac{h_{i+1} - h_i}{3}v''_i + \frac{h_i^3 + h_{i+1}^3}{12(h_i + h_{i+1})}v'''_i + \frac{h_{i+1}^4 - h_i^4}{60(h_i + h_{i+1})}v^{(4)}_i \\ &+ O(\bar{h}^4), \end{aligned} \quad (2.28)$$

$$\frac{2}{h_{i-1} + h_i}(I_i - I_{i-1}) = v'_i - \frac{2h_i + h_{i-1}}{3}v''_i + \frac{3h_i^2 + 3h_i h_{i-1} + h_{i-1}^2}{12}v'''_i + O(\bar{h}^3). \quad (2.29)$$

Эдгээр тооцооны үр дүнд:

Теорем 2.1. $v(t) \in C^5[t_0, t_k]$ ба \tilde{S}_i нь (2.3b), (2.17)-р, \tilde{S}'_i нь (2.11), (2.12)-р, \tilde{S}''_i нь (2.15), (2.21)-р тус тодорхойлогдсон байг. Тэгвэл бараг жигд торны хувьд дараах үнэлгээ хүчинтэй

$$\tilde{S}''_i - v''_i = O(\bar{h}^2), \quad i = 0, 1, \dots, k, \quad (2.30)$$

$$\tilde{S}'_i - v'_i = O(\bar{h}^4), \quad i = 0, 1, \dots, k, \quad (2.31)$$

$$\tilde{S}_i - v_i = O(\bar{h}^4), \quad i = 0, 1, \dots, k. \quad (2.32)$$

Баталгаа. Эхлээд (2.30) үнэлгээг гаргая. (2.27) болон (2.29)-с

$$\frac{6}{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) = v''_i + \frac{b_i}{12}v'''_i + O(\bar{h}^2). \quad (2.33)$$

Энд

$$b_i = \frac{3(h_{i+1}^2 - h_i^2) + 3(h_{i+1}h_{i+2} - h_i h_{i-1}) + h_{i+2}^2 - h_{i-1}^2}{\hat{h}_i}. \quad (2.34)$$

Бараг жигд торны хувьд v'''_i гишүүний өмнөх коэффициент $O(h_{i+1} - h_i) = O(\bar{h}^2)$ утгатай. Иймд (2.21) болон (2.33) харьцаанаас

$$\tilde{S}''_i - v''_i = O(\bar{h}^2), \quad i = 2, 3, \dots, k - 2.$$

Үлдсэн i хувьд (2.30) үнэлгээ

$$\lambda_i(\tilde{S}''_{i-1} - v''_{i-1}) - (\tilde{S}''_i - v''_i) + \mu_i(\tilde{S}''_{i+1} - v''_{i+1}) = -(\lambda_i v''_{i-1} - v''_i + \mu_i v''_{i+1}) = O(\bar{h}^2),$$

харьцаанаас гарна. Сүүлийн харьцаа нь pot-a-knot захын нөхцөлөөс шууд мөрдөнө.

Хэрэв $v(t) \in C^5$ бол (2.28)-г (2.11) ашиглавал

$$\begin{aligned} \tilde{S}'_i - v'_i &= \frac{1}{12}(h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(h_i - h_{i+1})(\tilde{S}''_i - v''_i) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) \\ &+ \frac{2}{h_i + h_{i+1}}(I_{i+1} - I_i) - v'_i + \frac{1}{12}(h_i \mu_i v''_{i-1} + 3(h_i - h_{i+1})v''_i - h_{i+1} \lambda_i v''_{i+1}) \\ &= \frac{1}{12}(h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(h_i - h_{i+1})(\tilde{S}''_i - v''_i) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) + O(\bar{h}^4). \end{aligned}$$

(2.33) болон (2.34)-г $i - 1$ болон $i + 1$ хувьд ашиглавал

$$h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1}) = \frac{h_i \mu_i b_{i-1} - h_{i+1} \lambda_i b_{i+1}}{12} v_i^{(3)} + O(\bar{h}^4).$$

Иймд бид

$$\tilde{S}'_i - v'_i = O(\bar{h}^4), \quad i = 3, 4, \dots, k-2$$

үнэлгээг гаргаж чадна. Үлдсэн i утгуудын хувьд (2.31) нь

$$\begin{aligned} \mu_i(\tilde{S}'_{i-1} - v'_{i-1}) + 5(\tilde{S}'_i - v'_i) + \lambda_i(\tilde{S}'_{i+1} - v'_{i+1}) &= \frac{12}{h_i + h_{i+1}}(I_{i+1} - I_i) \\ + (h_i - h_{i+1})(\tilde{S}''_i - v''_i) + (h_i - h_{i+1})v''_i - \mu v'_{i-1} - 5v'_i - \lambda_i v'_{i+1} &= O(\bar{h}^4), \end{aligned}$$

харьцаанаас гарна. Сүүлийн харьцаа нь (2.12)-с мөрдөх нь тодорхой юм. Цаашилбал (2.17), (2.24) болон (2.25)-с

$$\begin{aligned} \tilde{S}_i - v_i &= \lambda_i I_i + \mu_i I_{i+1} - \frac{h_i h_{i+1}}{24} (\mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(\tilde{S}''_i - v''_i) + \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) \\ &\quad - v_i - \frac{h_i h_{i+1}}{24} (\mu_i v''_{i-1} + 3v''_i + \lambda_i v''_{i+1}) \tag{2.35} \\ &= \frac{h_i h_{i+1}}{24} (\mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(\tilde{S}''_i - v''_i) + \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) + O(\bar{h}^4). \end{aligned}$$

Ингэхээр (2.35) нь (2.32) үнэлгээг $i = 1, 2, \dots, k-1$ дугааруудын хувьд хүчинтэй болохыг харуулж байна. $i = 0$ эсвэл $i = k$ байвал (2.32) үнэлгээ нь (2.3)-с мөрдөнө. \square

Санамж 1. Теорем 2.1-н баталгаанаас үзэхэд дурын жигд бус торны хувь

$$\tilde{S}''_i - v''_i = O(\bar{h}), \quad \tilde{S}'_i - v'_i = O(\bar{h}^2), \quad \tilde{S}_i - v_i = O(\bar{h}^3)$$

үнэлгээ хүчинтэй.

Санамж 2. (2.30)–(2.32) үнэлгээ нь зөвхөн зангилааны цэгийн хувьд хүчинтэй. [30, Theorem 2] болон Теорем 2.1-г ашиглавал өгөгдсөн завсар дээрх дурын цэгийн хувьд үнэлгээг гаргах боломжтой. Үнэхээр бараг жигд торны хувьд Теорем 2.1-н нөхцөлүүд биелж байвал (2.22)-р тодорхойлогдсон $\tilde{S}(t)$ локаль интегро куб сплайны хувьд

$$\|\tilde{S}^{(r)}(t) - v^{(r)}(t)\|_\infty = O(\bar{h}^{4-r}), \quad r = 0, 1, 2,$$

үнэлгээ хүчинтэй.

Теорем 2.2. Хэрэв бараг жигд торны хувьд Теорем 2.1-н нөхцөлүүд биелж байвал \tilde{S}_i болон \tilde{S}''_i коэффициент бүхий (2.2)-р тодорхойлогдсон $S(t)$ интегро куб сплайны хувьд

$$\|S^{(r)}(t) - v^{(r)}(t)\|_\infty = O(\bar{h}^{4-r}), \quad r = 0, 1, 2,$$

үнэлгээ хүчинтэй.

Санамж 3. Хэрэв бараг жигд торны хувьд Теорем 2.1, 2.2 нөхцөлүүд биелж байвал

$$\|\tilde{S}^{(r)}(t) - S^{(r)}(t)\|_\infty = O(\bar{h}^{4-r}), \quad r = 0, 1, 2,$$

үнэлгээ хүчинтэй.

Байгуулсан сплайны гүдгэр чанарыг судалцгаая. I_i өгөгдлүүдийг

$$a_i - a_{i-1} \geq 0, \quad i = 2, 3, \dots, k-1, \quad (2.36)$$

нөхцөл биелэхээр өгөгдсөн байвал гүдгэр чанартай байна гэдэг. Энд $a_i = \frac{2(I_{i+1}-I_i)}{h_i+h_{i+1}}$. Үнэндээ энэ тодорхойлолт нь [23, 24, 32] ажлууд дахь тодорхойлолтой ижил утгатай юм.

Теорем 2.3. I_i өгөгдлийн олонлог гүдгэр ба

$$\hat{h}_i := \frac{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}}{3}$$

байг. Хэрэв

$$\begin{aligned} \hat{h}_2 a_2 + \hat{h}_3 a_3 &\geq \hat{h}_3 a_1 + \hat{h}_2 a_4, \\ \hat{h}_{k-2} a_{k-4} + \hat{h}_{k-3} a_{k-1} &\geq \hat{h}_{k-3} a_{k-3} + \hat{h}_{k-2} a_{k-2}, \end{aligned}$$

нөхцөл биелдэг бол $t \in [t_0, t_k]$ бүх утгын хувьд $\tilde{S}''(t) > 0$ байна. Өөрөөр хэлбэл $\tilde{S}(t)$ нь $[t_0, t_k]$ завсар дээр гүдгэр функц.

v_i өгөгдлүүд гүдгэр ([24]-г үз) байх нь

$$\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \geq 0, \quad i = 1, 2, \dots, k-1, \quad (2.38)$$

нөхцөл биелэхтэй адилхан.

Гөлгөр функцийн хувьд

$$\frac{2}{h_i + h_{i+1}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right) = v_i'' + O(\bar{h}^2), \quad (2.39)$$

үнэлгээ хүчинтэй. Иймд (2.21), (2.30) болон (2.39)-с

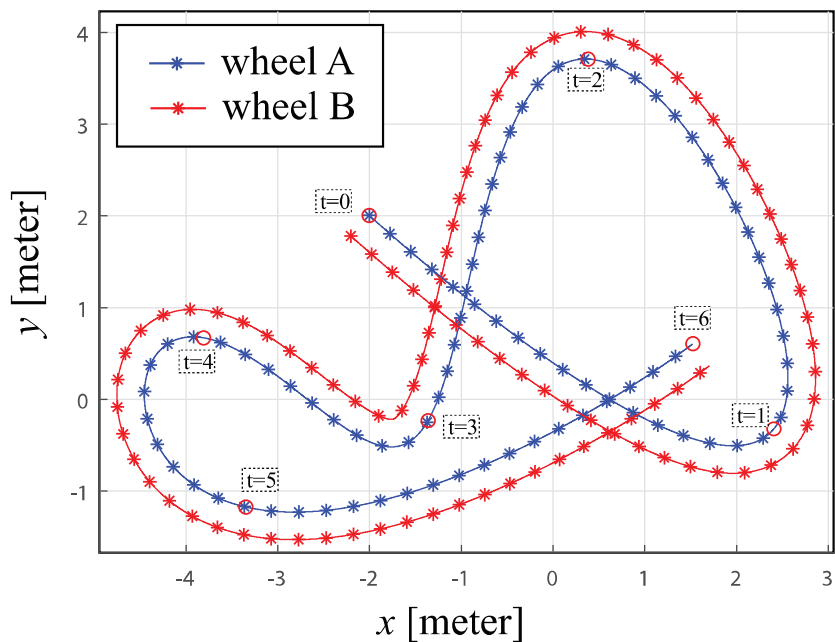
$$\frac{2}{\hat{h}_i} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) \approx \frac{2}{h_i + h_{i+1}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right),$$

нь $O(\bar{h}^2)$ нарийвчлалын хүрээнд хүчинтэй. Тэгэхээр (2.36) нь $O(\bar{h}^2)$ нарийвчлалын хүрээнд (2.38) биелэхийг агуулж байна. Теорем 2.2 ба 2.3 нь бидний байгуулсан локаль интегро куб сплайн нь дөхөлтийн болон гүдгэр байх чанартай болохыг харуулж байна. Үнэндээ интерполяцын куб сплайн $S(t) \in C^2$ гүдгэр байх хүрэлцээтэй нөхцөл [26]

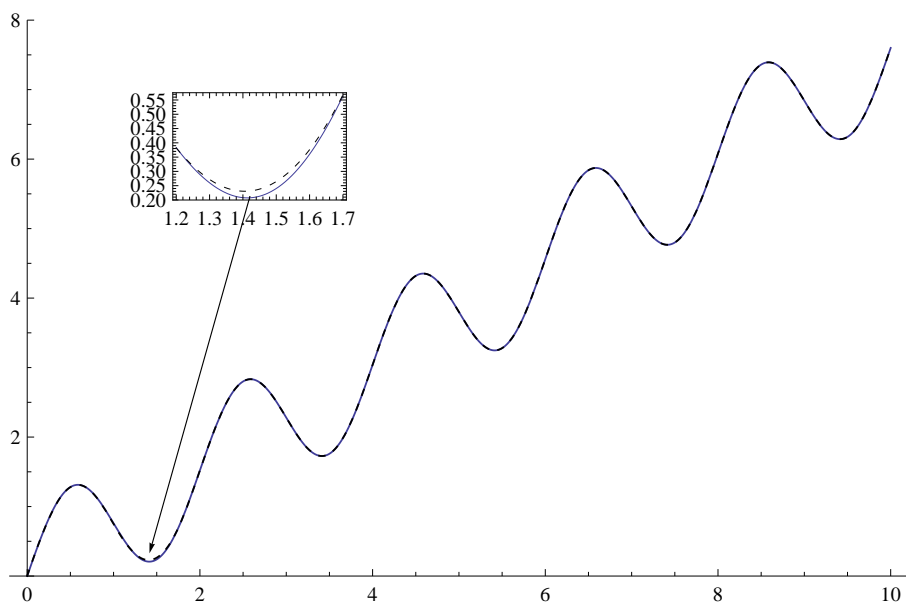
$$\begin{aligned} 2\Delta_i^2 - \mu_i \Delta_{i-1}^2 - \lambda_i \Delta_{i+1}^2 &\geq 0, \quad i = 1, 2, \dots, k-1, \\ 2\Delta_0^2 - \Delta_1^2 &\geq 0, \quad 2\Delta_k^2 - \Delta_{k-1}^2 \geq 0, \end{aligned}$$

байдаг. Энд $\Delta_i^2 = f[t_{i-1}, t_i, t_{i+1}]$. Эндээс үзвэл бидний байгуулсан локаль интегро куб сплайн (2.36) болон (2.37) нөхцөл биелэхэд л гүдгэр байгаа нь интерполяцын куб сплайнаас давуу чанартай болохыг илтгэнэ.

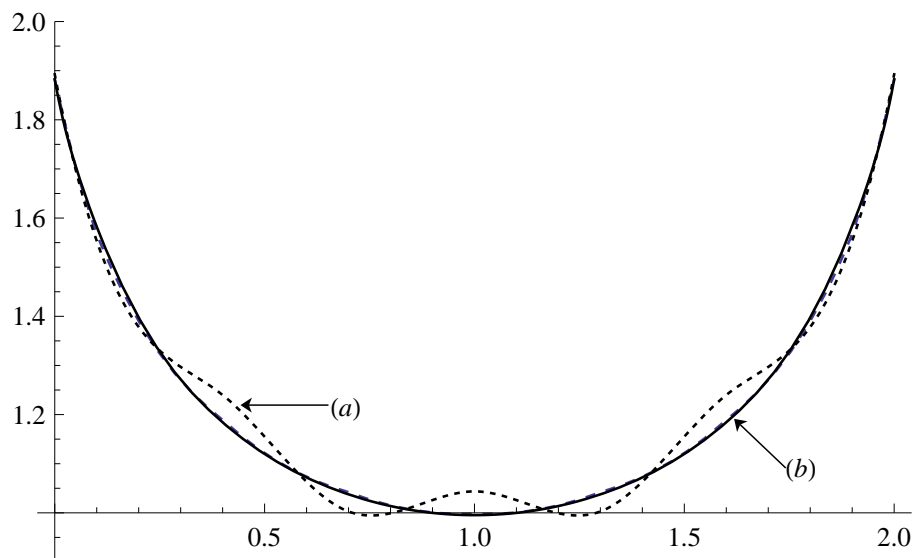
Эдгээр онолын үр дүн нь янз бүрийн тоон туршилт дээр баталгаажсан бөгөөд бид энд зөвхөн бодит хугацааны хувьд хурд олох болон байгуулсан сплайны гүдгэр чанарыг харуулах жишээг танилцуулъя. Хугацааны эхлэл $t_0 = 0$ ба дугуй $v(t) = 0.9 \sin(\pi t) + 0.76t$ хурдаар хөдөлсөн байг. Энд $\square = 0.1$. Бидний байгуулсан сплайн жинхэнэ функцийг сайн дөхөж чадсан болохыг Зураг 2.4-с харж болно. Хурдыг нарийвчлал сайтай олж чадвал роботын байрлал зөв тодорхойлогдож болохыг Зураг 2.3 харуулж байна. Энэ аргын давуу тал



Зураг 2.3: Байршил тодорхойлсон байдал, М. Баярпүрэв нарын судалгааны гар бичмэлээс авав.



Зураг 2.4: Үргэлжилж буй хөдөлгөөний хувьд хурд.



Зураг 2.5: Сплайнуудын гүдгэр чанарын харьцуулалт.

нь хурдыг бодит хугацаанд тооцоолж байгаагаараа ач холбогдолтой юм. Өмнө тасралтгүй чанарыг илэрхийлэх зураг дахь сплайн нь $\mathcal{T}_{10} = \{0, 0.05, 0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 1.95, 2\}$ тор дээр байгуулагдсан. Гүдгэр чанарыг Зураг 2.5-с харна уу. Энэ нь $v(t) = 2 - \sqrt{t(2-t)}$ функцийг дөхсөн бөгөөд (a) зураг нь (2.2) сплайн тасралтгүй чанартай буюу систем тэгшитгэл бодож гарсан үр дүн юм. Тэгвэл (b) зураг дээр (2.2) ба (2.22) сплайныг $\tilde{S}_i, \tilde{S}'_i, \tilde{S}''_i$ болон $\tilde{\alpha}_i$ коэффициентуудтай байгуулсныг харуулсан. Зургаас харвал (b) нь гүдгэр чанартай бол (a) гүдгэр биш хэлбэлзэн савласныг харуулж байна.

2.2 Интегро сплайны харьцуулалт, бусад чанар

Энэ хэсэгт [ZM7, ZM8, ZM9] ажлууд дахь гол үр дүнгүүдийг тоймловол:

(2.22) дүрслэлийн коэффициентуудын хувьд

$$\begin{aligned} \mu_{-1} &= S_0 - h_0 w m_0 + \frac{h_0^2 w^2}{3} M_0, \\ \mu_i &= S_i + \frac{h_i - h_{i-1}}{3} m_i - \frac{h_i h_{i-1}}{6} M_i, \quad i = 0(1)N, \\ \mu_{N+1} &= S_N - h_{N-1} w m_N + \frac{h_{N-1}^2 w^2}{3} M_N, \end{aligned} \tag{2.40}$$

байдгийг [30] ажилд тогтоосон. Энэ чанарт тулгуурлан куб сплайны коэффициентуудыг ил хэлбэрээр тооцоолох аргыг боловсруулж куб сплайнуудын хооронд харьцуулалт хийсэн [ZM7] болно.

Куб сплайны хувьд $C^2[a, b]$ ангид байх нь гөлгөр чанар сайтай болохыг илтгэдэг хэдий ч хэлбэр хадгалах чанар нь алдагддаг дутагдалтай. Хэлбэр хадгалахгүй байх нь зарим үед алдаа ихтэй болж хэлбэлзэл үүсгэдэг. [ZM9] ажилд гүдгэр эсвэл өсөх чанартай өгөгдөл $\{I_j\}$ хувьд C^1 ангийн хэлбэр хадгалах чанартай интегро сплайныг байгуулан холбогдох теоремуудыг баталсан болно.

Локаль сплайны коэффициентуудыг тооцоолох ил томъёонуудыг ашиглан гурав болон таван диагональтай матрицын урвууг хялбараар тооцоолох аргыг [ZM8] ажилд авч үзсэн. Үндсэн үр дүнг сийрүүлбэл:

$\mathbf{A} = \text{Tri-diag}\{1, d, 1\}$ нь $|d| > 2$ байх утгатай гурван диагональтай матриц байг. Тэгвэл түүний урвуу матрицын элемент $\alpha_{i,j}$ -г дараах ил томъёогоор тооцоолж болно.

$$\alpha_{ii} = \frac{1}{d - \frac{1}{\frac{d}{2} + \frac{1}{3d+8}}} = \frac{3d^2 + 8d + 2}{3d^3 + 8d^2 - 4d - 16}, \quad \alpha_{i,i\pm 1} = \frac{1 - d\alpha_{ii}}{2}, \quad (2.41)$$

$$\alpha_{i,i\pm 2} = \frac{3}{3d^3 + 8d^2 - 4d - 16}, \quad \alpha_{ij} = O(h^4), \quad |i - j| \geq 3.$$

Бүлэг 3

Шредингерийн тэгшитгэлийн шийдийн тоон ба чанарын судалгаа

Судалгааны энэ чиглэлээр академич О. Чулуунбаатар олон улсын эрдэмтэдтэй хамтран ИБР-2М хэмээх нейтроны импульсэн реактор дээр нейтроны ба цөмийн физик болон хатуу ба шингэн төлөвт орших материалын судалгаа, нуклотрон хэмээх хурдасгуур дээр хүнд ионыг асар өндөр энергитэй болтол нь хурдасгах их энергитэй цөм, эгэл бөөмсийн физикийн туршилт судалгаа, У-400 ба У-400М хэмээх хүнд ионы хурдасгуур дээр хэт хүнд химийн элементүүдийг синтезлэн гаргаж авах судалгаа, фазатрон хэмээх протоны хурдасгуур дээр цацрагийн анагаах ухааны судалгааг явуулж байна. Уг хамтарсан судалгааны үр дүн дэлхийд нэр хүндтэй “Nature physics” сэтгүүлд хэвлэгдсэн нь онцлох үйл явдал юм¹.

1. Москвагийн их сургуулийн болон Бельги Улсын Хатуу бие, нано-технологийн институтийн эрдэмтэдтэй хамтран “идэвхтэй” электрон цөмтэй, эсвэл импульсийн огторгуйд атом цөмүүд болон молекулууд тусгаарлагдсан потенциалуудтай харилцан үйлчлэх математик загварыг судалсан. Уг загварыг устөрөгчийн атом, устөрөгчийн молекул гаднын хэт богино импульстэй лазертай харилцан үйлчлэх үйлчлэлийг судлахад хэрэглэв. Бидний загвар нь сөрөг устөрөгчийн ион харилцан үйлчлэлийн бага радиусын мужид хангалттай сайн ажиллаж байгааг харуулсан
2. Хоёроос зургаа хүртэлх хэмжээст симплекс мужаар авсан интегралыг бодох найм хүртэлх эрэмбийн, эерэг коэффициентууд бүхий бүх зангилааны цэгүүд нь симплекс дотроо оршдог Гауссын төрлийн квадратурын томьёонуудыг байгуулсан. Эдгээр квадратурын томьёонуудыг хоёроос зургаа хүртэлх хэмжээст эллипслэг төрлийн захын бодлогын шийдийг төгсгөлөг элементийн аргаар өндөр эрэмбийн нарийвчлалтайгаар бодоход хэрэглэж онолын физикийн зарим асуудлуудын математик загварууд дээр тоон туршилтуудыг хийсэн. Жишээ болгон дурдахад, аксиал

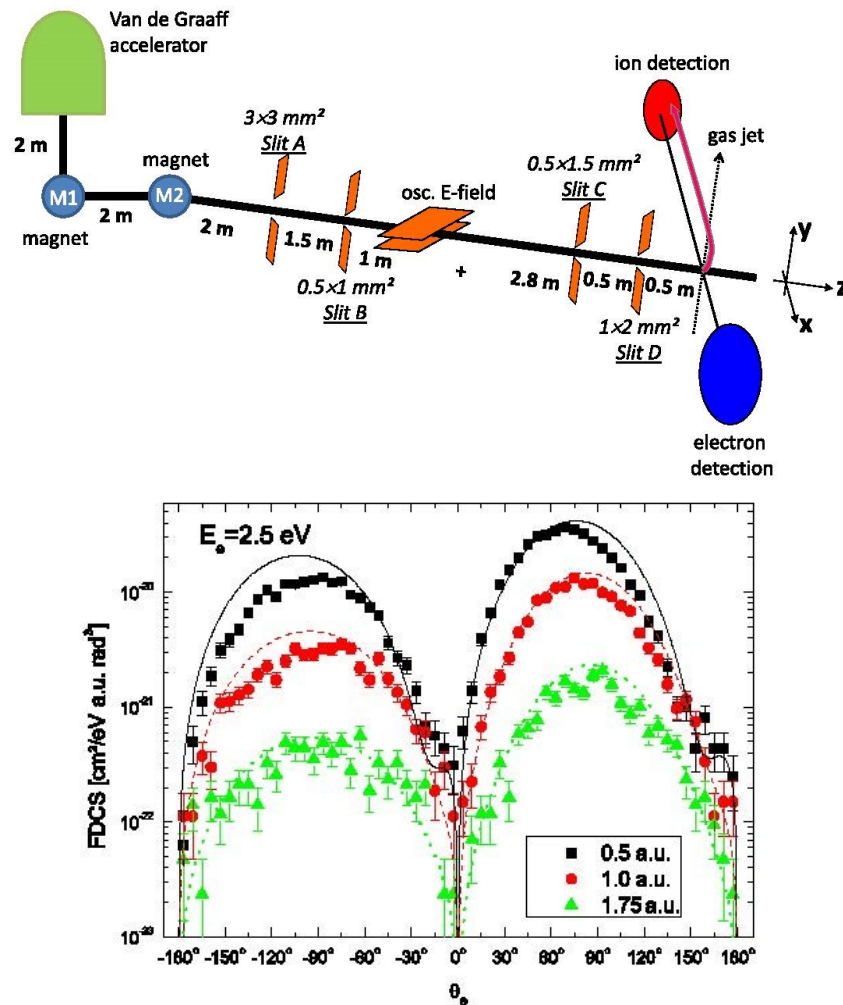
¹Тайлбар: Манай төслийн багийн судлаач О.Чулуунбаатар уг бүтээлийн хамтран зохиогч бөгөөд төслөөс уг судалгаанд хөрөнгө зарцуулаагүй болно. Энэ бүтээлийг Монгол улсынхаа судлаачдад түгээн дэлгэрүүлэх зорилгоор зохиогчдын албан ёсны зөвшөөрлийн дагуу уг тайланд бүрэн эхээр нь хавсралтаар оруулсан болно.

тэгш хэмтэй потенциалуудын хувьд сарнилын амплитудыг (хоёр хэмжээст бодлого), мөн гелийн атомын үндсэн төлөвийг бодсон (гурван хэмжээст бодлого) тоон үр дүнгүүдийг 4 дүгээр эрэмбийн нийлэлтэй Нумеровын аргаар бодсон үр дүнгүүдтэй харьцуулж, нарийвчлалын төдийгүй тооцоолох хугацааны хувьд бидний арга илт давуу байгааг харуулсан. Харгалзах комплекс программуудыг Fortran болон Maple хэл дээр бичсэн.

3. Гетегийн нэрэмжит Франкфуртын их сургуулийн судлаачид Petra III синхротрон (DESY, Гамбург) дээр COLTRIMS (COLd Target Recoil Ion Momentum Spectroscopy) детекторын тусламжтайгаар чөлөөт гелийн атомууд дээрх комптоны сарнилын шинж чанарыг судлах туршилт хийж, 2.1 кэВ энергитэй фотонуудын комптоны сарнилыг гелийн атомын иончлолын босго энергийн (ө.х. гелийн атомын дан иончлолын процессын шилжилтийн энерги 24.6 эВ) орчимд судалсан. Уг туршилтын онолын загварыг Москвагийн Их Сургуулийн судлаачидтай хамтран боловсруулж, харгалзах тооцоог хийв. Холбоост электронуудаар сарнисан цацрагийн өнцгийн тархалт нь Томсоны томьёогоор өгөгддөг чөлөөт электронуудаар сарнисан цацрагийн өнцгийн тархалтаас эрс ялгаатай болохыг харуулсан. Онол туршилтын үр дүнгүүд хангалттай сайн тохирч байгааг үзүүлсэн.
4. Хятадын Атомын энергийн институтийн судлаачидтай хамтран хүнд бөөмүүд ($^{64}\text{Ni} + ^{100}\text{Mo}$, $^{36}\text{S} + ^{48}\text{Ca}$) нэгдэх урвалын (сарнилын) онолын тооцоог KANTBP² комплекс программд зарим нэмэлт өөрчлөлт оруулж хийв. KANTBP комплекс программ нь бага энергийн мужид 4 эрэмбийн нарийвчлалтай Нумеровын аргаар боддог CCFULL³ программаас илүү тогтвортой үр дүн өгч байгааг харуулсан. Бид CCFULL программын алдааг олж засвар хийсэн. Мөн энэ тухай уг программын зохиогчид нь мэдэгдсэн, программын алдааг хүлээн зөвшөөрч засварласан шинэ хувилбараа ирүүлсэн. Гэсэн хэдий ч бага энергийн мужид дахь үр дүн сайжраагүй.
5. Францын Метц хотын Атом молекулын мөргөлдөөний лабораторийн судлаачидтай хамтран H_3^+ молекулын (гурвалсан устөрөгчийн молекулын ион буюу 5 биеийн бодлого) үндсэн болон өдөөгдсөн төлөвүүдийн энергийг Slater төрлийн суурь функцүүд ашиглан бодсон. Долгионы функцууд нь D_{3h} группэд тэгш хэмтэй байх ёстой. өгөгдсөн 7 параметртэй суурь функц нь 2 электрон, 3 протоны хувьд тэгш хэмтэй буюу $2 \times 3! = 12$ гишүүнээс бүрдэнэ. Уг суурь функцүүд нь электронууд протонуудаас хол байхад зөв асимптотик өгдөг учраас (нийт $7 + 7 = 14$ параметртэй) 24 гишүүн бүхий долгионы функц нь хэдэн зуун Gaussian төрлийн суурь функцүүд хэрэглэж бодсон үр дүнтэй хангалттай сайн тохирч байгаа болно. Slater төрлийн

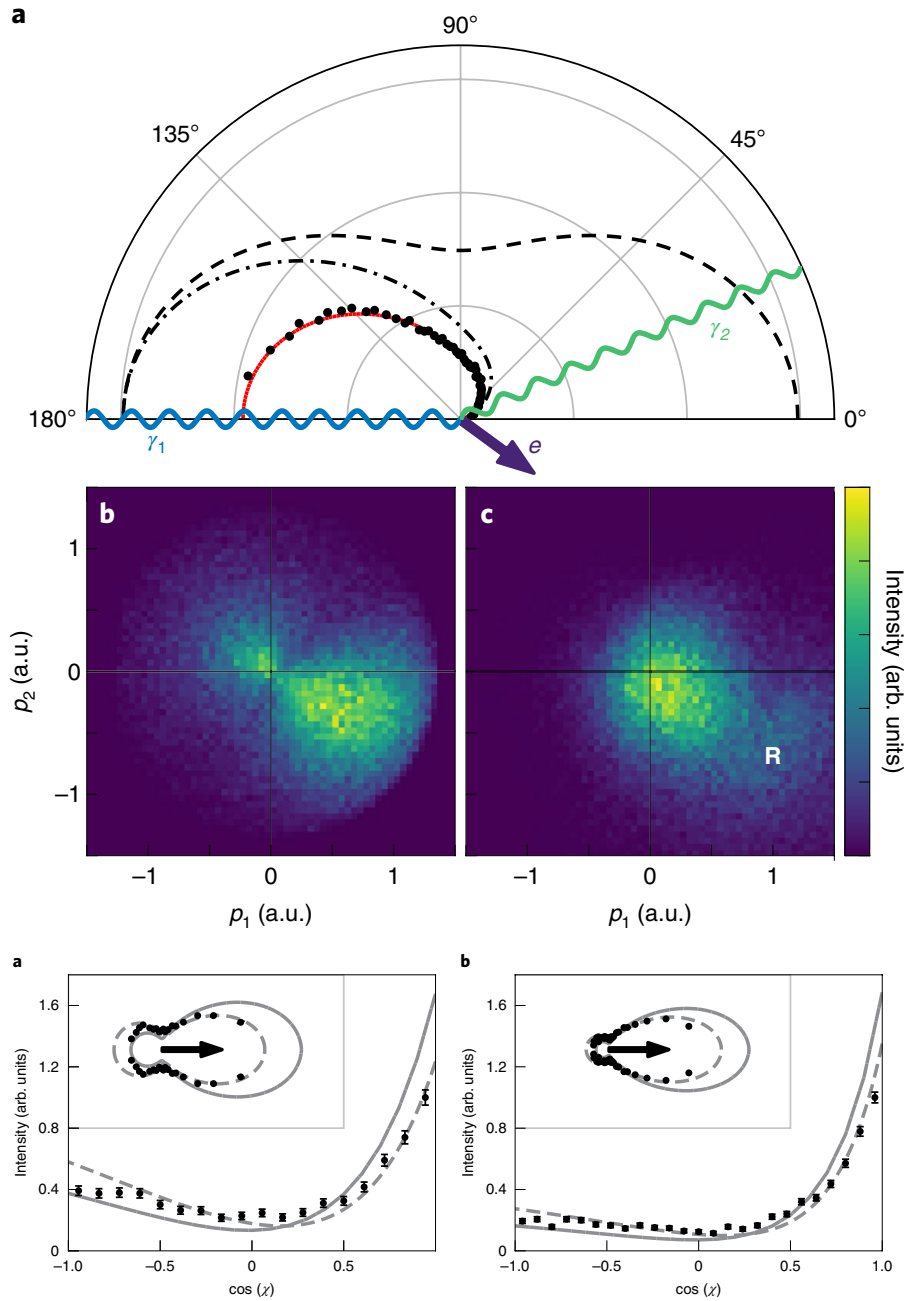
²A.A. Gusev, O. Chuluunbaatar, S.I. Vinitzky and A.G. Abrashkevich, KANTBP 3.0: New version of a program for computing energy levels, reflection and transmission matrices, and corresponding wave functions in the coupled-channel adiabatic approach, *Comput. Phys. Commun.* 185, 3341 (2014)

³K.Hagino, N.Rowley, T.Kruppa, A program for coupled-channel calculations with all order couplings for heavy-ion fusion reactions, 123, 143 (1999)



Зураг 3.1: Электроны сарнилын энерги 2.5 ± 1 eV, шилжилтийн импульс $q = 0.5 \pm 0.15$ au., $q = 1.0 \pm 0.25$ au, $q = 1.75 \pm 0.4$ au.

суурь функцүүд хэрэглэхэд гардаг гол хүндрэл нь бүх матрицын элементүүд нь 6D интегралуудаар илэрхийлэгддэг.



Зураг 3.2: Nature сэтгүүлд хэвлэгдсэн ажлын үр дүнгээс.

ХЭВЛҮҮЛСЭН ӨГҮҮЛЭЛ:

- [ZO1] T. Zhanlav, Kh. Otgondorj, *A new family of optimal eighth-order methods for solving nonlinear equations*, Am. J. Comput. Appl. Math. 8. pp. 15–19 (2018).
- [ZOC2] T. Zhanlav, Kh. Otgondorj, O. Chuluunbaatar, *Families of Optimal Derivative-Free Two- and Three-Point Iterative Methods for Solving Nonlinear Equations*, Comput. Math. Math. Phys. 59 pp. 864–880 (2019) (Impact Factor:0.584).
- [ZO3] T. Zhanlav, Changbum Chun, Kh. Otgondorj, V. Ulziibayar, *High-order iterations for systems of nonlinear equations*, Int. J. Comput. Math. 97 (2019) 1704–1724 (Impact Factor: 1.60).
- [ZO4] T. Zhanlav, Kh. Otgondorj, *Comparison of some optimal derivative-free three-point iterations*, J. Numer. Anal. Approx. Theor. 49 pp. 76–90 (2020).
- [ZO5] T. Zhanlav, Kh. Otgondorj, *On the Optimal Choice of Parameters in Two-Point Iterative Methods for Solving Nonlinear Equations.*, Comput. Math. and Math. Phys. pp. 61, 29–42 (2021). <https://doi.org/10.1134/S0965542520120180>. (Impact Factor:0.584).
- [ZOM6] T. Zhanlav, Kh. Otgondorj, R. Mijiddorj, *Constructive theory of designing optimal eighth-order derivative-free methods for solving nonlinear equations*, Am. J. Comput. Math. 10 pp. 100–117 (2020).
- [ZM7] T. Zhanlav and R. Mijiddorj, *A comparative analysis of local cubic splines*, Comp. Appl. Math. 37, 5576–5586 (2018) (Impact Factor: 0.863).
- [ZM8] T. Zhanlav and R. Mijiddorj and H. Behforooz, *On the approximation of inverse of some band matrices and their applications in local splines*, Commun. in Numer. Anal., 1 (2018) 56–65, doi.org/10.5899/2018/cna-00302.
- [ZM9] T. Zhanlav and R. Mijiddorj, *Construction of a Family of C^1 Convex Integro Cubic Splines*, Communications in Mathematics and Applications, 11 (4) (2020) pp. 527–538, doi.org/10.26713/cma.v11i4.1386 (Web of Science emerging indexed journal).
- [ZM10] T. Zhanlav and R. Mijiddorj, *Integro Cubic Splines on Non-Uniform Grids and Their Properties*, East Asian J. Appl. Math., 11 (2021) 406–420, [doi:10.4208/eajam.030920.251220](https://doi.org/10.4208/eajam.030920.251220), (Impact Factor:1.848).
- [CHU11] A.A. Gusev, V.P. Gerdt, O. Chuluunbaatar, G. Chuluunbaatar, S.I. Vinitisky, V.L. Derbov, A. Gózdź, P.M. Krassovitskiy, *Symbolic-numerical algorithms for solving elliptic boundary-value problems using multivariate simplex lagrange elements*, Lecture Notes in Computer Science 11077, pp. 197–213 (2018).
- [CHU12] M. Kircher, F. Trinter, S. Grundmann, I. Vela-Perez, S. Brennecke, N. Eicke, J. Rist,

S. Eckart, S. Houamer, O. Chuluunbaatar, Yu.V. Popov, I.P. Volobuev, K. Bagschik, M.N. Piancastelli, M. Lein, T. Jahnke, M.S. Schöffler and R. Dörner, *Kinematically complete experimental study of Compton scattering at helium atoms near the threshold*, Nature Physics 16, pp. 756–760 (2020), **(Impact Factor: 21.797)**.

[CHU13] P.W. Wen, O. Chuluunbaatar, A.A. Gusev, R.G. Nazmitdinov, A.K. Nasirov, S.I. Vinitzky, C.J. Lin, and H.M. Jia, *Near-barrier heavy-ion fusion: Role of boundary conditions in coupling of channels*, Phys. Rev. C 101, pp. 014618–1–10 (2020), (Impact Factor: 2.988).

[CHU14] O. Chuluunbaatar, S. Obeid, B.B. Joulakian, A.A. Gusev, P.M. Krassovitskiy, L.A. Sevastianov, *D_{3h} symmetry adapted correlated three center wave functions of the ground and the first five excited states of H_3^+* , Chem. Phys. Lett. **746**, pp. 137304–1–8 (2020), (Impact Factor: 2.029).

Хурлын илтгэлүүд

Олон эрдэм шинжилгээний хурал

1. V. Derbov, G. Chuluunbaatar, A. Gusev, O. Chuluunbaatar, S. Vinitzky et al, “Metastable states of diatomic beryllium molecule, Compton ionization of atoms near threshold as a method of spectroscopy of outer shells”, LXX International conference NUCLEUS-2020, Nuclear physics and elementary particle physics. Nuclear physics technologies, 11-17 October, 2020 Saint Petersburg, <https://indico.cern.ch/event/839985/overview>
2. O. Chuluunbaatar et al, “A Maple implementation of the finite element method for solving boundary problems of the systems of ordinary second order differential equations”, Maple Conference, Waterloo Maple Inc., 2-6 November, 2020, Canada, <https://www.maplesoft.com/mapleconference>
3. T. Zhanlav, Kh. Otgondorj, “High-order iteration for systems of nonlinear equations”, Japan-Mongolia joint workshop on pure and applied mathematics, Ulaanbaatar, Mongolia, October 24-25, 2019.
4. R. Mijiddorj, A. Enkhbayar, “Research Results on A Study of a Discrete Surface Using by Integro Spline”, Japan-Mongolia joint workshop on pure and applied mathematics, Ulaanbaatar, Mongolia, October 24-25, 2019.
5. T. Zhanlav, Kh. Otgondorj, “On the optimal choices of parameters in the two-point iterative methods for solving nonlinear equations” , The 6th International Conference on Optimization, Simulation and Control (COSC2019), Ulaanbaatar, Mongolia, June 21-23, 2019.
6. T. Zhanlav, R. Mijiddorj, “Numerical Solution of Burgers’ Equation by Local Integro Spline”, The 6th International Conference on Optimization, Simulation and Control (COSC2019), Ulaanbaatar, Mongolia, June 21-23, 2019.

7. T. Zhanlav, Kh. Otgondorj, “Numerical and graphic comparison between optimal derivative-free methods”, The first International Conference on Applied Sciences and Engineering, 2019, SHUTIS, UB Mongolia.
8. T. Zhanlav, R. Mijiddorj, Kh. Otgondorj, “An optimal derivative-free iterations for solving nonlinear equations”, The Third Mongolia-Russia-Vietnam Workshop on Numerical Solution of Integral and Differential Equations, (NSIDE 2018), Hanoi, Vietnam, October 22–27, 2018.
9. T. Zhanlav, Kh. Otgondorj, “Comparison of some optimal derivative-free three-point iterations”, The Third Mongolia-Russia-Vietnam Workshop on Numerical Solution of Integral and Differential Equations, (NSIDE 2018), Hanoi, Vietnam, October 22–27, 2018.
10. S. Obeid, O. Chuluunbaatar and B. Joulakian, “About the numerical calculations of the energy of H_3^+ ”, International Conference “The Molecular Electronic Structure” (MES2018), 28-31 August 2018, Metz, France. <https://mesm.event.univ-lorraine.fr>.

Дотоодын эрдэм шинжилгээний хурал

11. Т. Жанлав, Х. Отгондорж, “Optimization by parameters in the iterative methods for solving nonlinear equations”, МУИС-ийн Хэрэглээний Шинжлэх Ухаан, Инженерчлэлийн Сургууль, “Хэрэглээний математик 2019” 2019.11.23.
12. Т. Жанлав, Р. Мижиддорж, “Integro cubic spline on arbitrary spaced sub-intervals”, МУИС-ийн Хэрэглээний Шинжлэх Ухаан, Инженерчлэлийн Сургууль, “Хэрэглээний математик 2018” 2018.11.17.

Ашигласан материалын жагсаалт

- [1] F. A. Potra, *Nondiscrete induction and iterative processes*, Pitman, London, 1984.
- [2] M. Petković, B. Neta, L. Petković J. Dzunić, *Multipoint methods for solving nonlinear equations*, Elsevier, 2013.
- [3] T. Zhanlav, V. Ulziibayar, O. Chuluunbaatar, Necessary and sufficient conditions for the convergence of two- and three-point Newton-type iterations, *Comput. Math. Math. Phys.* 57 (2017) 1090–1100.
- [4] T. Zhanlav, O. Chuluunbaatar, V. Ulziibayar, Generating function method for constructing new iterations, *Appl. Math. Comput.* 315 (2017) 414–423.
- [5] A. Cordero, J. L. Hueso, E. Martinez, J. R. Torregrosa, A new technique to obtain derivative-free optimal iterative methods for solving nonlinear equations, *J. Comput. Appl. Math.* 252 (2013) 95–102.
- [6] I. K. Argyros, M. Kansal, V. Kanwar, S. Bajaj, Higher-order derivative-free families of Chebyshev-Halley type methods with or without memory for solving nonlinear equations, *Appl. Math. Comput.* 315 (2017) 224–245.
- [7] S. K. Khattri, T. Steihaug, Algorithm for forming derivative-free optimal methods, *Numer. Algor.* 65 (2014) 809–842.
- [8] R. Thukral, Eighth-order iterative methods without derivatives for solving nonlinear equations, *ISRN Appl. Math.* Article ID 693787, (2011) 12 pages.
- [9] Q. Zheng, J. Li, F. Huang, An optimal Steffensen-type family for solving nonlinear equations, *Appl. Math. Comput.* 217 (2011) 9592–9597.
- [10] Y. Peng, H. Feng, Q. Li, X. Zhang, A fourth-order derivative-free algorithm for nonlinear equations, *J. Comput. Appl. Math.* 235 (2011) 2551–2559.
- [11] T. Lotfi, F. Soleymani, M. Ghorbanzadeh, P. Assari, On the construction of some tri-parametric iterative methods with memory, *Numer. Algor.* 70 (2015) 835–845.
- [12] S. Sharifi, S. Siegmund, M. Salimi, Solving nonlinear equations by a derivative-free form of the King's family with memory, *Calcolo.* 53 (2016) 201–215.
- [13] J. R. Sharma, R. K. Guha, P. Gupta, Some efficient derivative free methods with memory for solving nonlinear equations, *Appl. Math. Comput.* 219 (2012) 699–707.
- [14] R. Behl, D. Gonzalez, P. Maraju, S. S. Motsa, An optimal and efficient general eighth-order derivative-free scheme for simple roots, *J. Comput. Appl. Math.* 330 (2018) 666–675. order with and without memory iterative method, *J. Nonlinear Sci. Appl.* 9 (2016) 1410–1423.

- [15] F. Soleymani, S. K. Vanani Optimal Steffensen-type methods with eighth order of convergence, *Comput. Math. Appl.* 62 (2011) 4619–4626.
- [16] C. Chun, B. Neta, Comparative study of eighth–Order methods for finding simple roots of nonlinear equations, *Numer. Algor.* 74 (2017) 1169–1201.
- [17] H. T. Kung, J. F. Traub, Optimal order of one–point and multi–point iteration. *J. Assoc. Comput. Math.* 21 (1974) 643–651.
- [18] F. Soleymani, S. Shateyi, Two optimal eighth-order derivative–free classes of iterative methods, *Abstr. Appl. Anal.* ID :318165, 14 (2012). doi:10.1155/2012/318165.
- [19] H. Behforooz, *Approximation by integro cubic splines*, *Appl. Math. Comput.* **175**, 8–15 (2006).
- [20] D. Barrera, S. Eddargani and A. Lamnii, *Uniform algebraic hyperbolic spline quasi-interpolant based on mean integral values*, *Comput. Math. Methods*, DOI: 10.1002/cmm4.1123.
- [21] S. Eddargani, A. Lamnii and M. Lamnii, *On algebraic trigonometric integro splines*, *Z. Angew Math. Mech.* e201900262, DOI: 10.1002/zamm.201900262 (2019).
- [22] X. Guo, X. Han and Y. Zhang, *The local integro splines with optimized knots*, *Comp. Appl. Math.* **38**:156 doi: 10.1007/s40314-019-0960-z (2019).
- [23] X. Han, *Convexity-preserving approximation by univariate cubic splines*, *J. Comput. Appl. Math.* **287**, 196–206 (2015).
- [24] T. Kim and B.I. Kvasov, *A shape-preserving approximation by weighted cubic splines*, *J. Comput. Appl. Math.* **236**, 4383–4397 (2012).
- [25] E. Kirsiaed, P. Oja and G.W. Shah, *Cubic spline histopolation*, *Mathematical modelling and analysis* **22**, 514–527 (2017).
- [26] Yu.S. Volkov, V.V. Bogdanov, V.L. Miroschnichenko and V.T. Shevaldin, *Shape-Preserving Interpolation by Cubic Splines*, *Math. Notes*, **88**, 798–805 (2010).
- [27] J. Wu and X. Zhang, *Integro quadratic spline interpolation*, *Appl. Math. Modell.* **39**, 2973–2980 (2015).
- [28] J. Wu and X. Zhang, *Integro sextic spline interpolation and its super convergence*, *Appl. Math. Comput.* **219**, 6431–6436 (2013).
- [29] Yu.S. Zavyalov, B.I. Kvasov and V.L. Miroschnichenko, *Methods of Spline Functions (in Russian)*, Nauka, Moscow (1980).
- [30] T. Zhanlav, *B-representation of interpolatory cubic splines (in Russian)*, *Vychislitel'nye Systemy*, Novosibirsk, **87**, 3–10 (1981).

- [31] T. Zhanlav and R. Mijiddorj, *The local integro cubic splines and their approximation properties*, Appl. Math. Comput. **216**, 2215–2219 (2010).
- [32] T. Zhanlav and R. Mijiddorj, *Convexity and monotonicity properties of the local integro cubic spline*, Appl. Math. Comput. **293**, 131–137 (2017).
- [33] T. Zhanlav and R. Mijiddorj, *Integro quintic splines and their approximation properties*, Appl. Math. Comput. **231**, 536–543 (2014).
- [34] T. Zhanlav and R. Mijiddorj, *On local integro quartic splines*, Appl. Math. Comput. **269**, 301–307 (2015).
- [35] A.A. Gusev, V.P. Gerdt, O. Chuluunbaatar, G. Chuluunbaatar, S.I. Vinitisky, V.L. Derbov, A. Gózdź, P.M. Krassovitskiy, *Symbolic-numerical algorithms for solving elliptic boundary-value problems using multivariate simplex lagrange elements*, Lecture Notes in Computer Science **11077**, pp. 197–213 (2018).
- [36] M. Kircher, F. Trinter, S. Grundmann, I. Vela-Perez, S. Brennecke, N. Eicke, J. Rist, S. Eckart, S. Houamer, O. Chuluunbaatar, Yu.V. Popov, I.P. Volobuev, K. Bagschik, M.N. Piancastelli, M. Lein, T. Jahnke, M.S. Schöffler and R. Dörner, *Kinematically complete experimental study of Compton scattering at helium atoms near the threshold*, Nature Physics **16**, pp. 756–760 (2020). (IF: 21.797)
- [37] P.W. Wen, O. Chuluunbaatar, A.A. Gusev, R.G. Nazmitdinov, A.K. Nasirov, S.I. Vinitisky, C.J. Lin, and H.M. Jia, *Near-barrier heavy-ion fusion: Role of boundary conditions in coupling of channels*, Phys. Rev. C **101**, pp. 014618–1–10 (2020). (IF: 2.988)
- [38] O. Chuluunbaatar, S. Obeid, B.B. Joulakian, A.A. Gusev, P.M. Krassovitskiy, L.A. Sevastianov, *D_{3h} symmetry adapted correlated three center wave functions of the ground and the first five excited states of H_3^+* , Chem. Phys. Lett. **746**, pp. 137304–1–8 (2020). (IF: 2.029)

Хавсралт

A New Family of Optimal Eighth-order Methods for Solving Nonlinear Equations

T. Zhanlav¹, Kh. Otgondorj^{2,*}

¹Institute of Mathematics, National University of Mongolia, Mongolia

²School of Applied Sciences, Mongolian University of Science and Technology, Mongolia

Abstract We propose a new family of optimal eight-order methods for solving nonlinear equations. The order of convergence of proposed methods verified using sufficient convergence conditions given in [6]. Using of sufficient convergence condition allows us to develop new optimal three-point iterations. Various numerical examples are considered to check the performance and to verify the theoretical results. Numerical comparisons of proposed methods with some existing methods are made. The test results are in good accordance with our study.

Keywords Nonlinear equations, Iterations methods, Optimal order of convergence

1. Introduction

At present, there are many optimal eighth-order methods for solving nonlinear equations (see, e.g., [1, 2, 3]). They require complicated convergence analysis that is feasible only by symbolic computation, although they produce high accuracy. The interest for these methods has renewed in recent years due to the rapid development of digital computers, advanced computer arithmetic and symbolic computation. In this note, we develop a family of three-point methods with optimal eighth-order convergence. In Section 2, the new family is developed and its convergence analysis is discussed. Unlike the usually considered convergence analysis here we first time used the sufficient conditions under which the three-point iteration have the eighth order of convergence [7, 8]. This allows to simplify the proof of theorem and to reduce tedious calculations. We also discussed similar theorems given by Sharma and Arora in [4, 5] and by Petrovic *et al* in [3]. The theoretical results proved in Section 2 are verified in Section 3 by considering various numerical examples. A comparison of the new methods with the existing methods is also given in this section.

2. The Family of Methods

Let x^* be a simple zero of the function $f(x): D \subset R \rightarrow R$ and x_0 be an initial approximation

to x^* . We consider the following simple three-point iteration

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ z_n &= \phi_4(x_n, y_n), \\ x_{n+1} &= z_n - \alpha_n \frac{f(z_n)}{f'(x_n)}, \quad n = 0, 1, \dots \end{aligned} \quad (1)$$

Here $\phi_4(x_n, y_n)$ is any two-point optimal fourth order scheme and α_n is given by the following formula

$$\begin{aligned} \alpha_n &= \frac{f'(x_n)}{(2f[z_n, y_n] - f[z_n, x_n]) \left(\frac{f'(x_n) - f[y_n, x_n]}{f'(x_n) + f[z_n, x_n]} + \right.} \\ &\quad \left. + \tau_n \frac{f[z_n, x_n]}{f'(x_n)} - 1 \right), \end{aligned} \quad (2)$$

where

$$f[r, s] = \frac{f(s) - f(r)}{s - r}. \quad (3)$$

Note that any two-point optimal fourth order iteration can be written as [8]:

$$z_n = y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n)}, \quad (4)$$

or

$$z_n = x_n - \tau_n \frac{f(x_n)}{f'(x_n)}, \quad (5)$$

* Corresponding author:

otgondorj@gmail.com (Kh. Otgondorj)

Published online at <http://journal.sapub.org/ajcam>

Copyright © 2018 Scientific & Academic Publishing. All Rights Reserved

where

$$\begin{aligned}\bar{\tau}_n &= \frac{\tau_n - 1}{\theta_n}, \quad \theta_n = \frac{f(y_n)}{f(x_n)}, \\ \tau_n &= 1 + \theta_n + 2\theta_n^2 + \beta\theta_n^3 + \gamma\theta_n^4 + \dots,\end{aligned}\quad (6)$$

with some constants β and γ .

Theorem 1 *Let the function $f(x)$ be sufficiently differentiable in a neighborhood of its simple zero x^* and $\phi_n(x_n, y_n)$ is an optimal fourth order method. If the initial approximation x_0 is sufficiently close to x^* , then the order of convergence of iteration (1), (2) is 8.*

Proof Using (1), (6) we have

$$\begin{aligned}f[x_n, y_n] &= (1 - \theta_n)f'(x_n), \\ f[y_n, z_n] &= \frac{1}{\bar{\tau}_n} f'(x_n)(1 - \nu_n), \\ f[x_n, z_n] &= \frac{1}{\tau_n} f'(x_n)(1 - \theta_n \nu_n),\end{aligned}\quad (7)$$

where

$$\nu_n = \frac{f(z_n)}{f(y_n)}.\quad (8)$$

Substituting (7) and (8) into (2) we get

$$\alpha_n = \frac{\tau_n \bar{\tau}_n}{(2\tau_n - \bar{\tau}_n)(1 - \frac{1 + \tau_n}{2\tau_n - \bar{\tau}_n} \nu_n)} \left(\frac{\tau_n}{\bar{\tau}_n(1 + \frac{\nu_n}{\bar{\tau}_n})} - \theta_n \nu_n \right),\quad (9)$$

Using the well-known expansion

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots,\quad (10)$$

One can write (9) as

$$\alpha_n = \frac{\tau_n^2}{2\tau_n - \bar{\tau}_n} (1 + (1 + 2\theta_n)\nu_n) + O(\theta_n^4).\quad (11)$$

Since

$$\frac{\tau_n^2}{2\tau_n - \bar{\tau}_n} = 1 + 2\theta_n + (\beta + 1)\theta_n^2 + (2\beta + \gamma - 4)\theta_n^3 + \dots,$$

then from (11), we obtain

$$\begin{aligned}\alpha_n &= (1 + 2\theta_n + (\beta + 1)\theta_n^2 + (2\beta + \gamma - 4)\theta_n^3 + \dots)(1 + (1 + 2\theta_n)\frac{f(z_n)}{f(y_n)} \\ &+ O(\theta_n^4)) = 1 + 2\theta_n + (\beta + 1)\theta_n^2 + (2\beta + \gamma - 4)\theta_n^3 + \\ &+ (1 + 2\theta_n)\frac{f(z_n)}{f(y_n)} + O(\theta_n^4),\end{aligned}\quad (12)$$

in which we have used $\frac{f(z_n)}{f(y_n)} = O(\theta_n^2)$. Then by

Theorem 1 and 2 in [7, 8] the order of convergence of (1) and (2) is 8.

It is often used two-point iterations (5) with functions [1-5]

$$\tau_n = \frac{1 + (\beta - 1)\theta_n + \beta\theta_n^2}{1 + (\beta - 2)\theta_n},\quad (13)$$

and

$$\tau_n = \frac{1 - \theta_n + \theta_n^2}{(1 - \theta_n)^2}.\quad (14)$$

In [10] it was developed optimal fourth-order method with parameter

$$\tau_n = \frac{1 - \sqrt{1 - 4\theta_n}}{2\theta_n}.\quad (15)$$

According to (15) we call the iteration (1), (2) with (15) the eighth-order iteration based on Zhanlav's fourth-order method. Similar theorems for (1) presented by Sharma and Arora in [4, 5] under choices

$$\alpha_n = \frac{f'(x)f[z_n, y_n]}{f[z_n, x_n](2f[z_n, y_n] - f[z_n, x_n])},\quad (16)$$

and

$$\alpha_n = \frac{f'(x) - f[y_n, x_n] + f[z_n, y_n]}{2f[z_n, y_n] - f[z_n, x_n]},\quad (17)$$

and by Petkovic *et al* in [3] under choice

$$\alpha_n = \frac{f'(x_n)}{2(f[x_n, z_n] - f[x_n, y_n]) + f[y_n, z_n] + (y_n - z_n)f[y_n, x_n, x_n]},\quad (18)$$

where

$$f[y_n, x_n, x_n] = \frac{f[y_n, x_n] - f'(x_n)}{y_n - x_n}.$$

Now we consider the three-point iterations (1) with α_n given by formula

$$\alpha_n = \sum_{i=1}^m \omega_i \alpha_n^i, \quad \omega_i \in \mathbb{R}, \quad \sum_{i=1}^m \omega_i = 1.\quad (19)$$

Here by α_n^i we denote any functions satisfying the condition (12). In particular, as α_n^i one can take functions (2), (16), (17) and (18). As before, using the sufficient convergence conditions (12) it is easy to prove that the convergence order of three-point iterations (1) with α_n given by (19) is 8.

3. Numerical Experiments

In this section, we have made some numerical experiments on our proposed method and given some numerical comparisons with existing optimal eighth order methods as various examples. We consider the following test functions used in [5, 6]:

$$f_1(x) = \sin^2(x) - x^2 + 1,$$

$$x^* \approx 1.404491648215$$

$$f_2(x) = xe^{x^2} - \sin^2(x) + 3\cos(x) + 5,$$

$$x^* \approx 1.207647827130$$

$$f_3(x) = \ln(x^2 + x + 2) - x + 1,$$

$$x^* \approx 4.152590736757$$

$$f_4(x) = (x-1)^6 - 1, \quad x^* = 2$$

All computations have been carried out using Maple 18 computer algebra system with 1500 significant digits and the

fixed stopping criterion $\varepsilon = 10^{-250}$. In Tables 1-4, α_n and τ_n with some parameters are considered in first and in second columns, respectively. The number of iterations N , the absolute value $|x_n - x^*|$ and the computational order of convergence (COC) are displayed in these Tables as well. To verify the theoretical order of convergence of our methods, we calculate the computational order of convergence using the formula [6]

$$\rho \approx \frac{\ln(|x_n - x^*| / |x_{n-1} - x^*|)}{\ln(|x_{n-1} - x^*| / |x_{n-2} - x^*|)}.$$

For a comparison, we employed the Sharma-Arora methods with function (16), (17) and method given in [3] with function (18). From Tables 1-4, we see that the COC perfectly coincides with theoretical order and the new method (1), (2) with function (15) is comparable with existing methods.

Table 1: Performance of methods as α_n chosen by (2)

α_n	τ_n	$f_1(x), x_0 = 1.2$		$f_2(x), x_0 = -1.3$			
		N	$ x^* - x_n $	COC	N	$ x^* - x_n $	COC
	(15)	3	1.483(-474)	8.00000	3	4.481(-439)	8.00000
	(14)	3	7.146(-422)	8.00000	3	3.594(-453)	8.00000
(2)	$\beta = 2$	3	2.952(-299)	8.00000	3	2.953(-382)	8.00000
	(13) $\beta = 1$	3	3.907(-402)	8.00000	3	8.460(-417)	8.00000
	$\beta = 0$	3	7.140(-487)	8.00000	3	1.257(-612)	8.00000
		$f_3(x), x_0 = 3.0$		$f_4(x), x_0 = 1.9$			
	(15)	3	3.155(-494)	7.99999	3	1.101(-339)	8.00000
	(14)	3	2.122(-431)	8.00000	3	3.749(-276)	8.00000
(2)	$\beta = 2$	3	2.390(-377)	8.00000	4	1.244(-1326)	8.00000
	(13) $\beta = 1$	3	2.357(-409)	8.00000	4	2.210(-1921)	8.00000
	$\beta = 0$	3	7.977(-456)	8.00000	3	3.021(-365)	8.00000

Table 2: Performance of methods as α_n chosen by (16)

α_n	τ_n	$f_1(x), x_0 = 1.2$		$f_2(x), x_0 = -1.3$			
		N	$ x^* - x_n $	COC	N	$ x^* - x_n $	COC
	(15)	3	7.289(-480)	8.00000	3	1.361(-500)	8.00000
	(14)	3	5.093(-409)	8.00000	3	8.654(-503)	8.00000
(16)	$\beta = 2$	3	6.557(-467)	7.99999	3	5.817(-470)	8.00000
	(13) $\beta = 1$	3	2.658(-397)	8.00000	3	1.311(-486)	8.00000
	$\beta = 0$	3	1.870(-433)	8.00000	3	2.294(-566)	8.00000
		$f_3(x), x_0 = 3.0$		$f_4(x), x_0 = 1.9$			
	(15)	3	1.369(-566)	7.99999	3	1.177(-380)	8.00000
	(14)	3	4.047(-539)	8.00000	3	8.423(-360)	8.00000
(16)	$\beta = 2$	3	1.754(-527)	8.00000	3	4.742(-262)	8.00000
	(13) $\beta = 1$	3	6.122(-532)	8.00000	3	7.752(-357)	8.00000
	$\beta = 0$	3	8.825(-550)	8.00000	3	1.013(-386)	8.00000

Table 3: Performance of methods as α_n chosen by (17)

α_n	τ_n	$f_1(x), x_0 = 1.2$			$f_2(x), x_0 = -1.3$		
		N	$ x^* - x_n $	COC	N	$ x^* - x_n $	COC
(17)	(15)	3	7.289(-480)	8.00000	3	3.414(-445)	8.00000
	(14)	3	5.093(-408)	8.00000	3	1.170(-437)	8.00000
	$\beta = 2$	3	6.557(-467)	7.99999	3	1.391(-371)	8.00000
	$\beta = 1$	3	2.658(-396)	8.00000	3	3.217(-406)	8.00000
	$\beta = 0$	3	1.870(-433)	8.00000	3	2.930(-546)	8.00000
		$f_3(x), x_0 = 3.0$			$f_4(x), x_0 = 1.9$		
(17)	(15)	3	6.255(-505)	8.00000	3	4.917(-339)	8.00000
	(14)	3	1.740(-443)	8.00000	4	3.787(-1936)	8.00000
	$\beta = 2$	3	1.017(-390)	8.00000	4	4.192(-1033)	8.00000
	$\beta = 1$	3	6.576(-422)	8.00000	4	1.241(-1515)	8.00000
	$\beta = 0$	3	2.260(-467)	8.00000	3	1.407(-324)	8.00000

Table 4: Performance of methods as α_n chosen by (18)

α_n	τ_n	$f_1(x), x_0 = 1.2$			$f_2(x), x_0 = -1.3$		
		N	$ x^* - x_n $	COC	N	$ x^* - x_n $	COC
(18)	(15)	3	1.583(-466)	8.00000	4	1.543(-476)	8.00000
	(14)	3	1.441(-335)	8.00000	4	2.105(-420)	8.00000
	$\beta = 2$	4	1.617(-1945)	7.99999	3	7.034(-363)	8.00000
	$\beta = 1$	3	1.978(-293)	8.00000	3	2.930(-393)	8.00000
	$\beta = 0$	3	3.409(-402)	8.00000	3	2.356(-506)	8.00000
		$f_3(x), x_0 = 3.0$			$f_4(x), x_0 = 1.9$		
(18)	(15)	3	7.134(-489)	8.00000	3	1.913(-353)	8.00000
	(14)	3	4.392(-461)	8.00000	3	4.134(-252)	8.00000
	$\beta = 2$	3	6.107(-393)	8.00000	4	4.158(-1133)	8.00000
	$\beta = 1$	3	5.661(-421)	8.00000	4	3.632(-1597)	8.00000
	$\beta = 0$	4	4.392(-461)	8.00000	3	2.717(-330)	8.00000

4. Conclusions

In this paper, a new family of optimal eight-order methods for solving nonlinear equations is introduced and studied. This family (1), (19) includes the three-point methods given by Sharma and Arora in [4, 5] and by Petkovic *et al* in [3] and our proposed method (1), (2) as particular cases. Finally, the theoretical proofs and numerical experiments have shown that new iterative method is of eight-order and effective.

ACKNOWLEDGEMENTS

This work was supported by the Foundation of Science and Technology of Mongolian under grant SST_007/2015.

REFERENCES

[1] C. Chun, B. Neta, Comparative study of eighth-order methods for finding simple roots of nonlinear equations, *Numer. Algor.* 74 (2017) 1169-1201.

[2] Chun, C. & Neta, B. On the new family of optimal eighth order methods developed by Lotfi *et al.* *Numer Algor* (2016) 72: 363.

[3] M.S. Petkovic, B. Neta, L.D. Petkovic, J. Dzunic, *Multipoint Methods for Solving Nonlinear Equations*, Elsevier, 2013.

[4] M.S. Petkovic, B. Neta, L.D. Petkovic, J. Dzunic, *Multipoint Methods for Solving Nonlinear Equations*, A survey. *Appl. Math. Comput.* 226 (2014) 635-660.

[5] J.R. Sharma, H. Arora, A new family of optimal eighth order methods with dynamics for nonlinear equations, *Appl. Math. Comput.* 273 (2016) 924-933.

[6] J.R. Sharma, H. Arora, An efficient family of weighted-Newton methods with optimal eighth order convergence, *Appl. Math. Lett.* 29 (2014) 1-6.

[7] X. Wang, L. Liu, Modified Ostrowski's method with eighth-order convergence and high efficiency index, *Appl. Math. Lett.* 23 (2010) 549-554.

[8] T. Zhanlav, O. Chuluunbaatar, V. Ulziibayar, Generating function method for constructing new iterations, *Appl. Math. Comput.* 315 (2017) 414-423.

[9] T. Zhanlav, O. Chuluunbaatar, V. Ulziibayar, The necessary and sufficient conditions for some two and three-point

- Newton's type iterations, *Comput Math. Math. Phys.* 57 (2017) 1090--1100.
- [10] T. Zhanlav, V. Ulziibayar, Modified King's Methods with Optimal Eighth-order of Convergence and High Efficiency Index, *American Journal of Comput and Applied Math.*6(5) (2016) 177-181.
- [11] T. Zhanlav, O. Chuluunbaatar, V. Ulziibayar, Accelerating the convergence of Newton-type iterations, *J. Numerical analysis and approximation theory.* 46 (2) (2017) 162–180.

Families of Optimal Derivative-Free Two- and Three-Point Iterative Methods for Solving Nonlinear Equations

T. Zhanlav^{a,*}, Kh. Otgondorj^{b,**}, and O. Chuluunbaatar^{a,c,***}

^a Institute of Mathematics, National University of Mongolia, Ulan-Bator, 14201 Mongolia

^b Division of Applied Sciences, Mongolian University of Science and Technology, Ulan-Bator, 14191 Mongolia

^c Joint Institute for Nuclear Research, Dubna, Moscow oblast, 141980 Russia

*e-mail: tzhanlav@yahoo.com

**e-mail: otgondorj@gmail.com

***e-mail: chuka@jinr.ru

Received September 9, 2018; revised January 16, 2019; accepted February 8, 2019

Abstract—Necessary and sufficient conditions for derivative-free two- and three-point iterative methods to have the optimal convergence order are obtained. These conditions can be effectively used not only for determining the order of convergence of iterative methods but also for designing new methods. Furthermore, the use of the method of generating functions makes it possible to construct a wide class of optimal derivative-free two- and three-point methods that includes many well-known methods as particular cases. An analytical formula for the optimal choice of the parameter of iterations improving the order of convergence is derived.

Keywords: nonlinear equations, two- and three-point iterations, necessary and sufficient conditions, optimal methods

DOI: 10.1134/S0965542519060149

1. INTRODUCTION

Presently, there are a lot of iterative methods for solving nonlinear equations and systems of equations (see [1–6]). Among them, there are derivative-free methods, which are helpful if the derivative of the function is difficult or impossible to calculate. The simplest of them are the well-known secant method and Steffensen's method, which have a low order of convergence. Nowadays, we need new optimal methods with the eighth order of convergence because their index of efficiency is $8^{1/4} \approx 1.682$. Such methods have applications in experimental mathematics, number theory, high energy physics, nonlinear simulation, finite element methods used in CAD, 3D graphics, statistics, security, and cryptography (see [7–9]). In the last decade, various derivative-free two- and three-point methods having good convergence properties have been developed (e.g., see [1–33]). The construction of iterative methods with a high order of convergence became possible due to the rapid progress in computing, computer arithmetic, and symbolic computations. In this paper, we propose some families of derivative-free methods based on the method of generating functions proposed in [5] and on the optimal choice of parameters of iterations [6]. A novel direct approach to proving the order of convergence of such methods that does not use symbolic computations is proposed.

The paper is organized as follows. In Section 2, we consider derivative-free two-point iterative methods and obtain necessary and sufficient conditions for these methods to have the fourth order of convergence. The choice of generating functions for the iteration parameter τ is discussed. In particular, optimal finite difference versions of the well-known Kung–Traub, King, and Maheshwari methods are obtained. In Section 3, we consider derivative-free three-point iterative methods and obtain necessary and sufficient conditions for these methods to have the eighth order of convergence. A wide class of optimal three-point iterative methods that includes many known methods as its special cases is proposed. The local convergence of these methods is proved without using symbolic computations. Section 4 presents the results of numerical computations confirming the theoretical results concerning the order of convergence, and these results are compared with the results obtained using other methods.

2. DERIVATIVE-FREE TWO-POINT ITERATIVE METHODS

Consider the derivative-free two-point iterative method

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \tag{2.1a}$$

$$x_{k+1} = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi(x_k)}, \tag{2.1b}$$

where

$$f'(x) \approx \phi(x) = \frac{f(x + \gamma f(x)) - f(x)}{\gamma f(x)}, \quad \gamma \in R, \tag{2.2}$$

γ is a free nonzero parameter, and $\bar{\tau}_k$ is a parameter to be determined. Here the function $\phi(x) \equiv \phi(x, \gamma)$ depends not only on x but also on the parameter γ ; by the definition of derivative, we have

$$f'(x) = \phi(x, \gamma), \quad \gamma \rightarrow 0. \tag{2.3}$$

To determine the order of convergence of the iterative method (2.1a), (2.1b), define

$$w_k = \frac{f'(x_k)}{\phi(x_k)} \neq 0. \tag{2.4}$$

Let $f(x) \in C^3(I)$, where I is an interval containing the root x^* of the equation $f(x) = 0$. Then, the Taylor expansions of the functions $f(y_k)$ and $f(x_k + \gamma f(x_k))$ give

$$f(y_k) = (1 - w_k)f(x_k) + \frac{f''(x_k)}{2} \left(\frac{f(x_k)}{f'(x_k)} \right)^2 w_k^2 + O(f^3(x_k)), \tag{2.5}$$

$$\phi(x_k) = f'(x_k) \left(1 + \gamma \frac{f''(x_k) f(x_k)}{2 f'(x_k)} \right) + O(f^2(x_k)). \tag{2.6}$$

Substitute (2.6) into (2.4) to obtain

$$w_k = \frac{1}{1 + \gamma \frac{f''(x_k) f(x_k)}{2 f'(x_k)}} + O(f^2(x_k)) = 1 - \gamma \frac{f''(x_k) f(x_k)}{2 f'(x_k)} + O(f^2(x_k)), \tag{2.7}$$

or

$$w_k = 1 + O(f(x_k)). \tag{2.8}$$

Taking into account (2.8), we have in (2.5)

$$f(y_k) = O(f^2(x_k)). \tag{2.9}$$

As in [6], we use the notation

$$\theta_k = \frac{f(y_k)}{f(x_k)}. \tag{2.10}$$

Formulas (2.9) and (2.10) imply that $\theta_k = O(f(x_k))$. Using (2.5) in (2.10), we obtain

$$\theta_k = 1 - w_k + \frac{1}{2} w_k \frac{f''(x_k) f(x_k)}{f'(x_k) \phi(x_k)} + O(f^2(x_k)). \tag{2.11}$$

By eliminating $\frac{f''(x_k) f(x_k)}{f'(x_k)}$ from (2.7) and (2.11), we obtain

$$w_k^2 - (1 - \gamma \phi(x_k)) w_k - (1 - \theta_k) \gamma \phi(x_k) = O(f^2(x_k)). \tag{2.12}$$

It is seen from (2.12) that w_k depends on θ_k . Due to (2.8), we may seek w_k in the form

$$w_k = 1 - a_k \theta_k + O(f^2(x_k)). \tag{2.13}$$

By substituting (2.13) into (2.12), we obtain

$$\theta_k(\gamma\phi_k - a_k(1 + \gamma\phi_k)) = O(f^2(x_k)), \quad \phi_k = \phi(x_k). \quad (2.14)$$

Now (2.14) implies that

$$a_k = \frac{\gamma\phi_k}{1 + \gamma\phi_k} + O(f(x_k)). \quad (2.15)$$

By substituting (2.15) into (2.13), we obtain

$$w_k = 1 - \frac{\gamma\phi_k}{1 + \gamma\phi_k}\theta_k + O(f^2(x_k)). \quad (2.16)$$

On the other hand, the Taylor expansion of $f(x_{k+1})$ gives

$$f(x_{k+1}) = \left(1 - \frac{f'(y_k)}{\phi_k}\bar{\tau}_k\right)f(y_k) + O(f(y_k)^2). \quad (2.17)$$

Due to (2.1a), we have

$$f'(y_k) = f'(x_k)\left(1 - \frac{f''(x_k)f(x_k)}{f'(x_k)\phi(x_k)}\right) + O(f^2(x_k)). \quad (2.18)$$

The elimination of the term $\frac{f''(x_k)f(x_k)}{f'(x_k)\phi(x_k)}$ from (12) and (19) yields

$$f'(y_k) = -f'(x_k)\frac{w_k + 2(\theta_k - 1)}{w_k} + O(f^2(x_k)). \quad (2.19)$$

By substituting w_k given by (2.16) into (2.19) and using the expansion

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots, \quad |x| < 1, \quad (2.20)$$

we obtain

$$f'(y_k) = f'(x_k)\left(1 - \frac{2}{1 + \gamma\phi_k}\theta_k\right) + O(f^2(x_k)). \quad (2.21)$$

Using (2.21) in (2.17), we have

$$f(x_{k+1}) = (1 - (1 - \hat{d}_k\theta_k)\bar{\tau}_k)f(y_k) + O(f(y_k)^2), \quad \hat{d}_k = \frac{2 + \gamma\phi_k}{1 + \gamma\phi_k}. \quad (2.22)$$

Now we can prove the following result.

Theorem 1. *Let $f(x) \in C^3(I)$, and let the initial approximation x_0 be sufficiently close to the simple root $x^* \in I$ of the function $f(x)$. Then, the iterative method (2.1) has the fourth order of convergence if and only if the parameter $\bar{\tau}_k$ in (2.1) satisfies the condition*

$$\bar{\tau}_k = \frac{1}{1 - \hat{d}_k\theta_k} + O(f^2(x_k)) = 1 + \hat{d}_k\theta_k + O(f^2(x_k)). \quad (2.23)$$

Proof. Suppose that $\bar{\tau}_k$ in (2.1) satisfies condition (2.23). Then

$$1 - (1 - \hat{d}_k\theta_k)\bar{\tau}_k = O(f^2(x_k)),$$

and $f(y_k) = O(f^2(x_k))$ due to (2.8). Therefore, due to (2.22) we have

$$f(x_{k+1}) = O(f(x_k)^4); \quad (2.24)$$

i.e., the order of convergence of (2.1) is four under condition (2.23). Conversely, let method (2.1) have the fourth order of convergence, i.e., let (2.24) hold. Then, (2.24) and (2.22) imply that $f(y_k) = O(f^2(x_k))$ and $1 - (1 - \hat{d}_k\theta_k)\bar{\tau}_k = O(f^2(x_k))$; i.e., $\bar{\tau}_k$ satisfies condition (2.23).

The iterative method (2.4) uses $f(x_k), f(y_k)$, and $\phi(x_k)$ at each iteration step; therefore, it is optimal in the sense of the Kung–Traub conjecture. The second step in (2.1) can be rewritten as

$$x_{k+1} = x_k - \tau_k \frac{f(x_k)}{\phi(x_k)}, \tag{2.25}$$

where

$$\tau_k = 1 + \bar{\tau}_k \theta_k = 1 + \theta_k + \hat{d}_k \theta_k^2 + O(f^3(x_k)). \tag{2.26}$$

If $\phi(x_k, \gamma) = f'(x_k)$ as $\gamma \rightarrow 0$, then $w_k = 1$, and formulas (2.23) and (2.26) take the form

$$\begin{aligned} \bar{\tau}_k &= 1 + 2\theta_k + O(f^2(x_k)), \\ \tau_k &= 1 + \theta_k + 2\theta_k^2 + O(f^3(x_k)), \end{aligned}$$

respectively. Thus, the iterative method (2.1) has the form

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}, \quad x_{k+1} = x_k - \tau_k \frac{f(x_k)}{f'(x_k)}; \tag{2.27}$$

therefore, it is an optimal fourth-order two-point iterative method [6]. As in [5], the generation function method can be applied for constructing new iterative methods (2.1). Certainly, there are various versions of the generation functions $\bar{\tau}_k = H(\theta_k)$ satisfying the conditions

$$H(0) = 1, \quad H'(0) = \hat{d}_k. \tag{2.28}$$

In this paper, we consider the simple form

$$H(x) = \frac{c + (\hat{d}_k c + d)x + \omega x^2}{c + dx + bx^2}, \quad c + d + b \neq 0, \quad c, d, b, \omega \in R. \tag{2.29}$$

We consider some interesting special cases of H .

1. Let $c = 1, d = \beta - 2$, and $b = \omega = 0$ in (2.29). Then, we obtain

$$H(x) = \frac{1 + \left(\beta - \frac{\gamma \phi_k}{1 + \gamma \phi_k} \right) x}{1 + (\beta - 2)x}.$$

The iterative method (2.1) with $\bar{\tau}_k = H(\theta_k)$ has the form

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \quad x_{k+1} = y_k - \frac{1 + \left(\beta - \frac{\gamma \phi_k}{1 + \gamma \phi_k} \right) \theta_k}{1 + (\beta - 2)\theta_k} \frac{f(y_k)}{\phi(x_k)}. \tag{2.30}$$

As $\gamma \rightarrow 0$, (2.30) gives the well-known King method. We call (2.30) the finite difference version of the King method.

2. Let $c = b = 1, d = -2$, and $\omega = 0$ in (2.29). Then, we obtain

$$H(x) = \frac{1 - \frac{\gamma \phi_k}{1 + \gamma \phi_k} x}{(1 - x)^2}.$$

The iterative method (2.1) with $\bar{\tau}_k = H(\theta_k)$ has the form

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \quad x_{k+1} = y_k - \frac{1 - \frac{\gamma \phi_k}{1 + \gamma \phi_k} \theta_k}{(1 - \theta_k)^2} \frac{f(y_k)}{\phi(x_k)}. \tag{2.31}$$

As $\gamma \rightarrow 0$, (2.31) gives the well-known fourth-order Kung–Traub method. For this reason, we call (2.31) the finite difference version of the Kung–Traub method.

3. Let $c = 1$, $\omega = d = -1$, and $b = 0$ in (2.29). Then, we obtain

$$H(x) = \frac{1 + \frac{1}{1 + \gamma\phi_k}x - x^2}{1 - x}.$$

The iterative method (2.1) with $\bar{\tau}_k = H(\theta_k)$ has the form

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \quad x_{k+1} = y_k - \frac{1 + \frac{1}{1 + \gamma\phi_k}\theta_k - \theta_k^2}{1 - \theta_k} \frac{f(y_k)}{\phi(x_k)}. \quad (2.32)$$

As $\gamma \rightarrow 0$, (2.32) gives the Maheshwari method. For this reason, we call (2.32) the finite difference version of the Maheshwari method.

Note that an attempt to construct derivative-free versions of the Kung and Traub methods was made in [30]. However, the method obtained in [30] differs from our extensions (2.30) and (2.31).

Thus, using the generating function method, we obtain a wide class of optimal derivative-free two-point methods (2.1) with $\bar{\tau}_k = H(\theta_k)$ specified by (2.29). This class has five parameters $(\gamma, c, d, b, \omega)$. The coefficients in (2.29) can depend on the iteration index k . Note that many derivative-free two-point methods were constructed in [1, 2, 7, 14–18]. The class of iterative methods (2.1) proposed in this paper, which is specified by formula (2.29) with the parameter $\bar{\tau}_k$, includes some well-known iterative methods as special cases. Some of them are listed in Table 1. Only $\bar{\tau}_k$ in Ren's method [16, 34] does not belong to the class $H(\theta_k)$ given by (2.29). Thus, the proposed family (2.1) with the parameter specified by (2.29) is a considerable generalization of the methods described in [2, 7, 9, 11–18, 20–23, 26, 27].

The two-point iterative method (2.1) includes one free nonzero parameter γ . It is well known that the convergence can be accelerated by a proper variation of the free parameter $\gamma = \gamma_k$ at each iteration step. This approach is helpful for constructing high order iterative methods with memory (see [9, 22, 23, 25]). We now try to find the optimal free parameter from the accuracy viewpoint. Consider the Taylor expansion of the function $f(\eta_k) = f(x_k + \gamma f(x_k))$ in the neighborhood of x_k

$$f(\eta_k) = (1 + \gamma f'(x_k))f(x_k) + \frac{f''(x_k)}{2} \gamma^2 f^2(x_k) + O(f^3(x_k)). \quad (2.33)$$

Hence, we see that at each step γ can be chosen as

$$\gamma_k = -\frac{1}{f'(x_k)}. \quad (2.34)$$

Then, (2.33) takes the form

$$f(\eta_k) = \frac{f''(x_k)}{2} \frac{f^2(x_k)}{f'(x_k)^2} + O(f^3(x_k)), \quad (2.35)$$

where

$$\eta_k = x_k - \frac{f(x_k)}{f'(x_k)}. \quad (2.36)$$

Therefore, due to (2.35) η_k specified by formula (2.36) can be considered as a new approximation that is better than x_k (2.35). Taking into account (2.34) and (2.35), formulas (2.5) and (2.7) can be written as

$$f(y_k) = (1 - w_k)f(x_k) + f(\eta_k)w_k^2 + O(f^3(x_k)), \quad (2.37)$$

$$w_k = 1 + \frac{f(\eta_k)}{f(x_k)} + O(f^2(x_k)), \quad (2.38)$$

respectively. Substitute (2.38) into (2.37) and use (2.35) to obtain

$$f(y_k) = O(f^3(x_k)). \quad (2.39)$$

Table 1. Iterative methods

	Methods	$\bar{\tau}_k$	Special cases of H determined in (30)
1.	Methods described in [18, 20] and $h(t, s) = (1 + t)(1 + s)$ [2], $P1$ in [21]	$1 + \hat{d}_k \theta_k$	$c \neq 0, d = b = \omega = 0$
2.	Methods described in [10, 15, 12] and $h(t, s) = \frac{1}{1 - t - s}$ [2], CTM in [26]	$\frac{1}{1 - \hat{d}_k \theta_k}$	$c = 1, d = -\hat{d}_k, b = \omega = 0$
3.	$h(t, s) = \frac{1+t}{1-s}$ [2, 34]	$\frac{1 + \gamma\phi_k + (\hat{d}_k(1 + \gamma\phi_k) - 1)\theta_k}{1 + \gamma\phi_k - \theta_k}$	$c = 1 + \gamma\phi_k, d = -1, b = \omega = 0$
4.	Methods described in [7]	$\frac{1 + \phi_k + (\hat{d}_k(1 + \phi_k) - (2 - \phi_k))\theta_k}{1 + \phi_k - (2 + \phi_k)\theta_k}$	$c = 1 + \phi_k, d = -(2 + \phi_k), b = -1, \omega = 0$
5.	Chebyshev–Halley family [9]	$\frac{1 + \left(\hat{d}_k - \left(2\alpha + \frac{1}{1 + \gamma\phi_k}\right)\right)\theta_k + \omega\theta_k^2}{1 - \left(2\alpha + \frac{1}{1 + \gamma\phi_k}\right)\theta_k + \frac{2\alpha}{1 + \gamma\phi_k}\theta_k^2}$	$c = 1, d = -\left(2\alpha + \frac{1}{1 + \gamma\phi_k}\right), b = \frac{2\alpha}{1 + \gamma\phi_k}, \omega = H(\theta_k)$
6.	Kung–Traub’s method and method in [11]	$\frac{1}{1 - d_k \theta_k + \frac{1}{1 + \gamma\phi_k} \theta_k^2}$	$c = 1, d = -\hat{d}_k, b = \frac{1}{1 + \gamma\phi_k}, \omega = 0$
7.	Potra–Ptak’s method [13, 22]	$1 + d_k \theta_k + \frac{d_k a}{2} \theta_k^2$	$c = 1, d = b = 0, \omega = \frac{\hat{d}_k a}{2}$
8.	King-type method [23]	$\frac{1 + (\gamma - 1)\theta_k - \gamma\theta_k^2}{1 + \left(\gamma - 2 - \frac{1}{1 - \beta\phi_k}\right)\theta_k + \frac{\gamma - 2}{1 - \beta\phi_k}\theta_k^2}$	$c = 1, d = \gamma - 2 - \frac{1}{1 - \beta\phi_k}, b = \frac{2 - \gamma}{1 - \beta\phi_k}, a = -\gamma$
9.	Methods described in [17]	$\frac{1 - \frac{\gamma\phi_k}{1 + \gamma\phi_k} \theta_k}{1 - 2\theta_k + \theta_k^2}$	$c = b = 1, d = -2, \omega = 0$
10.	Methods described in [16]	$\frac{1}{1 - \hat{d}_k \theta_k + a \frac{1 + \phi_k}{\phi_k} f^2(x_k)}$	Does not belong to (30)
11.	Methods $P2$ described in [21]	$\frac{1 + \theta_k}{1 - \frac{\theta_k}{1 + \gamma\phi_k}}$	$c = 1, d = -\frac{1}{1 + \gamma\phi_k}, b = \omega = 0$
12.	Methods described in [27]	$\frac{1 + \phi_k + 2\theta_k}{1 + \phi_k - \phi_k \theta_k}$	$c = 1 + \phi_k, d = -\phi_k, b = \omega = 0$
13.	Methods described in [16]	$1 + \hat{d}_k \theta_k + \left(\alpha_1 + \frac{\alpha_2}{(1 + \gamma\phi_k)^2}\right)\theta_k^2$	$c = 1, d = b = 0, \omega = \alpha_1 + \frac{\alpha_2}{(1 + \gamma\phi_k)^2}$

Let the parameter $\bar{\tau}_k$ be chosen by the formula

$$\bar{\tau}_k = \frac{1}{1 - \hat{d}_k \theta_k} = 1 + \hat{d}_k \theta_k + \hat{d}_k^2 \theta_k^2 + O(f^3(x_k)). \tag{2.40}$$

Using (2.39) and (2.40) in (2.22), we obtain

$$f(x_{k+1}) = O(f(x_k)^6). \tag{2.41}$$

This implies that the choice of the variable parameter (2.34) significantly accelerates the two-point method (2.1). The order of convergence increases from two to six. In this case, method (2.1) actually is a three-point one, i.e.,

$$\eta_k = x_k - \frac{f(x_k)}{f'(x_k)}, \quad y_k = x_k - \frac{f(x_k)}{\phi_k}, \quad x_{k+1} = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi_k}. \tag{2.42}$$

If we replace $\gamma_k = -\frac{1}{f'(x_k)} \approx -\frac{1}{N_3'(x_k)}$, then we obtain a two-point iterative method with memory (x_0 and γ_0 are given). Then $\eta_0 = x_0 + \gamma_0 f(x_0)$ and

$$\begin{aligned} \gamma_k &= -\frac{1}{N_3'(x_k)}, & \eta_k &= x_k + \gamma_k f(x_k), & k &= 1, 2, \dots, \\ y_k &= x_k - \frac{f(x_k)}{\phi_k}, & x_{k+1} &= y_k - \bar{\tau}_k \frac{f(y_k)}{\phi_k}, & \phi_k &= \phi(x_k, \gamma_k). \end{aligned} \quad (2.43)$$

Here, $N_3(t) = N_3(t, x_k, x_{k-1}, y_{k-1}, \eta_{k-1})$ is the Newton cubic interpolation polynomial specified by the node points x_k, x_{k-1}, y_{k-1} , and η_{k-1} [9], [23]. It is clear that the order R of methods (2.43) is at least six.

Note that sometimes asymmetric derivative-free iterations that require additional computations were used. For example, in [33] the optimal iterative families of the King type were proposed:

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi_k}, & \phi_k &= f[x_k, \eta_k], & \eta_k &= x_k + \gamma f(x_k), \\ x_{k+1} &= y_k - \frac{f(y_k)}{f[y_k, \eta_k]} \frac{1 + \beta \theta_k}{1 + (\beta - 1)\theta_k}, \end{aligned} \quad (2.44)$$

where $f[x_k, \eta_k]$ is the first divided difference. The second substep in (2.44) can be written as

$$x_{k+1} = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi_k}, \quad (2.45)$$

where

$$\bar{\tau}_k = \frac{1 + \beta \theta_k}{1 + (\beta - \hat{d}_k)\theta_k - \frac{(\beta - 1)\theta_k^2}{1 + \gamma \phi_k}}; \quad (2.46)$$

i.e., in this case $\bar{\tau}_n$ is determined by a more complicated formula than in (2.30). Moreover, as $\gamma \rightarrow 0$, we have

$$\bar{\tau}_k \rightarrow \frac{1 + \beta \theta_k}{1 + (\beta - 2)\theta_k - (\beta - 1)\theta_k^2},$$

while

$$\bar{\tau}_k \rightarrow \frac{1 + \beta \theta_k}{1 + (\beta - 2)\theta_k}$$

in (2.30). Another example of the finite difference version of the optimal Hansen–Patrick family proposed in [32] can be written as

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi_k + \lambda f(\eta_k)}, & \eta_k &= x_k + \gamma f(x_k), & \gamma, \lambda &\in R \setminus \{0\}, \\ x_{k+1} &= y_k - \frac{f(y_k)}{f[y_k, \eta_k] + \lambda f(\eta_k)} \bar{\tau}_k, \end{aligned} \quad (2.47)$$

where

$$\begin{aligned} \bar{\tau}_k &= \frac{1}{\theta_k} \left(-1 + \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_k}} \right) H(\theta_k), & \alpha &\neq -1, \\ H(0) &= 1, & H'(0) &= -\frac{\alpha + 1}{2}, & |H''(0)| &< \infty. \end{aligned} \quad (2.48)$$

Remove the asymmetry in (2.47) and consider the iterative method

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{\phi_k + \lambda f(\eta_k)}, \\ x_{k+1} &= y_k - \frac{f(y_k)}{\phi_k + \lambda f(\eta_k)} \bar{\tau}_k, \end{aligned} \tag{2.49}$$

where $\bar{\tau}_k$ as before is determined by formula (2.48). The following result is easy to prove.

Theorem 2. *Let $f(x) \in C^3(I)$ have a simple root $x^* \in I$. If the initial approximation x_0 is sufficiently close to $x^* \in I$, then the iterative method (2.49) has the optimal fourth order of convergence if*

$$H(0) = 1, \quad H'(0) = \hat{d}_k - \frac{\alpha + 3}{2}, \quad |H''(0)| < \infty. \tag{2.50}$$

Proof. Assume that $H(0) = a$ and $H'(0) = b$. Then, we obtain from (2.48) that

$$\begin{aligned} \bar{\tau}_k &= \left(1 + \frac{\alpha + 3}{2} \theta_k + \left(\frac{(\alpha + 1)^2}{2} + \alpha + 2 \right) \theta_k^2 + \dots \right) (a + b\theta_k + O(f^2(x_k))) \\ &= a + \left(\frac{\alpha + 3}{2} a + b \right) \theta_k + O(f^2(x_k)). \end{aligned}$$

By comparing this with the sufficient convergence condition (2.23), we conclude that

$$a = 1, \quad \frac{\alpha + 3}{2} + b = \hat{d}_k \rightarrow b = \hat{d}_k - \frac{\alpha + 3}{2}.$$

Therefore, by Theorem 1, the iterative method (2.49) has the fourth order of convergence under condition (2.50).

3. DERIVATIVE-FREE THREE-POINT ITERATIVE METHODS

Consider the derivative-free three-point methods

$$y_k = x_k - \frac{f(x_k)}{\phi(x_k)}, \quad z_k = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi(x_k)}, \quad x_{k+1} = z_k - \alpha_k \frac{f(z_k)}{\phi(x_k)}, \tag{3.1}$$

which are obtained from the three-point methods studied in [6] by replacing $f'(x_k)$ with $\phi(x_k)$. Note that the first two steps in (3.1) determine optimal two-point fourth-order methods if $\bar{\tau}_k$ is given by (2.23). Our aim is to find α_k such that the order of convergence of iterations (3.1) is eight. To this end, we use the Taylor expansion of $f(x_{k+1})$:

$$\begin{aligned} f(x_{k+1}) &= f(z_k) - f'(z_k) \alpha_k \frac{f(z_k)}{\phi(x_k)} + O(f(z_k)^2) \\ &= \left(1 - \alpha_k \frac{f'(z_k)}{\phi(x_k)} \right) f(z_k) + O(f^2(z_k)). \end{aligned} \tag{3.2}$$

This implies that

$$f(x_{k+1}) = O(f^8(x_k)) \tag{3.3}$$

under the condition

$$\alpha_k = \frac{\phi(x_k)}{f'(z_k)} + O(f^4(x_k)). \tag{3.4}$$

Now we approximate $f'(z_k)$ in (3.4) using $f(x_k)$, $f(y_k)$, $f(z_k)$, and $\phi(x_k)$ such that

$$f'(z_k) = a_k f(x_k) + b_k f(y_k) + c_k f(z_k) + d_k \phi(x_k) + O(f(x_k)^4). \tag{3.5}$$

Using the Taylor expansion of $f(x)$ about the point z_k , we obtain the system of equations

$$\begin{aligned} a_k + b_k + c_k &= 0, \\ a_k w_k + b_k \gamma_k + d_k &= 1, \\ a_k w_k^2 + b_k \gamma_k^2 + 2d_k \left(w_k + \frac{1}{2} \gamma f(x_k) \right) &= 0, \\ a_k w_k^3 + b_k \gamma_k^3 + 3d_k \left(w_k^2 + w_k \gamma f(x_k) + \frac{1}{3} \gamma^2 f^2(x_k) \right) &= 0, \end{aligned} \quad (3.6)$$

where

$$w_k = x_k - z_k, \quad \gamma_k = y_k - z_k. \quad (3.7)$$

System (3.6) has the unique solution

$$\begin{aligned} c_k &= -a_k - b_k, \\ d_k &= 1 - a_k w_k - b_k \gamma_k, \\ b_k &= \frac{w_k(w_k + \gamma f(x_k))}{\gamma_k(\gamma_k - w_k)(\gamma_k - w_k - \gamma f(x_k))}, \\ a_k &= \frac{\gamma_k}{w_k(\gamma_k - w_k)} \frac{(w_k - \gamma_k)(2w_k + \gamma f(x_k)) + (w_k + \gamma f(x_k))^2}{(w_k + \gamma f(x_k))(w_k - \gamma_k + \gamma f(x_k))}. \end{aligned} \quad (3.8)$$

Substitute (3.8) into (3.5) to obtain

$$f'(z_k) = \phi_k \left(1 + a_k w_k \left(\frac{f[z_k, x_k]}{\phi_k} - 1 \right) + b_k \gamma_k \left(\frac{f[z_k, y_k]}{\phi_k} - 1 \right) \right) + O(f^4(x_k)), \quad (3.9)$$

where

$$\phi_k = \phi(x_k) = f[x_k, \xi_k], \quad \xi_k = x_k + \gamma f(x_k).$$

According to (3.2) and (3.7), we have

$$w_k = \frac{f(x_k)}{\phi_k} \tau_k, \quad \gamma_k = (\tau_k - 1) \frac{f(x_k)}{\phi_k}, \quad \gamma_k - w_k = -\frac{f(x_k)}{\phi_k} = y_k - x_k, \quad (3.10)$$

$$\frac{\gamma_k}{w_k - \gamma_k} = \frac{y_k - z_k}{x_k - y_k} = \tau_k - 1 \rightarrow \tau_k = \frac{x_k - z_k}{x_k - y_k}, \quad (3.11)$$

$$w_k + \gamma f_k = \frac{f(x_k)}{\phi_k} (\tau_k + \gamma \phi_k), \quad w_k - \gamma_k + \gamma f(x_k) = \frac{f(x_k)}{\phi_k} (1 + \gamma \phi_k). \quad (3.12)$$

Using (3.10)–(3.12) in (3.8), we conclude that

$$b_k \gamma_k = \frac{\tau_k (\tau_k + \gamma \phi_k)}{1 + \gamma \phi_k}, \quad a_k w_k = (1 - \tau_k) \frac{2\tau_k + \gamma \phi_k + (\tau_k + \gamma \phi_k)^2}{(\tau_k + \gamma \phi_k)(1 + \gamma \phi_k)}. \quad (3.13)$$

Substitute (3.9) into (3.4) and neglect the small term $O(f^4(x_k))$ to find that

$$\alpha_k = \frac{1}{1 + a_k w_k \left(\frac{f[z_k, x_k]}{\phi_k} - 1 \right) + b_k \gamma_k \left(\frac{f[z_k, y_k]}{\phi_k} - 1 \right)}, \quad (3.14)$$

where $a_k w_k$ and $b_k \gamma_k$ are determined by formula (3.13). The expressions in parentheses in (3.14) can be rewritten in terms of the second divided differences as

$$\frac{f[z_k, x_k]}{\phi_k} - 1 = \frac{1}{\phi_k} f[z_k, x_k, \xi_k] (z_k - \xi_k) = -\frac{f(x_k)}{\phi_k^2} f[z_k, x_k, \xi_k] (\tau_k + \gamma \phi_k), \quad (3.15)$$

$$\frac{f[z_k, y_k]}{\phi_k} - 1 = -\frac{f(x_k)}{\phi_k^2} (f[y_k, z_k, x_k] + f[z_k, x_k, \xi_k] (\tau_k + \gamma \phi_k)). \quad (3.16)$$

By substituting (3.13), (3.15), and (3.16) into (3.14), we obtain another representation of α_k :

$$\alpha_k = \frac{1}{1 - \frac{f(x_k)}{\phi_k^2(1 + \gamma\phi_k)} F_k}; \tag{3.17}$$

here $F_k = (\tau_k(\tau_k + \gamma\phi_k)f[y_k, z_k, x_k] + ((1 - \tau_k)(2\tau_k + \gamma\phi_k) + (\tau_k + \gamma\phi_k)^2)f[z_k, x_k, \xi_k])$.

Now, we are going to find an asymptotic formula for α_k defined by (3.14). To this end, we use the formulas

$$\frac{f[z_k, x_k]}{\phi_k} - 1 = -\frac{(\bar{\tau}_k + v_k)\theta_k}{1 + \bar{\tau}_k\theta_k}, \tag{3.18}$$

$$\frac{f[z_k, y_k]}{\phi_k} - 1 = \frac{1 - \bar{\tau}_k - v_k}{\bar{\tau}_k}, \quad v_k = f(z_k)/f(y_k), \tag{3.19}$$

$$\tau_k = 1 + \bar{\tau}_k\theta_k. \tag{3.20}$$

Similarly, (3.13) can be rewritten in terms $\bar{\tau}_k$ as (3.18) and (3.19). Taking this into account and using (3.18) and (3.19), we can rewrite (3.14) as

$$\alpha_k = \frac{1}{1 + \frac{A_1 + A_2 + A_3v_k}{(1 + \gamma\phi_k)(1 + \gamma\phi_k + \bar{\tau}_k\theta_k)(1 + \bar{\tau}_k\theta_k)\bar{\tau}_k}}, \tag{3.21}$$

where

$$A_1 = (1 + \theta_k\bar{\tau}_k)^2(1 + \gamma\phi_k + \bar{\tau}_k\theta_k)^2(1 - \bar{\tau}_k), \tag{3.22a}$$

$$A_2 = (2 + \gamma\phi_k + 2\bar{\tau}_k\theta_k + (1 + \gamma\phi_k + \bar{\tau}_k\theta_k)^2)\bar{\tau}_k^3\theta_k^2, \tag{3.22b}$$

$$A_3 = -(1 + \theta_k\bar{\tau}_k)^2(1 + \gamma\phi_k + \bar{\tau}_k\theta_k)^2 + (2 + \gamma\phi_k + 2\bar{\tau}_k\theta_k + (1 + \gamma\phi_k + \bar{\tau}_k\theta_k)^2)\bar{\tau}_k^2\theta_k^2. \tag{3.22c}$$

Due to (2.23), we may write $\bar{\tau}_k$ as

$$\bar{\tau}_k = 1 + \hat{d}_k\theta_k + \tilde{\beta}_k\theta_k^2 + \tilde{\gamma}_k\theta_k^3 + \dots, \tag{3.23}$$

where $\tilde{\beta}_k$ and $\tilde{\gamma}_k$ are constants. Then, by Theorem 1 we have

$$f(y_k) = O(f^2(x_k)), \quad f(z_k) = O(f^4(x_k)), \quad v_k = O(f^2(x_k)). \tag{3.24}$$

Using (3.23) and (3.24) in (3.22), we obtain

$$A_1 = -\theta_k(1 + \gamma\phi_k)^2(a_1 + a_2\theta_k + a_3\theta_k^2 + \dots), \tag{3.25}$$

$$A_2 = \theta_k^2(1 + \gamma\phi_k)^2(b_1 + b_2\theta_k + \dots), \tag{3.26}$$

$$A_3 = -(1 + \gamma\phi_k)^2(c_1 + c_2\theta_k + \dots), \tag{3.27}$$

where

$$a_1 = \hat{d}_k, \quad a_2 = \tilde{\beta} + 2\hat{d}_k^2, \quad a_3 = \tilde{\gamma} + 2\tilde{\beta}\hat{d}_k + \left(3\hat{d}_k^2 + \frac{2}{1 + \gamma\phi_k}\right)\hat{d}_k, \tag{3.28}$$

$$b_1 = 1 + \frac{\hat{d}_k}{1 + \gamma\phi_k}, \quad b_2 = \hat{d}_k - \hat{d}_k^2 + 3\hat{d}_k^3, \quad c_1 = 1, \quad c_2 = 2\hat{d}_k.$$

Similarly, we have

$$\frac{1}{\left(1 + \frac{\bar{\tau}_k\theta_k}{1 + \gamma\phi_k}\right)(1 + \bar{\tau}_k\theta_k)\bar{\tau}_k} = 1 - 2\hat{d}_k\theta_k + \left(2\hat{d}_k^2 - \tilde{\beta} - \frac{1}{1 + \gamma\phi_k}\right)\theta_k^2 + O(f^3(x_k)). \tag{3.29}$$

Then

$$\begin{aligned} \frac{A_1 + A_2 + A_3 v_k}{(1 + \gamma \phi_k)(1 + \gamma \phi_k + \bar{\tau}_k \theta_k)(1 + \bar{\tau}_k \theta_k) \bar{\tau}_k} &= \left(1 - 2\hat{d}_k \theta_k + \left(2\hat{d}_k^2 - \tilde{\beta} - \frac{1}{1 + \gamma \phi_k} \right) \theta_k^2 \right) \\ &\times (-a_1 \theta_k + (b_1 - a_2) \theta_k^2 + (b_2 - a_3) \theta_k^3 - (c_1 + c_2 \theta_k) v_k) \\ &= -a_1 \theta_k + (b_1 - a_2 + 2a_1 \hat{d}_k) \theta_k^2 + (b_2 - a_3 - 2\hat{d}_k (b_1 - a_2) \\ &- a_1 \left(2\hat{d}_k^2 - \tilde{\beta} - \frac{1}{1 + \gamma \phi_k} \right)) \theta_k^3 - (c_1 + (c_2 - 2c_1 \hat{d}_k) \theta_k) v_k + O(f^4(x_k)). \end{aligned}$$

Substitute this expression into (3.21) and use the known expansion (2.20) to obtain

$$\begin{aligned} \alpha_k &= 1 + a_1 \theta_k - (b_1 - a_2 + 2a_1 \hat{d}_k) \theta_k^2 + (2\hat{d}_k (b_1 - a_2) + a_1 \left(2\hat{d}_k^2 - \tilde{\beta} - \frac{1}{1 + \gamma \phi_k} \right) - (b_2 - a_3)) \theta_k^3 \\ &+ (c_1 + (c_2 - 2c_1 \hat{d}_k) \theta_k) v_k + a_1^2 \theta_k^2 - 2a_1 (b_1 - a_2 + 2a_1 \hat{d}_k) \theta_k^3 + 2a_1 c_1 \theta_k v_k + a_1^3 \theta_k^3 + O(f^4(x_k)) \\ &= 1 + a_1 \theta_k + (a_1^2 - (b_1 - a_2) - 2a_1 \hat{d}_k) \theta_k^2 + \left((a_1^3 + 2\hat{d}_k (b_1 - a_2) + a_1 \left(2\hat{d}_k^2 - \tilde{\beta} - \frac{1}{1 + \gamma \phi_k} \right)) \right. \\ &\quad \left. - (b_2 - a_3) - 2a_1 (b_1 - a_2 + 2a_1 \hat{d}_k) \right) \theta_k^3 + (c_1 + (c_2 - 2c_1 \hat{d}_k) \theta_k) v_k + O(f^4(x_k)) \end{aligned}$$

or

$$\alpha_k = 1 + \hat{d}_k \theta_k + \left(\tilde{\beta} + \frac{1}{1 + \gamma \phi_k} \right) \theta_k^2 + \left(\tilde{\gamma} + \hat{d}_k \tilde{\beta} - \hat{d}_k - \frac{\hat{d}_k}{(1 + \gamma \phi_k)^2} \right) \theta_k^3 + (1 + 2\hat{d}_k \theta_k) v_k + O(f^4(x_k)). \quad (3.30)$$

As $\gamma \rightarrow 0$, formula (3.30) reduces to the form

$$\alpha_k = 1 + 2\theta_k + (\tilde{\beta} + 1) \theta_k^2 + (\tilde{\gamma} + 2\tilde{\beta} - 4) \theta_k^3 + (1 + 4\theta_k) v_k + O(f^4(x_k)), \quad (3.31)$$

which describes the asymptotic behavior of α_k in three-point iterative methods (see [6]).

Theorem 3. *Let all assumptions of Theorem 1 be fulfilled. Then, the three-point iterative methods (3.1) have the eighth order of convergence if and only if the iteration parameters $\bar{\tau}_k$ and α_k are specified by formulas (3.23) and (3.30), respectively.*

Proof. Let $\bar{\tau}_k$ and α_k be defined by formulas (3.23) and (3.30), respectively. Then, by Theorem 1 the first two steps in (3.1) determine an optimal fourth-order method, i.e., $f(z_k) = O(f^4(x_k))$. The value α_k specified by formula (3.30) satisfies condition (3.4). Therefore, we have (3.3). Conversely, assume that the order of convergence of (3.1) is eight. Then (3.1) and (3.3) imply that $f(z_k) = O(f^4(x_k))$ and formula (3.4) is valid. Therefore, by Theorem 1 formula (3.23) is valid for certain constants $\tilde{\gamma}$ and $\tilde{\beta}$. Using approximation (3.9) in (3.4), we obtain (3.14) accurate to $O(f^4(x_k))$. Due to (3.23), we obtain from (3.14) the asymptotic formula (3.30).

Assume that in (3.1)

$$\bar{\tau}_k = H(\theta_k) = \frac{c + (\hat{d}_k c + d) \theta_k + \omega \theta_k^2}{c + d \theta_k + b \theta_k^2}, \quad c + d + b \neq 0, \quad c, d, b, \omega \in R, \quad (3.32)$$

$$\alpha_k = H(\theta_k) + \frac{1}{1 + \gamma \phi_k} \theta_k^2 + \hat{d}_k \left(\tilde{\beta} - \frac{2}{1 + \gamma \phi_k} \right) \theta_k^3 + (1 + 2\hat{d}_k \theta_k) v_k. \quad (3.33)$$

Then, we obtain a family of optimal derivative-free three-point iterative methods because $\bar{\tau}_k$ and α_k determined by (2.23) and (3.33) satisfy conditions (3.23) and (3.30) with the constants

$$\tilde{\beta} = \frac{\omega - b}{c} - \frac{d}{c} \left(\frac{d}{c} + \hat{d}_k \right), \quad \tilde{\gamma} = -\frac{(b + \omega)d}{c^2} + \frac{d^2 - bc}{c^2} \hat{d}_k,$$

respectively. Therefore, the generation function method described in [5] makes it possible to construct the family of optimal three-point iterations.

Table 2. Nonlinear functions

	Functions	Root
1.	$f_1(x) = e^{(x^2+x \cos x-1)} \sin x + x \log(x \sin x + 1)$, [9]	$x^* = 0$
2.	$f_2(x) = \log(x^2 - 2x + 2) + e^{(x^2-5x+4)} \sin(x - 1)$,	$x^* = 1$
3.	$f_3(x) = \begin{cases} x(x + 1) & \text{if } x < 0, \\ -2x(x - 1) & \text{if } x \geq 0, \end{cases}$ [7, 32]	$x^* = 0$ $x^* = 1$ $x^* = -1$
4.	$f_4(x) = x^2 - 4 $, [9]	$x^* = \pm 2$

Now consider the three-point iterative method

$$\eta_k = x_k + \mathcal{V}(x_k), \quad y_k = x_k - \frac{f(x_k)}{f[x_k, \eta_k]}, \quad z_k = \Psi_4(x_k, y_k, \eta_k), \tag{3.34}$$

$$x_{k+1} = z_k - \frac{f(z_k)}{f[z_k, y_k] + (z_k - y_k)f[z_k, y_k, x_k] + (z_k - y_k)(z_k - x_k)f[z_k, y_k, x_k, \eta_k]}.$$

Here the function Ψ_4 is taken from any optimal derivative-free fourth-order method and $f[z_k, y_k, x_k, \eta_k]$ is the third divided difference. Theorem 2 implies the following result.

Theorem 4. *Let all assumptions of Theorem 1 be fulfilled. Then, the order of convergence of the iterative method (3.34) is eight.*

Proof. Since Ψ_4 is a fourth-order iteration, z_k can be rewritten as

$$z_k = y_k - \bar{\tau}_k \frac{f(y_k)}{\phi(x_k)}, \quad \phi(x_k) = f[x_k, \eta_k].$$

By Theorem 1, we have $\bar{\tau}_k = 1 + \hat{d}_k \theta_k + O(f^2(x_k))$. This implies the Taylor expansion (3.23) for $\bar{\tau}_k$. By comparing (3.1) with (3.34), we obtain

$$\alpha_k = \frac{\phi_k}{f[z_k, y_k] + (z_k - y_k)f[z_k, y_k, x_k] + (z_k - y_k)(z_k - x_k)f[z_k, y_k, x_k, \eta_k]} \tag{3.35}$$

It is easy to verify that the parameter α_k defined by formula (3.35) satisfies condition (3.30). Then, Theorem 3 implies that the order of convergence of (3.34) is eight 8.

Remark 1. The order of convergence of the three-point iterative methods proposed in [12, 22–24] immediately follows from Theorem 4, which is an extension of the theorems in [12, 22–24].

Note that all existing optimal derivative-free methods can be unambiguously written in form (3.1).

It is easy to verify that the parameters $\bar{\tau}_k$ and α_k in these methods have the same asymptotics (3.23) and (3.30) with specific constants $\tilde{\gamma}$ and $\tilde{\beta}$. Thus, the convergence of all existing optimal derivative-free methods can be proved using the sufficient convergence conditions (3.23) and (3.30) without symbolic computations. Furthermore, the application of these sufficient convergence conditions makes it possible to construct new optimal iterative methods [29]. It is seen from Table 1 that the parameter $\bar{\tau}_k$ in all optimal three-point methods listed in it is obtained using the generating functions $H(\theta_k)$ determined by (3.32); the only exception is the method proposed in [16]. It is seen from (3.32) and (3.34) that the function Ψ_4 can contain free parameters. This implies that the iterative methods (3.34) form a wide class of optimal derivative-free three-point methods. This class includes many well-known methods as special cases (see [4–6, 9, 12]). As in the preceding section, we can vary γ at each iteration step using the information

Table 3. Two-point iterative methods

Methods	$\bar{\tau}_k$	k	$ x^* - x_k $	COC
Numerical results for the smooth function $f_1(x)$ with $x_0 = 1$				
(2.1)	$c = 1, d = -\hat{d}_k, b = -\frac{1}{1 + \gamma\varphi_k}, \omega = 0$	4	0.4180e-33	3.99
King-type [23]	$c = 1, d = -\hat{d}_k, b = \frac{1}{1 + \gamma\varphi_k}, \omega = 0$	5	0.5272e-96	4.00
Potra–Ptak’s [13, 22]	$c = 1, d = b = 0, \omega = \frac{\hat{d}_k}{2}$	5	0.9744e-80	3.99
P1 [21]	$c = 1, b = d = \omega = 0$	5	0.1887e-65	4.00
P2 [21]	$c = 1, d = -\frac{1}{1 + \gamma\varphi_k}, b = \omega = 0$	5	0.1022e-95	4.00
Zheng’s [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	4	0.1655e-35	4.00
(2.31)	$c = b = 1, d = -2, \omega = 0$	5	0.1416e-95	4.00
(2.32)	$c = 1, d = \omega = -1, b = 0$	5	0.3838e-82	3.99
Steffensen’s	$x_{k+1} = x_k - \frac{f(x_k)}{\phi(x_k)}$	9	0.8745e-58	2.00
Numerical results for the smooth function $f_2(x)$ with $x_0 = 0.5$				
(2.1)	$c = 1, d = -\hat{d}_k, b = -\frac{1}{1 + \gamma\varphi_k}, \omega = 0$	4	0.1673e-104	4.00
King-type [23]	$c = 1, d = -\hat{d}_k, b = \frac{1}{1 + \gamma\varphi_k}, \omega = 0$	5	0.8607e-112	4.00
Potra–Ptak’s [13, 22]	$c = 1, d = b = 0, \omega = \frac{\hat{d}_k}{2}$	5	0.4066e-70	4.00
P1 [21]	$c = 1, b = d = \omega = 0$	5	0.1325e-62	4.00
P2 [21]	$c = 1, d = -\frac{1}{1 + \gamma\varphi_k}, b = \omega = 0$	5	0.5680e-88	4.00
Zheng’s [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	4	0.4934e-58	3.99
(2.31)	$c = b = 1, d = -2, \omega = 0$	5	0.6144e-109	4.00
(2.32)	$c = 1, d = \omega = -1, b = 0$	5	0.6129e-73	4.00
Steffensen’s	$x_{k+1} = x_k - \frac{f(x_k)}{\phi(x_k)}$	8	0.4282e-30	2.00

obtained at the preceding and the current steps. This enables us to increase the order of convergence without using additional computations. More precisely, we can obtain three-point iterative methods with memory (x_0 and γ_0 are given). Then, $\eta_0 = x_0 + \gamma_0 f(x_0)$ and

$$\begin{aligned}
 \gamma_k &= -\frac{1}{N_4^1(x_k)}, \quad \eta_k = x_k + \gamma_k f(x_k), \quad k = 1, 2, \dots \\
 y_k &= x_k - \frac{f(x_k)}{\phi(x_k, \gamma_k)}, \quad z_k = \Psi_4(x_k, y_k, \eta_k), \\
 x_{k+1} &= z_k - \frac{f(z_k)}{f[z_k, y_k] + (z_k - y_k)f[z_k, y_k, x_k] + (z_k - y_k)(z_k - x_k)f[z_k, y_k, x_k, \eta_k]}.
 \end{aligned}
 \tag{3.36}$$

Table 4. Three-point iterative methods

Methods	$\bar{\tau}_k = H(\theta_k)$	k	$ x^* - x_k $	COC
	choice of parameters			
Numerical results for the smooth function $f_1(x)$ with $x_0 = 1$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	3	0.1710e-38	8.38
(3.34)	$c = b = 1, d = -2, \omega = 0$	3	0.3900e-57	7.94
(3.34)	$c = 1, d = \omega = -1, b = 0$	3	0.4900e-44	7.99
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	3	0.4362e-43	7.99
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\phi_k}, (\beta = 2)$	3	0.1024e-54	7.98
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	3	0.5610e-62	7.97
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\phi_k}, b = \omega = 0$	3	0.9068e-48	8.00
Numerical results for the smooth function $f_2(x)$ with $x_0 = 0.5$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	3	0.3321e-33	7.96
(3.34)	$c = b = 1, d = -2, \omega = 0$	3	0.1543e-44	8.07
(3.34)	$c = 1, d = \omega = -1, b = 0$	3	0.4989e-36	7.98
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	3	0.2769e-35	7.99
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\phi_k}, (\beta = 2)$	3	0.5302e-44	8.00
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	3	0.6281e-64	7.97
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\phi_k}, b = \omega = 0$	3	0.7441e-40	8.02

Here $N_4(t) = N_4(t, x_k, z_{k-1}, y_{k-1}, \eta_{k-1}, x_{k-1})$ the fourth degree interpolation Newton polynomial specified by the node points $x_k, z_{k-1}, y_{k-1}, \eta_{k-1}, x_{k-1}$. As in [9], it is easy to prove that the order R of convergence of method (3.36) is at least 12.

4. NUMERICAL RESULTS

In this section, we describe the results of numerical computations for comparing the effectiveness of different methods. The computations were performed in Maple. To ensure high accuracy and avoid losing significant digits, the computations were performed with 300 significant digits. The computations were performed for smooth and nonsmooth functions (see Table 2) with $\gamma = -0.01$. To check the convergence of Newtons, the computational order of convergence (COC) was calculated by the formula

$$p \approx \frac{\ln(|x_k - x^*|/|x_{k-1} - x^*|)}{\ln(|x_{k-1} - x^*|/|x_{k-2} - x^*|)},$$

Table 5. Numerical results for the nonsmooth function $f_3(x)$. Three-point iterative methods

Methods	$\bar{\tau}_k = H(\theta_k)$	k	$ x^* - x_k $	COC
$x_0 = 0.1, x^* = 0$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	4	0.7235e-30	2.00
(3.34)	$c = b = 1, d = -2, \omega = 0$	4	0.7186e-30	2.00
(3.34)	$c = 1, d = \omega = -1, b = 0$	4	0.7222e-30	2.00
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	4	0.7221e-30	2.00
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\varphi_k}, (\beta = 2)$	4	0.7185e-30	2.00
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	4	0.7167e-30	2.00
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\varphi_k}, b = \omega = 0$	4	0.7205e-30	2.00
$x_0 = 5, x^* = 1$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	4	0.2191e-236	7.99
(3.34)	$c = b = 1, d = -2, \omega = 0$	3	0.8113e-39	7.77
(3.34)	$c = 1, d = \omega = -1, b = 0$	3	0.8754e-32	7.60
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	3	0.2144e-31	7.60
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\varphi_k}, (\beta = 2)$	3	0.2249e-37	7.76
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	3	0.5377e-47	7.86
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\varphi_k}, b = \omega = 0$	3	0.4975e-34	7.67
$x_0 = -10, x^* = -1$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	4	0.4791e-102	7.99
(3.34)	$c = b = 1, d = -2, \omega = 0$	4	0.2067e-141	7.99
(3.34)	$c = 1, d = \omega = -1, b = 0$	4	0.9351e-112	7.99
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	4	0.5302e-110	7.99
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\varphi_k}, (\beta = 2)$	4	0.2101e-135	7.99
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	4	0.8976e-178	7.99
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\varphi_k}, b = \omega = 0$	4	0.2099e-121	7.99

where x_k, x_{k-1} , and x_{k-2} are three successive approximations. The iterative process is stopped when $|x_k - x^*| < 10^{-30}$.

Table 2 presents the example taken from [9]. The third example with the nonsmooth function (see [7, 14, 32]) is often used for checking the validity of derivative-free iterative methods. Tables 3–6 show the number of iteration steps (k), the absolute errors $|x_k - x^*|$, and COC for the methods with $\gamma = -0.01$. The numerical results confirm the theoretical conclusion about the order of convergence. It is seen from Table 6 that the high order methods work well not only for sufficiently smooth functions but also for nonsmooth ones. Note that the derivative of the nonlinear function $f_3(x)$ has a discontinuity at the point $x^* = 0$; for

Table 6. Three-point iterative methods

Methods	$\bar{\tau}_k = H(\theta_k)$	k	$ x^* - x_k $	COC
	choice of parameters			
Numerical results for the nonsmooth function $f_4(x)$ with $x_0 = 3$				
(3.34)	$c = 1, d = \beta - 2, b = \omega = 0, (\beta = 2)$	2	0.1365e-35	7.70
(3.34)	$c = b = 1, d = -2, \omega = 0$	2	0.3071e-40	7.79
(3.34)	$c = 1, d = \omega = -1, b = 0$	2	0.8144e-37	7.72
Lotfi's [22]	$c = 1, d = b = 0, \omega = \frac{\tilde{d}_k}{2}$	2	0.1228e-36	7.72
King-type [23]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_k, b = \frac{2 - \beta}{1 + \gamma\phi_k}, (\beta = 2)$	2	0.3717e-40	7.80
Zheng's [12]	$c = 1, d = -\hat{d}_k, b = \omega = 0$	2	0.1675e-44	7.84
Sharma's [14]	$c = 1, d = -\frac{1}{1 + \gamma\phi_k}, b = \omega = 0$	2	0.2114e-38	7.75
Steffensen's	$x_{k+1} = x_k - \frac{f(x_k)}{\phi(x_k)}$	6	0.5556e-45	2.00

this reason, the COC = 2 in this case for all the methods discussed in this paper (see the first part of Table 5 and [7]). The proposed methods (3.34) can be successfully used in the computations that require high accuracy.

5. CONCLUSIONS

The necessary and sufficient convergence conditions for two- and three-point iterative methods obtained in [6] are extended for the case of derivative-free methods. The latter methods can be effectively used not only for determining the order of convergence but also for constructing new methods. Based on the generating function method, wide classes of optimal methods that include many known methods as special cases are proposed.

FUNDING

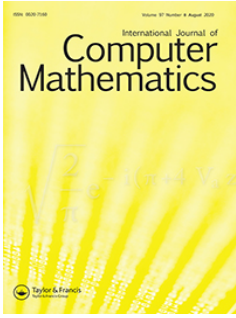
This work was supported by the Foundation of Science and Technology of Mongolia, project no. SST_18/2018 and by the program JINR–Romania–Hulubei–Meshcheryakov of the Joint Institute for Nuclear Research.

REFERENCES

1. L. Liu and X. Wang, "Eighth-order methods of high efficiency index for solving nonlinear equations," *Appl. Math. Comput.* **215**, 3449–3454 (2010).
2. M. S. Petković, B. Neta, L. D. Petković, and J. Džunić, "Multipoint methods for solving nonlinear equations: A survey," *Appl. Math. Comput.* **226**, 635–660 (2014).
3. R. Thukral and M. S. Petković, "A family of three-point methods of optimal order for solving nonlinear equations," *J. Comput. Appl. Math.* **233**, 2278–2284 (2010).
4. X. Wang and L. Liu, "New eighth-order iterative methods for solving non-linear equations," *J. Comput. Appl. Math.* **234**, 1611–1620 (2010).
5. T. Zhanlav, V. Ulziibayar, O. Chuluunbaatar, and "Generating function method for constructing new iterations," *Appl. Math. Comput.* **315**, 414–423 (2017).
6. T. Zhanlav, V. Ulziibayar, and O. Chuluunbaatar, "Necessary and sufficient conditions for the convergence of two- and three-point Newton-type iterations," *Comput. Math. Math. Phys.* **57**, 1090–1100 (2017).
7. A. Cordero, J. L. Hueso, E. Martinez, and J. R. Torregrosa, "A new technique to obtain derivative-free optimal iterative methods for solving nonlinear equations," *J. Comput. Appl. Math.* **252**, 95–102 (2013).
8. M. D. Junjua, F. Zafar, N. Yasmin, and S. Akram, "A general class of derivative free with memory root solvers," *Appl. Math. Phys.* **79**, 19–28 (2017).

9. I. K. Argyros, M. Kansal, V. Kanwar, and S. Bajaj, “Higher-order derivative-free families of Chebyshev–Halley type methods with or without memory for solving nonlinear equations,” *Appl. Math. Comput.* **315**, 224–245 (2017).
10. S. K. Khattri and T. Steihaug, “Algorithm for forming derivative-free optimal methods,” *Numer. Algorithms* **5**, 809–842 (2014).
11. R. Thukral, “Eighth-order iterative methods without derivatives for solving nonlinear equations,” *ISRN Appl. Math.* Article ID 693787 (2011).
12. Q. Zheng, J. Li, and F. Huang, “An optimal Steffensen-type family for solving nonlinear equations,” *Appl. Math. Comput.* **217**, 9592–9597 (2011).
13. F. Soleymani, R. Sharma, X. Li, and E. Tohidi, “An optimized derivative-free form of the Potra-Ptak method,” *Math. Comput. Model.* **56**, 97–104 (2012).
14. J. R. Sharma and R. K. Goyal, “Fourth-order derivative-free methods for solving non-linear equations,” *Int. J. Comput. Math.* **83**, 101–106 (2006).
15. A. Cordero and J. R. Torregrosa, “A class of Steffensen type methods with optimal order of convergence,” *Appl. Math. Comput.* **217**, 7653–7659 (2011).
16. H. Ren, Q. Wu, and W. Bi, “A class of two-step Steffensen type methods with fourth-order convergence,” *Appl. Math. Comput.* **209**, 206–210 (2009).
17. Z. Liu, Q. Zheng, and P. Zhao, “A variant of Steffensen’s method of fourth-order convergence and its applications,” *Appl. Math. Comput.* **216**, 1978–1983 (2010).
18. Y. Peng, H. Feng, Q. Li, and X. Zhang, “A fourth-order derivative-free algorithm for nonlinear equations,” *J. Comput. Appl. Math.* **235**, 2551–2559 (2011).
19. M. Kansal, V. Kanwar, and S. Bhatia, “An optimal eighth-order derivative-free family of Potra-Ptak’s method,” *Algorithms* **8**, 309–320 (2015).
20. F. Soleymani and S. K. Khattri, “Finding simple roots by seventh-and eighth-order derivative-free methods,” *Int. J. Math. Models Meth. Appl. Sci.* **6**, 45–52 (2012).
21. R. Thukral, “A family of three-point derivative-free methods of eighth-order for solving nonlinear equations,” *J. Mod. Meth. Numer. Math.* **3**, 11–21 (2012).
22. T. Lotfi, F. Soleymani, M. Ghorbanzadeh, and P. Assari, “On the construction of some tri-parametric iterative methods with memory,” *Numer. Algorithms* **70**, 835–845 (2015).
23. S. Sharifi, S. Siegmund, and M. Salimi, “Solving nonlinear equations by a derivative-free form of the King’s family with memory,” *Calcolo* **53**, 201–215 (2016).
24. J. R. Sharma, R. K. Guha, and P. Gupta, “Some efficient derivative free methods with memory for solving nonlinear equations,” *Appl. Math. Comput.* **219**, 699–707 (2012).
25. R. Behl, D. Gonzalez, P. Maroju, and S. S. Motsa, “An optimal and efficient general eighth-order derivative-free scheme for simple roots,” *J. Comput. Appl. Math.* **330**, 666–675 (2018).
26. F. Soleymani, “Optimal fourth-order iterative methods free from derivative,” *Miskolc Math. Notes.* **12**, 255–264 (2011).
27. S. K. Khattri and R. P. Agarwal, “Derivative-free optimal iterative methods,” *Comput. Meth. Appl. Math.* **10**, 368–375 (2010).
28. T. Zhanlav and Kh. Otgondorj, “A new family of optimal eighth-order methods for solving nonlinear equations,” *Am. J. Comput. Appl. Math.* **8**, 15–19 (2018).
29. F. Zafar, N. Yasmin, M. A. Kutbi, and M. Zeshan, “Construction of Tri-parametric derivative free fourth order with and without memory iterative method,” *J. Nonlinear Sci. Appl.* **9**, 1410–1423 (2016).
30. F. Soleymani and S. K. Vanani, “Optimal Steffensen-type methods with eighth order of convergence,” *Comput. Math. Appl.* **62**, 4619–4626 (2011).
31. F. Soleymani, “On a bi-parametric class of optimal eighth-order derivative-free methods,” *Int. J. Pure. Appl. Math.* **72**, 27–37 (2011).
32. M. Kansal, V. Kanwar, and S. Bhatia, “Efficient derivative-free variants of Hansen-Patrick’s family with memory for solving nonlinear equations,” *Numer. Algor.* **73**, 1017–1036 (2016).
33. F. I. Chicharro, A. Cordero, J. R. Torregrosa, and M. P. Vassileva, “King-Type derivative-free iterative families: Real and memory dynamics,” *Complexity* **2017**, Article ID 2713145 (2017). <https://doi.org/10.1155/2017/2713145>
34. M. S. Petković, S. Ilić, and J. Džunić, “Derivative-free two-point methods with and without memory for solving nonlinear equations,” *Appl. Math. Comput.* **217**, 1887–1895 (2010).

Translated by A. Klimontovich



High-order iterations for systems of nonlinear equations

T. Zhanlav , Changbum Chun , Kh. Otgondorj & V. Ulziibayar

To cite this article: T. Zhanlav , Changbum Chun , Kh. Otgondorj & V. Ulziibayar (2020) High-order iterations for systems of nonlinear equations, International Journal of Computer Mathematics, 97:8, 1704-1724, DOI: [10.1080/00207160.2019.1652739](https://doi.org/10.1080/00207160.2019.1652739)

To link to this article: <https://doi.org/10.1080/00207160.2019.1652739>



Accepted author version posted online: 15 Aug 2019.
Published online: 25 Aug 2019.



Submit your article to this journal [↗](#)



Article views: 78



View related articles [↗](#)



View Crossmark data [↗](#)

High-order iterations for systems of nonlinear equations

T. Zhanlav^a, Changbum Chun^b, Kh. Otgondorj^c and V. Ulziibayar^{a,c}

^aInstitute of Mathematics, National University of Mongolia, Mongolia; ^bDepartment of Mathematics, Sungkyunkwan University, Suwon, Republic of Korea; ^cSchool of Applied Sciences, Mongolian University of Science and Technology, Mongolia

ABSTRACT

In this paper, several families of order p ($4 \leq p \leq 6$) for the solution of systems of nonlinear equations are developed and compared to existing methods. The necessary and sufficient conditions for p th order of convergence are given in terms of parameter matrices τ_k and α_k . Several choices of parameter matrix Θ_k determining τ_k are suggested. The proposed families include some well-known methods as particular cases. The comparison is made based on the total cost of an iteration and the CPU time.

ARTICLE HISTORY

Received 22 November 2018
Revised 24 June 2019
Accepted 28 July 2019

KEYWORDS

Systems of nonlinear equations; Newton-type methods; order of convergence

2010 MATHEMATICS SUBJECT CLASSIFICATIONS

65H10; 65Y20; 41A58

1. Introduction

For a given nonlinear system $F(x) : D \subset R^n \rightarrow R^n$, the problem is to find a vector $(x_{(1)}^*, x_{(2)}^*, \dots, x_{(n)}^*)^T$ such that $F(x) = 0$. Such problem is important and interesting and often appears in numerical analysis and engineering. The most widely used method for solving this problem is the quadratically convergent Newton's method given by

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \dots, \quad (1)$$

where x_0 is the initial approximation and $F'(x)^{-1}$ is the inverse of Fréchet derivative $F'(x)$ of the function $F(x)$. In recent years, a number of methods with higher order of convergence for systems of nonlinear equations have been developed in the literature, for example, see [1,3–9,11,12,16–18] and references therein. The aim of this paper is to extend higher order methods presented in [23] to systems of nonlinear equations. This paper is organized as follows. In Section 2, we study the convergence of the proposed two-step iterative methods. Section 3 is devoted to the convergence of three-step methods with parameter $a = 1$ and $a \neq 1$. In Section 4, we consider the total cost comparison between methods. Finally, numerical results supporting theoretical ones and some comparison of methods are given in Section 5.

2. Two-step iterative methods

We consider the two-step iterative method

$$y_k = x_k - F'(x_k)^{-1}F(x_k), \quad (2a)$$

CONTACT Kh. Otgondorj  otgondorj@gmail.com

$$x_{k+1} = y_k - \bar{\tau}_k F'(x_k)^{-1} F(y_k), \quad (2b)$$

where $\bar{\tau}_k$ is some $n \times n$ matrix to be determined properly. Using the Taylor expansion of $F(x_{k+1})$ around y_k , we have

$$F(x_{k+1}) = \left(I - F'(y_k) \bar{\tau}_k F'(x_k)^{-1} \right) F(y_k) + O(\|F(y_k)\|^2), \quad (3)$$

where I is identity matrix. From (3), it clear that $F(x_{k+1}) = O(\|F(x_k)\|^4)$ if we choose $\bar{\tau}_k$, such that

$$I - F'(y_k) \bar{\tau}_k F'(x_k)^{-1} = 0 \quad \text{or} \quad \bar{\tau}_k = F'(y_k)^{-1} F'(x_k). \quad (4)$$

Substituting (4) into (2b), we obtain

$$x_{k+1} = y_k - F'(y_k)^{-1} F(y_k). \quad (5)$$

Our task is to approximate $F'(y_k)^{-1}$ with accuracy $O(\|F(x_k)\|^2)$. To this end, we use the expansion of $F'(y_k)$ at point x_k

$$\begin{aligned} F'(y_k) &= F'(x_k) - F''(x_k) F'(x_k)^{-1} F(x_k) + O(h^2) \\ &= F'(x_k) (I - P_k) + O(h^2), \end{aligned} \quad (6)$$

where

$$P_k = F'(x_k)^{-1} F''(x_k) F'(x_k)^{-1} F(x_k), \quad (7)$$

and $h = \|F(x_k)\|$, $O(\|F_k\|) = O(h)$. We can assume that

$$\|P_k\| \leq 1, \quad (8)$$

for x_k sufficiently close to x^* . By virtue of Banach lemma of inverse, from (6) we get

$$F'(y_k)^{-1} = (I - P_k)^{-1} F'(x_k)^{-1} + O(h^2) = (I + P_k) F'(x_k)^{-1} + O(h^2), \quad (9)$$

because

$$(I - P_k)^{-1} = \sum_{j=0}^{\infty} P_k^j = I + P_k + O(h^2). \quad (10)$$

Using the approximate formula obtained from (9)

$$F'(y_k)^{-1} \approx (I + P_k) F'(x_k)^{-1}, \quad (11)$$

in (5) gives

$$x_{k+1} = y_k - (I + P_k) F'(x_k)^{-1} F(y_k). \quad (12)$$

Using (11) in (4) we have

$$\bar{\tau}_k = I + P_k + O(h^2) = I + 2\Theta_k + O(h^2). \quad (13)$$

Here

$$\Theta_k = \frac{P_k}{2} = \frac{1}{2} F'(x_k)^{-1} F''(x_k) F'(x_k)^{-1} F(x_k). \quad (14)$$

Since

$$F(y_k) = \frac{F''(x_k)}{2} \left(F'(x_k)^{-1} F(x_k) \right)^2 + O(h^3), \quad (15)$$

$$F'(x_k)^{-1} F(y_k) = \frac{F'(x_k)^{-1} F''(x_k)}{2} \left(F'(x_k)^{-1} F(x_k) \right)^2 + O(h^3). \quad (16)$$

Substituting (14) into (16), we obtain

$$F'(x_k)^{-1} F(y_k) = \Theta_k F'(x_k)^{-1} F(x_k) + O(h^3). \quad (17)$$

On the other hand, using (2a) and (17), one can rewrite (12) as

$$x_{k+1} = x_k - \left(I + (I + P_k) \frac{P_k}{2} \right) F'(x_k)^{-1} F(x_k),$$

or

$$x_{k+1} = x_k - \tau_k F'(x_k)^{-1} F(x_k), \quad (18)$$

where

$$\tau_k = I + (I + P_k) \frac{P_k}{2} = I + \Theta_k + 2\Theta_k^2 + O(h^3). \quad (19)$$

From (13) and (19), one can find the connection of τ_k and $\bar{\tau}_k$ as:

$$\bar{\tau}_k = (\tau_k - I) \Theta_k^{-1} = I + 2\Theta_k + O(h^2). \quad (20)$$

Thus Theorem 2.1 in [23] is extended to system of nonlinear equations.

Theorem 2.1: Let $F(x) : D \subset R^n \rightarrow R^n$ be a sufficiently Fréchet differentiable function in a convex set $D \subset R^n$ containing a zero x^* of $F(x)$. Suppose that $F'(x)$ is continuous and nonsingular in x^* . Then, for an initial approximation sufficiently close to x^* , the sequence $\{x_k\}_{k \geq 0}$, $x_0 \in D$ obtained by (18) has order four if and only if the parameter matrix τ_k is given by (19).

Proof: The Taylor expansion of function $F(x_{k+1})$ at point y_k and use of (20) and (17) gives

$$F(x_{k+1}) = F(y_k) + F'(y_k)(\tau_k - I)(y_k - x_k) + O((\tau_k - I)(y_k - x_k))^2. \quad (21)$$

Since $\tau_k - I = O(h)$ and $y_k - x_k = O(h)$, using (6) and (17) into (21) we have

$$\begin{aligned} F(x_{k+1}) &= F(y_k) - F'(x_k)(I - P_k)(\tau_k - I)\Theta_k^{-1}F'(x_k)^{-1}F(y_k) + O(h^4) \\ &= F'(x_k) \left(I - (I - 2\Theta_k)(\tau_k - I)\Theta_k^{-1} \right) F'(x_k)^{-1}F(y_k) + O(h^4). \end{aligned} \quad (22)$$

By choice (19) we have $F(x_{k+1}) = O(h^4)$. The converse follows from (22). ■

From Theorem 2.1 immediately follows that the two-step iterative method (2) has a order four if and only if the parameter matrix $\bar{\tau}_k$ is given by (20). The main practical difficulty related to parameter matrix Θ_k is the evaluation of the second-order Fréchet derivative. For a nonlinear system of n equations and n unknowns, the first Fréchet derivative is a matrix with n^2 values, while the second Fréchet derivative for continuous functions has $\frac{1}{2}(n^3 + n^2)$ values. This implies a huge amount of operations

in order to evaluate every iteration. To overcome these difficulties we will find approximate formula for Θ_k . In general, we assume that y_k is defined by

$$y_k = x_k - aF'(x_k)^{-1}F(x_k), \quad a \neq 0. \quad (23)$$

In order to compute Θ_k with some accuracy we will use the first order divided difference of $F(x)$ [11]

$$[x+h, x; F] = \int_0^1 F'(x+th)dt = F'(x) + \frac{1}{2}F''(x)h + \frac{1}{6}F'''(x)h^2 + O(h^3), \quad (24)$$

where $h^i = (h, h, \dots, h)$, $h \in R^n$. Using definitions (24) and (23) we have

$$\begin{aligned} F'(x_k)^{-1}[y_k, x_k; F] &= I - a\Theta_k \\ &+ \frac{a^2}{6}F'(x_k)^{-1}F'''(x_k)\left(F'(x_k)^{-1}F(x_k)\right)^2 + O(h^3). \end{aligned} \quad (25)$$

Analogously, using Taylor expansion of $F'(y_k)$ at point x_k , we obtain

$$\begin{aligned} F'(x_k)^{-1}F'(y_k) &= I - 2a\Theta_k \\ &+ \frac{a^2}{2}F'(x_k)^{-1}F'''(x_k)\left(F'(x_k)^{-1}F(x_k)\right)^2 + O(h^3). \end{aligned} \quad (26)$$

The inverse of $F'(x_k)^{-1}F'(y_k)$ exists and by virtue of (26) we have

$$F'(y_k)^{-1}F'(x_k) = I + 2a\Theta_k + O(h^2). \quad (27)$$

Then from (26) and (27) it follows that

$$F'(x_k)^{-1}F'(y_k) + F'(y_k)^{-1}F'(x_k) = 2I + O(h^2). \quad (28)$$

From (25) we find

$$\begin{aligned} \Theta_k &= \frac{1}{a}\left(I - F'(x_k)^{-1}[y_k, x_k; F] \right. \\ &\left. + \frac{a^2}{6}F'(x_k)^{-1}F'''(x_k)\left(F'(x_k)^{-1}F(x_k)\right)^2\right) + O(h^3), \end{aligned} \quad (29)$$

It follows from (26) and (29) that

$$\Theta_k = \frac{1}{a}F'(x_k)^{-1}\left([y_k, x_k; F] - F'(y_k)\right) + O(h^3). \quad (30)$$

Elimination of term with factor a^2 from (25) and (26) gives

$$\Theta_k = \frac{1}{a}F'(x_k)^{-1}\left(2F'(x_k) + F'(y_k) - 3[y_k, x_k; F]\right) + O(h^3). \quad (31)$$

As a consequence of (31) and (30) we get

$$\Theta_k = \frac{1}{a}F'(x_k)^{-1}\left(F'(x_k) - [y_k, x_k; F]\right) + O(h^3), \quad (32)$$

and

$$\Theta_k = \frac{1}{2a}\left(I - F'(x_k)^{-1}F'(y_k)\right) + O(h^3). \quad (33)$$

Thus, we have four approximate formulas for Θ_k that will be used for determining the parameter matrix $\bar{\tau}_k$ in (20). In some cases it may be useful to have less accurate formula for Θ_k . Using (28)

in (33) we get

$$\Theta_k = \frac{1}{2a} \left(-I + F'(y_k)^{-1} F'(x_k) \right) + O(h^2). \quad (34)$$

In a similar way, (28) can be used in (30), (31) to obtain less accurate formulas. The divided difference $[y, x; F]$ of F is an $n \times n$ matrix with elements (see [11])

$$[y, x; F]_{ij} = \frac{F_i(y_{(1)}, \dots, y_{(j)}, x_{(j+1)}, \dots, x_{(n)}) - F_i(y_{(1)}, \dots, y_{(j-1)}, x_{(j)}, \dots, x_{(n)})}{y_{(j)} - x_{(j)}}, \quad (35)$$

where $1 \leq i, j \leq n$.

Remark 2.1: Note that the fourth-order convergence of iteration (2) holds true if Θ_k in (20) is replaced by one of the approximate formulae (30)–(34) with $a = 1$.

Analogously, we can prove that

Theorem 2.2: Assume that all the assumptions of Theorem 2.1 are fulfilled. Then the two-step iteration (18) has a third order of convergence if and only if the iteration matrix τ_k is chosen such that

$$\tau_k = I + \Theta_k + O(\Theta_k^2). \quad (36)$$

In what follows we will use two step iterative method (2a)–(2b) as

$$y_k = x_k - F'(x_k)^{-1} F(x_k), \quad (37a)$$

$$x_{k+1} = x_k - \tau_k F'(x_k)^{-1} F(x_k), \quad (37b)$$

which can be considered as a continuous analogy of Newton's method (*CANM for short*) or damped Newton's method [24]. Following the idea of generating functions method in [22] we can state the following.

Theorem 2.3: Suppose that all assumptions of Theorem 2.1 are fulfilled. Then the two-step iteration (37) has a fourth order of convergence if and only if the iteration matrix τ_k is chosen such that

$$\tau_k = (I - \alpha \Theta_k)^{-1} \left(I + (1 - \alpha) \Theta_k + (2 - \alpha) \Theta_k^2 + \omega \Theta_k^3 \right), \quad \alpha, \omega \in R, \quad (38)$$

Proof: It is easy to show that τ_k given by (38) can be rewritten as:

$$\begin{aligned} \tau_k &= (I + \alpha \Theta_k + \alpha^2 \Theta_k^2 + \alpha^3 \Theta_k^3 + \dots) (I + (1 - \alpha) \Theta_k + (2 - \alpha) \Theta_k^2 + \omega \Theta_k^3) \\ &= I + \Theta_k + 2\Theta_k^2 + O(h^3). \end{aligned}$$

Thus by Theorem 2.1 the convergence order of iteration (37) is equal to four. ■

From the approximate formulas (30)–(33), we see that the Theorem 2.3 is valid when Θ_k is defined by one of the formulas (30)–(33) with $a = 1$. When $\alpha = 0$ the formula (38) leads to (19). Now we

consider another two-point iterative method

$$y_k = x_k - aF'(x_k)^{-1}F(x_k), \quad x_{k+1} = x_k - \tau_k F'(x_k)^{-1}F(x_k), \quad (39)$$

It is easy to show that

$$F(y_k) = O(h^\sigma), \quad \sigma = \begin{cases} 2 & a = 1, \\ 1 & a \neq 1. \end{cases} \quad (40)$$

We consider the Taylor expansion of $F(x_{k+1})$ at point x_k . We have

$$\begin{aligned} F(x_{k+1}) &= F(x_k) - F'(x_k)\tau_k F'(x_k)^{-1}F(x_k) + \frac{F''(x_k)}{2} (\tau_k F'(x_k)^{-1}F(x_k))^2 \\ &\quad - \frac{F'''(x_k)}{6} (\tau_k F'(x_k)^{-1}F(x_k))^3 + O(h^4). \end{aligned} \quad (41)$$

As the preceding case, we seek for τ_k in the form

$$\tau_k = I + \Theta_k + c_k \Theta_k^2 + d_k + O(h^3). \quad (42)$$

Substituting (42) into (41) and taking into account the following relations

$$\begin{aligned} (\tau_k F'(x_k)^{-1}F(x_k))^2 &= (I + \Theta_k)(F'(x_k)^{-1}F(x_k))^2 \\ &\quad + F'(x_k)^{-1}F(x_k)\Theta_k F'(x_k)^{-1}F(x_k) + O(h^4), \\ (\tau_k F'(x_k)^{-1}F(x_k))^3 &= (F'(x_k)^{-1}F(x_k))^3 + O(h^4), \end{aligned} \quad (43)$$

we get

$$\begin{aligned} F(x_{k+1}) &= -F'(x_k)(\Theta_k + c_k \Theta_k^2 + d_k)F'(x_k)^{-1}F(x_k) \\ &\quad + \frac{F''(x_k)}{2} \left((I + \Theta_k)(F'(x_k)^{-1}F(x_k))^2 \right. \\ &\quad \left. + F'(x_k)^{-1}F(x_k)\Theta_k F'(x_k)^{-1}F(x_k) \right) \\ &\quad - \frac{F'''(x_k)}{6} (F'(x_k)^{-1}F(x_k))^3 + O(h^4). \end{aligned} \quad (44)$$

Multiplying by $F'(x_k)^{-1}$ both sides of (44) and taking (14) into account we have

$$\begin{aligned} F'(x_k)^{-1}F(x_{k+1}) &= \left((1 - c_k)\Theta_k^2 + \frac{F'(x_k)^{-1}}{2} F''(x_k)\Theta_k F'(x_k)^{-1}F(x_k) \right. \\ &\quad \left. - \left(d_k + \frac{1}{6} F'(x_k)^{-1} F'''(x_k)(F'(x_k)^{-1}F(x_k))^2 \right) \right) \\ &\quad \times F'(x_k)^{-1}F(x_k) + O(h^4). \end{aligned} \quad (45)$$

From (45) we see that

$$F(x_{k+1}) = O(h^4), \quad (46)$$

if we choose c_k and d_k such that

$$c_k = I + \frac{1}{2} F'(x_k)^{-1} F''(x_k)\Theta_k F'(x_k)^{-1}F(x_k)\Theta_k^{-2} + O(h) \quad (47)$$

and

$$d_k = -\frac{1}{6}F'(x_k)^{-1}F''(x_k)\left(F'(x_k)^{-1}F(x_k)\right)^2 + O(h^3). \quad (48)$$

From (26) we find d_k given by (48) as

$$d_k = -\frac{1}{3a^2}\left(F'(x_k)^{-1}F'(y_k) - (I - 2a\Theta_k)\right) + O(h^3). \quad (49)$$

Substituting (47) and (49) into (42), we get

$$\begin{aligned} \tau_k &= I + \Theta_k + \Theta_k^2 + \frac{1}{2}F'(x_k)^{-1}F''(x_k)\Theta_kF'(x_k)^{-1}F(x_k) \\ &\quad - \frac{F'(x_k)^{-1}F'(y_k) - (I - 2a\Theta_k)}{3a^2} + O(h^3). \end{aligned} \quad (50)$$

If we propose the following assumption

$$\frac{1}{2}F'(x_k)^{-1}F''(x_k)\Theta_kF'(x_k)^{-1}F(x_k) = \Theta_k^2 + O(h^3), \quad (51)$$

then (50) is written as

$$\tau_k = I + \Theta_k + 2\Theta_k^2 + d_k + O(h^3). \quad (52)$$

Note that the assumption (51) fulfilled for scalar equation case and τ_k given by (52) coincides with that of scalar equation case [22]. Furthermore, the assumption (51) holds true for some method (37). As an example, we consider the generalized Jarratt's method given in [8], as (39) with $a = \frac{2}{3}$ and

$$\tau_k = \frac{1}{2}(3F'(y_k) - F'(x_k))^{-1}(3F'(y_k) + F'(x_k)).$$

The method $M43$ [12] is a special case of $M4$, which is when $a = 1$ and (32) is used for Θ_k . Using (27) it is easy to show that τ_k satisfies (52). It means that the assumption (51) fulfilled in this case. We summarize the above result in the following theorem:

Theorem 2.4: *Let the assumption (51) be fulfilled. Then the two-step method (39) has a fourth-order convergence if τ_k is given by (52).*

Analogously, using (45) one can prove the following:

Theorem 2.5: *The two-step iteration (39) has a p th order ($p = 2, 3$) of convergence if τ_k is given by*

$$\tau_k = I + \Theta_k + O(h^2),$$

and

$$\tau_k = I + O(h),$$

respectively.

3. Three-step iterative methods

Now we consider three-step iterative method.

$$y_k = x_k - F'(x_k)^{-1}F(x_k), \tag{53a}$$

$$z_k = y_k - \bar{\tau}_k F'(x_k)^{-1}F(y_k), \tag{53b}$$

$$x_{k+1} = z_k - \alpha_k F'(x_k)^{-1}F(z_k). \tag{53c}$$

The order of convergence of iteration (53) is given by the following theorem:

Theorem 3.1: *Suppose that all the assumptions of Theorem 2.1 are fulfilled. Then the iteration (53) has a order p if and only if the iteration matrices $\bar{\tau}_k$ and α_k are given by formulas in Table 1.*

Proof: Let τ_k be defined by formula (19). Then according to Theorem 2.1 we have $F(z_k) = O(h^4)$. Then using expansion of $F(x_{k+1})$ around z_k we have

$$F(x_{k+1}) = \left(I - F'(z_k)\alpha_k F'(x_k)^{-1} \right) F(z_k) + O(F(z_k)^2). \tag{54}$$

From (54) it follows that

$$F(x_{k+1}) = O(F(z_k)^2), \tag{55}$$

under condition

$$\alpha_k = F'(z_k)^{-1}F'(x_k). \tag{56}$$

Using Taylor expansion of $F'(z_k)$ around x_k , we have

$$F'(z_k) = F'(x_k) \left(I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k) \right) + O(h^2). \tag{57}$$

From (57), it follows that

$$F'(z_k)^{-1} = \left(I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k) \right)^{-1} F'(x_k)^{-1} + O(h^2).$$

Using equality type (10) we have

$$F'(z_k)^{-1} = \left(I + F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k) \right) F'(x_k)^{-1} + O(h^2). \tag{58}$$

If we take into account (19) and $\Theta_k = O(F(x_k))$, then substituting (58) into (56) we find α_k with accuracy $O(h^2)$ as

$$\alpha_k = I + 2\Theta_k + O(h^2). \tag{59}$$

As a result, from (54) we get

$$F(x_{k+1}) = O(F(x_k)^6). \tag{60}$$

Table 1. Choices of parameters.

p	τ_k	α_k	$\bar{\tau}_k$
5	$I + \Theta_k + 2\Theta_k^2 + \beta\Theta_k^3 + O(\Theta_k^4)$	$I + O(\Theta_k)$	$I + 2\Theta_k + \beta\Theta_k^2$
	$I + \Theta_k + O(\Theta_k^2)$	$I + 2\Theta_k + O(\Theta_k^2)$	$I + O(\Theta_k)$
6	$I + \Theta_k + 2\Theta_k^2 + O(\Theta_k^3)$	$I + 2\Theta_k + O(\Theta_k^2)$	$I + 2\Theta_k + O(\Theta_k^2)$
7	$I + \Theta_k + 2\Theta_k^2 + O(\Theta_k^3)$	$I + 2\Theta_k + 6\Theta_k^2 + 3d_k$	$I + 2\Theta_k + O(\Theta_k^2)$

Conversely, let (60) hold. From (56) and (58), it can be easily seen that

$$\alpha_k = F'(z_k)^{-1}F'(x_k) = I + F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k) + O(h^2).$$

Substituting it into (54) we get

$$F(x_{k+1}) = O(h^2)F(z_k) + O(F(z_k)^2).$$

From this we conclude that

$$F(z_k) = O(h^4),$$

because of (60). Hence, by Theorem 2.1, we obtain (19). Substituting (57) into (54) we have

$$\begin{aligned} F(x_{k+1}) &= \left(I - F'(x_k)(I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k))\alpha_k F'(x_k)^{-1} \right) F(z_k) \\ &\quad + O(h^6). \end{aligned} \quad (61)$$

Then from (60) and (61) it follows that

$$I - F'(x_k)(I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k))\alpha_k F'(x_k)^{-1} = O(h^2),$$

or

$$(I - F'(x_k)^{-1}F''(x_k)F'(x_k)^{-1}F(x_k))\alpha_k = I + O(h^2),$$

in which we have used (19). From this we obtain

$$\alpha_k = (I - 2\Theta_k)^{-1}(I + O(h^2)) = I + 2\Theta_k + O(h^2).$$

The proof for $p = 6$ is complete.

Let the parameter matrix τ_k be defined by (36) and α_k is given by (59). Then by Theorem 2.2 we have $F(z_k) = O(F(x_k)^3)$. Using (36) in (57) we get

$$F'(z_k) = F'(x_k)(I - 2\Theta_k) + O(h^2).$$

Hence

$$I - F'(z_k)\alpha_k F'(x_k)^{-1} = I - F'(x_k)(I - 2\Theta_k)(I + 2\Theta_k)F'(x_k)^{-1} + O(h^2) = O(h^2), \quad (62)$$

because of $\Theta_k = O(F(x_k))$. Using (62) in (53c) we get

$$F(x_{k+1}) = O(F(x_k)^5), \quad (63)$$

i.e $p = 5$ when τ_k and α_k are defined by (36) and (59), respectively. Conversely, let (63) hold. From (54), it is clear that

$$F(z_k) = O(F(x_k)^3), \quad (64)$$

and

$$I - F'(z_k)\alpha_k F'(x_k)^{-1} = O(h^2). \quad (65)$$

Hence, by Theorem 2.2, we get $\tau_k = 1 + \Theta_k + O(h^2)$. Then from (57), we obtain

$$F'(z_k) = F'(x_k)(I - 2\Theta_k) + O(h^2). \quad (66)$$

Substituting (66) into (65) we obtain

$$I - F'(x_k)(I - 2\Theta_k)\alpha_k F'(x_k)^{-1} = O(h^2).$$

From the last expression (59) immediately follows, i.e. we have proved that $p = 5$ when τ_k and α_k are defined by (36) and (59), respectively. In a similar way, one can show that $p = 5$ for τ_k and α_k defined

by the formulas in the first row of Table 1. It remains to prove for $p = 7$. To do this we use the first divided differences

$$\begin{aligned} [y_k, x_k; F] &= F'(x_k) - \frac{1}{2}F''(x_k)F'(x_k)^{-1}F(x_k) \\ &\quad + \frac{1}{6}F'''(x_k)(F'(x_k)^{-1}F(x_k))^2 + O(h^3), \end{aligned} \quad (67)$$

$$\begin{aligned} [z_k, x_k; F] &= F'(x_k) - \frac{1}{2}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k) \\ &\quad + \frac{1}{6}F'''(x_k)(\tau_k F'(x_k)^{-1}F(x_k))^2 + O(h^3). \end{aligned} \quad (68)$$

Not that in deriving (68) we used the relations (14), (15) and (20). According to (19) and Theorem 2.1, we have $F(z_k) = O(F(x_k)^4)$. Using (19), (20) and (43) we obtain

$$F'(x_k)^{-1}[y_k, x_k; F] = I - \Theta_k - D_k + O(h^3), \quad (69)$$

$$F'(x_k)^{-1}[z_k, x_k; F] = I - \Theta_k - \frac{1}{2}C_k - D_k + O(h^3), \quad (70)$$

where

$$C_k = F'(x_k)^{-1}F''(x_k)\Theta_k F'(x_k)^{-1}F(x_k), \quad (71)$$

$$D_k = -\frac{1}{6}F'(x_k)^{-1}F'''(x_k)(F'(x_k)^{-1}F(x_k))^2. \quad (72)$$

From (69) and (70) we find

$$C_k = 2F'(x_k)^{-1}([y_k, x_k; F] - [z_k, x_k; F]) + O(h^3), \quad (73)$$

and

$$D_k = I - \Theta_k - F'(x_k)^{-1}[y_k, x_k; F] + O(h^3). \quad (74)$$

Expanding $F'(z_k)$ around x_k and using (19) we have

$$F'(z_k) = F'(x_k)(I - (2\Theta_k + C_k + 3D_k) + O(h^3)). \quad (75)$$

Since $C_k = O(F(x_k)^2)$ and $D_k = O(F(x_k)^2)$ then there exists the inverse of matrix in brackets and by Banach lemma we have

$$(I - (2\Theta_k + C_k + 3D_k) + O(h^3))^{-1} = I + 2\Theta_k + C_k + 3D_k + 4\Theta_k^2 + O(h^3).$$

Hence, from (75) we have

$$F'(z_k)^{-1} = (I + 2\Theta_k + C_k + 3D_k + 4\Theta_k^2 + O(h^3))F'(x_k)^{-1}. \quad (76)$$

The Taylor expansion of $F(x_{k+1})$ around z_k gives

$$F(x_{k+1}) = (I - F'(z_k)\alpha_k F'(x_k)^{-1})F(z_k) + O(F(z_k)^2). \quad (77)$$

From (77) it is clear that

$$F(x_{k+1}) = O(F(x_k)^7), \quad (78)$$

provided that

$$I - F'(z_k)\alpha_k F'(x_k)^{-1} = O(h^3), \quad (79a)$$

$$F'(x_k) - F'(z_k)\alpha_k = O(h^3). \quad (79b)$$

From (79b) we get

$$\alpha_k = F'(z_k)^{-1}F'(x_k) + O(h^3). \quad (80)$$

Substituting (76) into (80) we obtain

$$\alpha_k = I + 2\Theta_k + C_k + 3D_k + 4\Theta_k^2 + O(h^3). \quad (81)$$

Using (73) and (74) in (81) we have

$$\alpha_k = I + 2\Theta_k + \left(4\Theta_k^2 + 3(I - \Theta_k) - F'(x_k)^{-1}([y_k, x_k; F] + 2[z_k, x_k; F])\right) + O(h^3).$$

If we use assumption (51) then, by formula (71) we get $C_k = 2\Theta_k^2$ and $\Theta_k^2 = F'(x_k)^{-1}([y_k, x_k; F] - [z_k, x_k; F]) + O(h^3)$. Finally, we get

$$\alpha_k = I + 2\Theta_k + 6F'(x_k)^{-1}([y_k, x_k; F] - [z_k, x_k; F]) + O(h^3). \quad (82)$$

Conversely, let (78) hold. Then from (77) it follows that $F(z_k) = O(F(x_k)^4)$ (Indeed, if $F(z_k) = O(h^3)$ then $F(x_{k+1}) = O(F(x_k)^6)$ for any choice of α_k . This contradicts (78)). By Theorem 2.1 we get (19). As a consequence, the above obtained relations (76), (77) and (82) are valid too. This completes the proof of Theorem 3.1. ■

Moreover, based on the generating functions method one can propose p th-order ($p = 5, 6$) iterative methods (53) with parameter matrices τ_k given by (38) and α_k given by formula

$$\alpha_k = (I - \beta\Theta_k)^{-1}(I + (2 - \beta)\Theta_k), \quad \beta \in R. \quad (83)$$

Remark 3.1: Quite recently Sharma et al. [15] proposed the fifth-order three-step iterative method

$$y_k = x_k - F'(x_k)^{-1}F(x_k), \quad (84a)$$

$$z_k = y_k - 5F'(x_k)^{-1}F(y_k), \quad (84b)$$

$$x_{k+1} = y_k - \frac{9}{5}F'(x_k)^{-1}F(y_k) - \frac{1}{5}F'(x_k)^{-1}F(z_k). \quad (84c)$$

If we compare (84) with (53), then $\bar{\tau}_k = 5I$. Hence by virtue of (20) we have $\tau_k = I + 5\Theta_k = I + O(F(x_k))$. Then according to Theorem 2.1 in [23] we have $F(z_k) = O(F(x_k)^2)$. On the other hand, using relation (17) one can write the third-step in (84) as

$$x_{k+1} = z_k - \frac{1}{5}(I - 16\Theta_k)F'(x_k)^{-1}F(z_k) + O(F(x_k)^3),$$

which does not have such a form as in (53). Hence our theory (Theorem 3.1) does not work in this case.

In Table 2 we cite some existing p th-order methods. It is easy to show that the matrices $\bar{\tau}_k$ and α_k for the methods in Table 2 meet p th-order sufficient conditions in Theorem 2.1 and Theorem 3.1. The proposed p th-order methods with different choices of parameter matrix Θ_k include some existing

Table 2. Some existing pth order methods.

N	Methods	Order	$\bar{\tau}_k$	α_k
1	Cordero et al. [4]		$2I - F'(x_k)^{-1}F(y_k)$	
2	Sharma [12]		$3I - 2F'(x_k)^{-1}[y_k, x_k; F]$	
3	Grau-Sanchez et al. [7]		$(2[y_k, x_k; F] - F'(x_k)^{-1})F'(x_k)$	
4	Madhu et al. [9]	4	$2[y_k, x_k; F]^{-1}F'(x_k) - I$	
4	Sharma et al. [16]		$\tau_k = I - \frac{3}{4}(s_k - I) + \frac{9}{8}(s_k - I)^2$	
5	Grau-Sanchez et al. [6]		$s_k = F'(x_k)^{-1}F'(y_k)$	
5	Xiao et al. [21]		$\tau_k = \frac{1}{2} \left(-I + \frac{9}{4}F'(y_k)^{-1}F'(x_k) + \frac{3}{4}F'(x_k)^{-1}F'(y_k) \right)$	
6	Cordero et al. [3]	5	$\tau_k = \frac{1}{2}(I + F'(y_k)^{-1}F'(x_k))$	$F'(y_k)^{-1}F'(x_k)$
7	Xiao et al. [21]		$2(I - F'(x_k)^{-1}F'(y_k))^{-1}$	$F'(y_k)^{-1}F'(x_k)$
8	Sharma et al. [14]		$y_k = x_k - aF'(x_k)^{-1}F'(x_k)$	
8	Sharma et al. [12]		$\bar{\tau}_k = \left(\left(1 - \frac{1}{2a}\right)I + \frac{1}{2a}F'(x_k)^{-1}F'(y_k) \right)^{-1}$	$-I + 2 \left(\frac{1}{2a}F'(y_k) + \left(1 - \frac{1}{2a}\right)F'(x_k) \right)^{-1}F'(x_k)$
9	Xiao et al. [21]		$a = \frac{1}{2}, (F'(y_k)^{-1})F'(x_k) - I \Theta_k^{-1}$	$2F'(y_k)^{-1}F'(x_k) - I$
10	Grau-Sanchez et al. [7]	6	$3I - 2F'(x_k)^{-1}[y_k, x_k; F]$	$3I - 2F'(x_k)^{-1}[y_k, x_k; F]$
11	Cordero et al. [5]		$y_k = x_k - aF'(x_k)^{-1}F'(x_k)$	
12	Madru et al. [9]		$\tau_k = \frac{1}{2} \left(-I + \frac{9}{4}F'(y_k)^{-1}F'(x_k) + \frac{3}{4}F'(x_k)^{-1}F'(y_k) \right)$	$\left(1 - \frac{1}{a}\right)I + \frac{1}{a}F'(y_k)^{-1}F'(x_k)$
			$(2[y_k, x_k; F] - F'(x_k)^{-1})F'(x_k)$	$(2[y_k, x_k; F] - F'(x_k)^{-1})F'(x_k)$
			$(2[y_k, x_k; F]^{-1} - F'(x_k)^{-1})F'(x_k)$	$2[y_k, x_k; F]^{-1}F'(x_k) - I$
			$a = \frac{1}{2}, (F'(x_k) - 2F'(y_k))^{-1}(3F'(x_k)\Theta_k^{-1} - 4F(x_k))$	$(F'(x_k) - 2F'(y_k))^{-1}F(x_k)$
			$a = \frac{2}{3}, \tau_k = H_1$	$I - \frac{3}{2}(s_k - I) + \frac{1}{2}(s_k - I)^2$

methods as special cases. For example, if we choose Θ_k by formula (32) with $a=1$ then we obtain fifth-order methods given in [4,14,21]. The choice (31) with $a=1$ gives the fourth and sixth-order methods obtained by Sharma et al. [12]. The fifth-order method presented by Grau-Sanchez et al. in [6] is obtained for the choice given by (31) with $a=1$.

Now we consider the following method:

$$\begin{aligned}y_k &= x_k - aF'(x_k)^{-1}F(x_k), \\z_k &= \phi_p(x_k, y_k), \\x_{k+1} &= z_k - \alpha_k F'(x_k)^{-1}F(z_k).\end{aligned}\tag{85}$$

Note that $z_k = \phi_p(x_k, y_k)$ is the iteration function of order $p \geq 2$.

Theorem 3.2: *Let $F : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a sufficiently Fréchet differentiable function in a neighbourhood $D \subseteq \mathbb{R}^n$ containing a zero x^* of $F(x)$. Suppose that $F'(x)$ is continuous and nonsingular in x^* . Then, for an initial approximation sufficiently close to x^* , the sequence $\{x_k\}$, $x_0 \in D$ obtained by (85) has order of convergence $p+2$ if and only if the parameter matrix α_k is given by*

$$\alpha_k = I + 2\Theta_k + O(h^2).\tag{86}$$

Proof: Since $F(z_k) = O(h^p)$, then from (85) we get

$$F(x_{k+1}) = (I - F'(z_k)\alpha_k F'(x_k)^{-1})F(z_k) + O(h^{2p}).\tag{87}$$

If we choose α_k such that

$$I - F'(z_k)\alpha_k F'(x_k)^{-1} = O(h^2),\tag{88}$$

or

$$\alpha_k = F'(z_k)^{-1}F'(x_k) + O(h^2),\tag{89}$$

then from (87) we get

$$F(x_{k+1}) = O(h^{p+2}), \quad p \geq 2.$$

On other hand, the second step in (85) can be rewritten as

$$z_k = \phi_p(x_k, y_k) = x_k - \tau_k F'(x_k)^{-1}F(x_k).$$

Then, we have an expansion

$$F'(z_k) = F'(x_k)(I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k)) + O(h^2).\tag{90}$$

For small $O(F(x_k)) = O(h)$ the matrix $I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k)$ is invertible and hence from (90) we have

$$\begin{aligned}F'(z_k)^{-1} &= (I - F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k))^{-1}F'(x_k)^{-1} \\ &\quad + O(h^2).\end{aligned}\tag{91}$$

Using (10) in (91) we have

$$F'(z_k)^{-1} = (I + F'(x_k)^{-1}F''(x_k)\tau_k F'(x_k)^{-1}F(x_k))F'(x_k)^{-1} + O(h^2).$$

By Theorems 2.4 and 2.5, we can replace τ_k by I in the last expansion without loss of accuracy. Hence by virtue of (14), the (89) has a form

$$\alpha_k = (I + F'(x_k)^{-1}F''(x_k)F'(x_k)^{-1}F(x_k)) + O(h^2) = I + 2\Theta_k + O(h^2).$$

The converse is obvious from (87) and (89). ■

There are many possibilities to obtain α_k satisfying (86). We consider some choices for α_k . First we use approximate formula (33). Substituting (33) into (86) we have

$$\alpha_k = \left(1 + \frac{1}{a}\right) I - \frac{1}{a} F'(x_k)^{-1} F'(y_k). \quad (92)$$

For instance, if $a = \frac{2}{3}$, then (92) leads to [13]

$$\alpha_k = \frac{1}{2} \left(5I - 3F'(x_k)^{-1} F'(y_k)\right).$$

For $a = \frac{1}{2}$, -1 and 1 we obtain

$$\begin{aligned} \alpha_k &= 3I - 2F'(x_k)^{-1} F'(y_k), \\ \alpha_k &= F'(x_k)^{-1} F'(y_k), \end{aligned}$$

and

$$\alpha_k = 2I - F'(x_k)^{-1} F'(y_k) = F'(y_k)^{-1} F'(x_k),$$

respectively. The last case is sixth-order method presented in [3]. Another possible case is to use less accurate formula (34). Using (34) in (86) we get [21]

$$\alpha_k = \left(1 - \frac{1}{a}\right) I + \frac{1}{a} F'(y_k)^{-1} F'(x_k). \quad (93)$$

The simplest generating functions method [22] for (86) is

$$\alpha_k = (I - 2\Theta_k)^{-1}. \quad (94)$$

Using (33) and (34) in (94) we get

$$\alpha_k = \left(\left(1 - \frac{1}{a}\right) F'(x_k) + \frac{1}{a} F'(y_k) \right)^{-1} F'(x_k), \quad (95)$$

and

$$\alpha_k = \left(\left(1 + \frac{1}{a}\right) F'(y_k) - \frac{1}{a} F'(x_k) \right)^{-1} F'(y_k). \quad (96)$$

Thus, we propose four-type choices (92), (93), (95) and (96) for α_k . When $a = \frac{2}{3}$ the formula (95) leads to that of in [2], while when $a = \frac{1}{2}$ the formula (95) leads to that of in [5]. Thus, the Theorem 3.2 is more general than the particular Theorems presented in [2,3,5,13,21].

Remark 3.2: The choices (92), (93), (95) and (96) for α_k and $\bar{\tau}_k$ also valid for three-step method (53) with sixth-order of convergence.

4. Total cost comparison between methods

We check Tables 3–5 to see the total cost of each iteration for each method, where n is the system dimension, $\alpha = n(n-1)(2n-1)/6$, $\beta = n(n-1)$, μ_0 and μ_1 are relative costs of evaluation of F and Jacobian, respectively, in terms of multiplications and ℓ is the relative cost of division in terms of multiplications. The total cost ranges from $n^3/3$ to n^3 , not including lower powers of the dimension n of the system. Here we denote our fourth-order method (2) with Θ_k given by (33) by M4, fifth-order

Table 3. The cost of each iteration for fourth-order methods.

Method	Evaluation of F and Jacobian	Scalar Vector Multiply	Matrix Vector Multiply	Linear Solve	Total
Jarratt [8]	$n\mu_0 + 2n^2\mu_1$	$4n$	0	$3\alpha + 3\beta + \left(\frac{3\beta}{2} + 3n\right)\ell$	$n^3 + \left(2\mu_1 + \frac{3+3\ell}{2}\right)n^2 + \left(\mu_0 + \frac{3+3\ell}{2}\right)n$
Jarratt-like [13]	$n\mu_0 + 2n^2\mu_1$	$4n$	$2n^2$	$\alpha + 3\beta + \left(\frac{\beta}{2} + 3n\right)\ell$	$n^3/3 + \left(2\mu_1 + \frac{9+\ell}{2}\right)n^2 + \left(\mu_0 + \frac{7+15\ell}{6}\right)n$
Cordero et al. [4]	$2n\mu_0 + 2n^2\mu_1$	$3n$	n^2	$\alpha + 3\beta + \left(\frac{\beta}{2} + 3n\right)\ell$	$n^3/3 + \left(2\mu_1 + \frac{7+\ell}{2}\right)n^2 + \left(2\mu_0 + \frac{1+15\ell}{6}\right)n$
M43 [12]	$2n\mu_0 + n^2\mu_1$	$3n$	n^2	$\alpha + 3\beta + \left(\frac{\beta}{2} + 3n\right)\ell$	$n^3/3 + \left(\mu_1 + \frac{7+\ell}{2}\right)n^2 + \left(2\mu_0 + \frac{1+15\ell}{6}\right)n$
M4	$2n\mu_0 + 2n^2\mu_1$	$3n$	n^2	$\alpha + 3\beta + \left(\frac{\beta}{2} + 3n\right)\ell$	$n^3/3 + \left(2\mu_1 + \frac{7+\ell}{2}\right)n^2 + \left(2\mu_0 + \frac{1+15\ell}{6}\right)n$

Table 4. The cost of each iteration for fifth-order methods.

Method	Evaluation of F and Jacobian	Scalar Vector Multiply	Matrix Vector Multiply	Linear Solve	Total
Sharma et al. [14]	$2n\mu_0 + 2n^2\mu_1$	$4n$	0	$2\alpha + 4\beta + (\beta + 4n)\ell$	$2n^3/3 + (2\mu_1 + 3 + \ell)n^2 + \left(2\mu_0 + \frac{1+9\ell}{3}\right)n$
Cordero et al. [3]	$2n\mu_0 + 2n^2\mu_1$	$3n$	0	$3\alpha + 3\beta + \left(\frac{3\beta}{2} + 3n\right)\ell$	$n^3 + \left(2\mu_1 + \frac{3+3\ell}{2}\right)n^2 + \left(2\mu_0 + \frac{1+3\ell}{2}\right)n$
Xiao et al. [19]	$2n\mu_0 + 2n^2\mu_1$	$4n$	0	$2\alpha + 4\beta + (\beta + 4n)\ell$	$2n^3/3 + (2\mu_1 + 3 + \ell)n^2 + \left(2\mu_0 + \frac{1+9\ell}{3}\right)n$
M5	$3n\mu_0 + n^2\mu_1$	$5n$	$2n^2$	$\alpha + 5\beta + \left(\frac{\beta}{2} + 5n\right)\ell$	$n^3/3 + \left(\mu_1 + \frac{13+\ell}{2}\right)n^2 + \left(3\mu_0 + \frac{1+27\ell}{6}\right)n$

Table 5. The cost of each iteration for sixth-order methods.

Method	Evaluation of F and Jacobian	Scalar Vector Multiply	Matrix Vector Multiply	Linear Solve	Total
Jarratt-like [13]	$2n\mu_0 + 2n^2\mu_1$	$6n$	$3n^2$	$\alpha + 5\beta + \left(\frac{\beta}{2} + 5n\right)\ell$	$n^3/3 + \left(2\mu_1 + \frac{15 + \ell}{2}\right)n^2 + \left(2\mu_0 + \frac{7 + 27\ell}{6}\right)n$
Sharma and Arora [12]	$3n\mu_0 + n^2\mu_1$	$5n$	$2n^2$	$\alpha + 5\beta + \left(\frac{\beta}{2} + 5n\right)\ell$	$n^3/3 + \left(\mu_1 + \frac{13 + \ell}{2}\right)n^2 + \left(3\mu_0 + \frac{1 + 27\ell}{6}\right)n$
M61 [12]	$3n\mu_0 + n^2\mu_1$	$3n$	0	$2\alpha + 3\beta + (\beta + 3n)\ell$	$2n^3/3 + (\mu_1 + 2 + \ell)n^2 + \left(3\mu_0 + \frac{1 + 6\ell}{3}\right)n$
M62 [12]	$3n\mu_0 + n^2\mu_1$	$5n$	0	$2\alpha + 5\beta + (\beta + 5n)\ell$	$2n^3/3 + (\mu_1 + 4 + \ell)n^2 + \left(3\mu_0 + \frac{1 + 12\ell}{3}\right)n$
M63 [12]	$3n\mu_0 + n^2\mu_1$	$5n$	$2n^2$	$\alpha + 5\beta + \left(\frac{\beta}{2} + 5n\right)\ell$	$n^3/3 + \left(\mu_1 + \frac{13 + \ell}{2}\right)n^2 + \left(3\mu_0 + \frac{1 + 27\ell}{6}\right)n$
M6	$3n\mu_0 + n^2\mu_1$	$6n$	$3n^2$	$\alpha + 6\beta + \left(\frac{\beta}{2} + 6n\right)\ell$	$n^3/3 + \left(\mu_1 + \frac{17 + \ell}{2}\right)n^2 + \left(3\mu_0 + \frac{1 + 33\ell}{6}\right)n$

method (53) with Θ_k at (32) by M5, and sixth-order method (53) with Θ_k defined by (32) by M6. We chose $a = 1$ for M4-M6. The most expensive method is Cordero et al. [3] closely followed by Jarratt [8] for which the total cost is n^3 . Four methods Sharma et al. [14], Xiao et al. [19,20], M61 [12] and M62 [12] cost $2n^3/3$. All other methods cost $n^3/3$. We note that our methods M4, M5 and M6 all cost $n^3/3$. Hence they are competitive under both of computational cost and order of convergence to existing methods. The efficiency index of methods is given by $E = \rho^{1/C}$, where p is the order of convergence and C is computational cost per iteration. For large system C tends to infinity as $n \rightarrow \infty$. Hence $E \rightarrow 1$ for all methods i.e. E is not distinguished each other, as the scalar equation case. To compare the efficiency of different methods at first needed to be estimated computational cost per iteration (see more details [12,19–21]). It mainly consists of evaluations of functions and derivatives (divided difference) and of estimation of products and quotients for matrix inversion, multiplications of a matrix by a matrix and a vector. Therefore, to estimate the computational cost we first keep in mind the number of above mentioned operations per iteration. Obviously, it is necessary to compare methods by CPU time. In Tables 3–5 we present the number of evaluations per iteration for different fourth-, fifth- and sixth-order methods.

5. Numerical results

In this section, we compare the performance of proposed methods (2) and (53) by several experiments. The numerical experiments have been carried out using Maple 18 computer algebra system with a multi-precision arithmetic 2500 digits. The computer specifications are Microsoft Windows 8.1 Intel(R), Core(TM) i3-4150M CPU, 3.50 GHz with 4046 MB of RAM. We use the following stopping criterion for the methods and test problems

$$\|x_k - x_{k-1}\|_2 \leq 10^{-150}.$$

In order to verify the theoretical order of convergence, we calculate the approximated computational order of convergence p by

$$p = \frac{\ln(\|x_{k+1} - x_k\| / \|x_k - x_{k-1}\|)}{\ln(\|x_k - x_{k-1}\| / \|x_{k-1} - x_{k-2}\|)},$$

(see [3,20]) with the last four approximations in the iterative process. In addition, we include CPU time utilized in the execution of program which is computed by the Maple command “time()”. To testify the order of convergence of the new proposed methods M4, M5, M6, we consider the following Example 5.1. The error $\|x_k - x_{k-1}\|$ of approximations to the corresponding solution of Example 5.1, the number of iterations k and the computational orders of convergence are displayed in Table 6, where (30), (32)–(34) are choices of Θ_k as mentioned in Section 2 and the factor h in the brackets denotes 10^h .

Example 5.1: We consider the system of trigonometric equations (see [12]):

$$x_i - \cos\left(2x_i - \sum_{j=1}^n x_j\right) = 0, \quad 1 \leq i \leq n.$$

For $n = 4$, initial value is $x_0 = (0.75, \dots, 0.75)^T$ and solution of this problem is, $x^* \approx (0.514933264, \dots, 0.514933264)^T$.

Example 5.2: We consider a large-scale nonlinear problem with $n = 51,101$ (selected from [20]):

$$x_{(1)} + x_{(2)} - 2 = 0,$$

Table 6. Computational order of convergence (COC) for Example 5.1.

Methods	Θ_k	k	$\ x_k - x_{k-1}\ $	COC
M4	(30)	5	4.01(-569)	4.00
	(32)	5	1.77(-476)	4.00
	(33)	5	6.79(-543)	4.00
	(34)	5	1.18(-643)	4.00
M5	(30)	5	8.17(-347)	5.00
	(32)	5	7.12(-295)	5.00
	(33)	5	6.39(-322)	5.00
	(34)	5	4.31(-352)	5.00
M6	(30)	4	2.02(-715)	6.00
	(32)	4	7.92(-554)	6.00
	(33)	4	3.05(-652)	6.00
	(34)	4	0	6.00

$$x_{(i)}x_{(i+1)} - 1 = 0, \quad 2 \leq i \leq n-1,$$

$$x_{(n)} + x_{(1)} - 2 = 0.$$

A solution $x^* = (1, 1, \dots, 1)^T$, initial value $x_0 = (0.85, 0.85, \dots, 0.85)^T$.

Example 5.3: Consider the planar 1D Bratu problem [9,10]:

$$\frac{d^2u}{dx^2} + \lambda e^u = 0, \quad \lambda > 0, \quad 0 < x < 1 \quad (97)$$

with the boundary conditions $u(0) = u(1) = 0$. The Bratu problem arises from the study of the radiative heat transfer, the fuel ignition model of thermal combustion, thermal reaction, chemical reactor theory [10].

Table 7. CPU time (in seconds) for Example 5.2 and methods.

Methods	Θ_k	Order	CPU time	
			$n = 51$	$n = 101$
M4	(30)	4	15.8	118.3
	(32)		16.8	122.6
	(33)		8.89	58.0
	(34)		3.93	25.0
M43 [12]	-		15.9	122.7
Jarratt [8]	-		9.18	65.8
Jarratt-like [13]	-		19.9	166
Cordero [3]	-		10.7	70.3
M5	(30)	5	19.6	160.7
	(32)		19.5	161.2
	(33)		14.9	108
	(34)		7.01	46.7
Jarratt [13]	-		10.8	76.6
Cordero et al. [3]	-		12.0	60.0
Xiao et al. [19]	-		15.9	98.0
M6	(30)	6	25.4	196.5
	(32)		26.5	221.1
	(33)		14.9	100
	(34)		6.70	41.7
M63 [12]	-		25.5	197.2
Cordero CM-16 [13]	-		10.8	144
Jarratt-like [13]	-		32.1	240
Cordero [3]	-		12.0	172

Table 8. CPU time (in seconds) for a large-scale nonlinear problem (Example 5.3).

Methods	Θ_k	Order	Error	CPU time
M4	(30)	4	0.4954e-202	57.0
	(32)		0.5148e-202	58.3
	(33)		0.9041e-231	63.1
	(34)		0.4143e-238	21.7
M43 [12]	-		0.4143e-238	59.6
Jarratt [8]	-		0.2405e-714	80.3
Jarratt-like [13]	-		0.5031e-260	96.3
Cordero [3]	-		0.9041e-231	76.3
M5	(30)	5	0.9309e-410	74.4
	(32)		0.9965e-410	76.3
	(33)		0.5954e-451	90.3
	(34)		0.2397e-458	49.2
Jarratt [13]	-		0.6058e-414	76.6
Cordero et al. [3]	-		0.2159e-396	60.0
Xiao et al. [19]	-		0.2093e-344	62.2
M6	(30)	6	0.2086e-662	111.1
	(32)		0.2021e-414	111.9
	(33)		0	121.8
	(34)		0.1047e-750	44.6
M63 [12]	-		0.1254e-611	98.2
Cordero CM-16 [13]	-		0	102.1
Jarratt-like [13]	-		0.1241e-611	118.8
Cordero [3]	-		0.3113e-129	91.2

Using a standard finite-difference scheme, the discrete version of Bratu problem will be

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \lambda e^{u_i} = 0, \quad i = 1, \dots, N-1,$$

with discrete boundary conditions $u_0 = u_N = 0$ and the stepsize $h = 1/N$. There are $N-1$ unknowns ($n = N-1$). It is known that the finite difference scheme converges to the lower solution of the 1D Bratu using the starting vector $u_0 = (0, 0, \dots, 0)^T$. We use $N=121$ and $\lambda = 3.513830719$ (see [9]). Table 8 shows the results about discrete version of 1D Bratu problem solved by the considered methods. In the Table 8, we show the absolute error $\|x_k - x_{k-1}\|$ and CPU time.

From Table 6 we can observe that computed results completely support the theory of convergence discussed in previous sections. In Tables 7 and 8 we have made comparison of methods $M4$, $M5$, $M6$ and existing methods listed in Tables 3–5 based on the CPU time. We have also compared the performance of related methods by the CPU time for large scale nonlinear problem with $n = 51,101$ in Table 7 and $n = 120$ in Table 8. The comparison for large system (see Tables 7 and 8) clearly shows that our methods with the choice given by (34) are the fastest as compared to the other methods with the same order of convergence.

6. Conclusions

We have developed several families of order four, five and six for the solution of systems of nonlinear equations which include some well-known methods as particular cases. Therefore the proposed family of iterations can be considered as extension of some existing iteration methods. The necessary and sufficient conditions for p th order of convergence ($3 \leq p \leq 6$) are given in terms of parameter matrices τ_k and α_k . We suggested several choices of parameter matrix Θ_k determining τ_k . Sufficient convergence conditions in Theorems 2.3, 2.4 and 3.2 can be used effectively to derive new two and three-step iterations with higher order of convergence. We have compared the performance of our methods to existing methods and found that the methods with the choice given

by (34) cost the least and are the fastest as compared to the other methods with the same order of convergence.

Acknowledgements

The authors wish to thank the editor and the anonymous referees for their valuable suggestions and comments on the first version of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported partially by the Foundation of Science and Technology of Mongolia [grant number SST_18/2018]. The research of the author Changbum Chun was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1D1A1A09917373].

References

- [1] S. Amat, S. Busquier, J.M. Gutiérrez, *Geometric constructions of iterative functions to solve nonlinear equations*. J. Comput. Appl. Math. Comput. Appl. Math. 157 (2003), pp. 197–205.
- [2] A. Cordero, J.L. Hueso, E. Martínez and J.R. Torregrosa, *A modified Newton–Jarratt’s composition*, Numer. Algor. 55 (2010), pp. 87–99.
- [3] A. Cordero, J.L. Hueso, E. Martínez and J.R. Torregrosa, *Increasing the convergence order of an iterative method for nonlinear systems*, Appl. Math. Lett. 25 (2012), pp. 2369–2374.
- [4] A. Cordero, E. Martínez and J.R. Torregrosa, *Iterative methods of order four and five for systems of nonlinear equations*, Appl. Math. Comput. 231 (2009), pp. 541–551.
- [5] A. Cordero, J.R. Torregrosa and M.P. Vassileva, *Pseudocomposition: a technique to design predictor-corrector methods for systems of nonlinear equations*, Appl. Math. Comput. 218 (2012), pp. 11496–11504.
- [6] M. Grau-Sánchez and A. Grau, *On the computational efficiency index and some iterative methods for solving systems of nonlinear equations*, J. Comput. App. Math. 236 (2011), pp. 1259–1266.
- [7] M. Grau-Sanchez, A. Grau and M. Noguera, *Ostrowski type methods for solving systems of nonlinear equations*, Appl. Math. Comput. 218 (2011), pp. 2377–2385.
- [8] P. Jarratt, *Some fourth order multipoint iterative methods for solving equations*, Math. Comput. 20(95) (1966), pp. 434–437.
- [9] K. Madhu and J. Jayaraman, *Some higher order Newton-Like methods for solving system of nonlinear equations and its applications*, Int. J. Appl. Comput. Math. 3(3) (2017), pp. 2213–2230. doi:10.1007/s40819-016-0234
- [10] A. Mohsen, *A simple solution of the Bratu problem*, Comput. Math. Appl. 67 (2014), pp. 26–33.
- [11] F.A. Potra, *Nondiscrete Induction and Iterative Processes*, Pitman, London, 1984.
- [12] J.R. Sharma and H. Arora, *On efficient weighted-Newton methods for solving systems of nonlinear equations*, Appl. Math. Comput. 222 (2013), pp. 497–506.
- [13] J.R. Sharma and H. Arora, *Efficient Jarratt-like methods for solving systems of nonlinear equations*, Calcolo 51 (2014), pp. 193–210.
- [14] J.R. Sharma and P. Gupta, *An efficient fifth order method for solving systems of nonlinear equations*, Comput. Math. Appl. 67 (2014), pp. 591–601.
- [15] J.R. Sharma and R.K. Guha, *Simple yet efficient Newton-like method for systems of nonlinear equations*, Calcolo 53 (2016), pp. 451–473.
- [16] J.R. Sharma, R.K. Guha and R. Sharma, *An efficient fourth-order weighted-Newton method for systems of nonlinear equations*, Numer. Algor. 62 (2013), pp. 307–323.
- [17] X. Wang, *A family of Newton-type iterative methods using some special self-accelerating parameters*, Int. J. Comput. Math. 95 (2018), pp. 2112–2127.
- [18] X. Wang, *An Ostrowski-type method with memory using a novel self-accelerating parameters*, J. Comput. App. Math. 330 (2018), pp. 710–720.
- [19] X. Xiao and H. Yin, *A new class of methods with higher order of convergence for solving systems of nonlinear equations*, Appl. Math. Comput. 264 (2015), pp. 300–309.
- [20] X. Xiao and H. Yin, *A simple and efficient method with high order convergence for solving systems of nonlinear equations*, Comput. Appl. Math. 69 (2015), pp. 1220–1231.

- [21] X.Y. Xiao and H.W. Yin, *Increasing the order of convergence for iterative methods to solve nonlinear systems*, *Calcolo* 53 (2016), pp. 285–300.
- [22] T. Zhanlav, O. Chuluunbaatar and V. Ulziibayar, *Generating functions method for construction new iterations*, *Appl. Math. Comput.* 315 (2017), pp. 414–423.
- [23] T. Zhanlav, O. Chuluunbaatar and V. Ulziibayar, *Necessary and sufficient conditions for the convergence of two- and three-point Newton-type iterations*, *Comput. Math. Math. Phys.* 57 (2017), pp. 1090–1100.
- [24] T. Zhanlav and I.V. Puzynin, *The convergence of iteration based on a continuous analogue of Newton's method*, *Comput. Math. Math. Phys.* 32 (1992), pp. 729–737.

COMPARISON OF SOME OPTIMAL DERIVATIVE-FREE
THREE-POINT ITERATIONS

T. ZHANLAV[†] and KH. OTGONDORJ^{†,*}

Abstract. We show that the well-known Khattri et al. [5] methods and Zheng et al. [14] methods are identical. In passing we propose suitable calculation formula for Khattri et al. methods. We also show that the families of eighth-order derivative-free methods obtained in [13] include some existing methods, among them the above mentioned ones as particular cases. We also give the sufficient convergence condition of these families. Numerical examples and comparison with some existing methods were made. In addition, the dynamical behavior of methods of these families is analyzed.

MSC 2010. 65H05.

Keywords. Nonlinear equations, Derivative-free methods, Optimal three point iterative methods.

1. INTRODUCTION

At present there exist many optimal derivative-free three-point iterations see, for example, [1–3, 5–9, 13, 14] and references therein. They mainly distinguished among themselves by approximations of $f'(z_n)$ at the last step. Let the values of $f(x)$ be known at points x_n, w_n, y_n and z_n . Often the following three approaches are used for approximation $f'(z_n)$. The most preferred approximation (see [1],[6, 7, 9],[14]) is

$$(1) \quad f'(z_n) \approx N'_3(z_n),$$

where $N_3(z)$ is Newton's interpolation polynomial of degree three at the point x_n, w_n, y_n and z_n . The second approach is [5]

$$(2) \quad f'(z_n) \approx \nu_1 f(x_n) + \nu_2 f(w_n) + \nu_3 f(y_n) + \nu_4 f(z_n).$$

The real constants ν_1, ν_2, ν_3 and ν_4 are determined such that the relation (2) holds with equality for the four functions $f(x) = 1, x, x^2, x^3$. While in [13]

[†]Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Mongolia e-mail: tzhanlav@yahoo.com.

*School of Applied Sciences, Mongolian University of Science and Technology, Mongolia e-mail: otgondorj@gmail.com.

was used the approximation

$$(3) \quad \begin{aligned} f'(z_n) &\approx af(x_n) + bf(y_n) + cf(z_n) + d\phi(x_n), \\ \phi(x_n) &= \frac{f(w_n) - f(x_n)}{w_n - x_n} = f[x_n, w_n]. \end{aligned}$$

The real constants a, b, c and d in (3) are determined such that the equality (3) holds with accuracy $\mathcal{O}(f(x_n)^4)$. Note that in last years have been appeared papers, in which were used another approximations such as Pade approximant [3] and rational approximations [2] and so on. As we seen from (1), (2) and (3) more suitable and guaranteed approximation is (3). In general, all these three approaches turn out to be identical. This is well-known long ago fact [16]. This idea motivated us to make detail comparison of methods based on (1), (2) and (3). Note that the detail comparison of optimal three-point methods was made in [4] and such comparison for optimal derivative-free methods is still needed. The paper organized as follows. In Section 2 we consider some methods based on the approximations (1), (2), (3) and made comparison of them. We obtain the sufficient convergence condition for these families in Section 3. Numerical and visual comparison some optimal derivative-free methods are made in Section 4.

2. SOME METHODS BASED ON THE APPROXIMATION (1), (2) AND (3)

The well-known Zheng et al. [14] methods (Z8) based on (1) and has a form

$$(4) \quad \begin{aligned} y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \quad w_n = x_n + \gamma f(x_n), \quad \gamma \in \mathbb{R} \setminus \{0\} \\ z_n &= y_n - \frac{f(y_n)}{f[x_n, y_n] + f[y_n, w_n] - f[x_n, w_n]}, \\ x_{n+1} &= z_n - \frac{f(z_n)}{f[z_n, y_n] + (z_n - y_n)f[z_n, y_n, x_n] + (z_n - y_n)F}, \end{aligned}$$

where $F = (z_n - x_n)f[z_n, y_n, x_n, w_n]$. Based on (2) the well-known Khattri et al. [5] methods (KS8) has the following form:

$$(5) \quad \begin{aligned} y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \\ z_n &= y_n - \frac{f(y_n)}{\frac{x_n - y_n + \gamma f(x_n)}{(x_n - y_n)\gamma} - \frac{(x_n - y_n)f(w_n)}{(w_n - y_n)\gamma f(x_n)} - \frac{(2x_n - 2y_n + \gamma f(x_n))f(y_n)}{(x_n - y_n)(w_n - y_n)}}, \\ x_{n+1} &= z_n - \frac{f(z_n)}{H_1 + H_2 + H_3 - H_4}. \end{aligned}$$

Here

$$(6) \quad \begin{aligned} H_1 &= -\frac{(y_n - z_n)(w_n - z_n)}{(x_n - z_n)\gamma(x_n - y_n)}, \\ H_2 &= \frac{(y_n - z_n)(x_n - z_n)f(w_n)}{(w_n - z_n)(w_n - y_n)\gamma f(x_n)}, \\ H_3 &= \frac{(x_n - z_n)(w_n - z_n)f(y_n)}{(y_n - z_n)(w_n - y_n)(x_n - y_n)}, \\ H_4 &= \frac{\gamma(x_n - 2z_n + y_n)f(x_n) + x_n^2 + (-4z_n + 2y_n)x_n + 3z_n^2 - 2y_n z_n}{(y_n - z_n)(x_n - z_n)(w_n - z_n)} f(z_n). \end{aligned}$$

In [5], the authors pointed out that these methods given by (5), (6) is similar to the already known methods proposed in [1, 6, 7, 9], in particular to method in [14], however, they are not the same methods. From (4) and (5) we see that the second and third substeps in (5) are much complicated as compared with (4). The formula, requiring many mathematical operations absolutely unfitted for numerical and stability points of view. Hence, the formula (5) needed further simplifications. The families of derivative-free optimal methods proposed in [13] are based on (3) and have a form

$$(7) \quad \begin{aligned} y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \\ z_n &= y_n - \bar{\tau}_n \frac{f(y_n)}{f[x_n, w_n]}, \\ x_{n+1} &= z_n - \alpha_n \frac{f(z_n)}{f[x_n, w_n]}, \end{aligned}$$

where

$$(8) \quad \bar{\tau}_n = \frac{c + (d_n c + d)\theta_n + \omega\theta_n^2}{c + d\theta_n + b\theta_n^2}, \quad c + d + b \neq 0, \quad c, d, b, \omega \in \mathbb{R}.$$

and

$$(9) \quad \alpha_n = \frac{1}{\left(1 + a_n w_n \left(\frac{f[z_n, x_n]}{f[x_n, w_n]} - 1\right) + b_n \gamma_n \left(\frac{f[z_n, y_n]}{f[x_n, w_n]} - 1\right)\right)},$$

with

$$(10) \quad \begin{aligned} a_n w_n &= (1 - \tau_n) \frac{2\tau_n + \gamma\phi_n + (\tau_n + \gamma\phi_n)^2}{(\tau_n + \gamma\phi_n)(1 + \gamma\phi_n)}, \\ b_n \gamma_n &= \frac{\tau_n(\tau_n + \gamma\phi_n)}{1 + \gamma\phi_n}, \quad \phi_n = f[x_n, w_n], \\ \tau_n &= 1 + \bar{\tau}_n \theta_n, \quad \theta_n = \frac{f(y_n)}{f(x_n)}. \end{aligned}$$

We call the representation (7) of three-point methods as canonical form. Each derivative-free three-point methods, in particular the methods (4) and (5) can be written in canonical form uniquely. Note that all the considered methods (4), (5) and (7) are optimal in the sense of Kung and Traub [17]. So they has an efficiency index $8^{1/4} \approx 1.68179$. The methods (4) and (5) contain one free parameter γ , whereas the methods (7) contain, in addition γ , yet four parameter c , d , b and w . Hence, in our opinion, the families (7) represent a wide class of optimal three-point methods. Our aim is to compare the above mentioned methods in detail. First, we will show that the optimal derivative-free methods (4) and (5) are identical. Namely, we obtain

THEOREM 1. *The optimal derivative-free methods (4) and (5) are equivalent.*

Proof. Using easily verifying relations

$$(11) \quad f[x_n, y_n] = \phi_n(1 - \theta_n), \quad f[y_n, w_n] = \phi_n\left(1 - \frac{\theta_n}{1 + \gamma\phi_n}\right),$$

the second-step in (4) and (5) can be easily rewritten as

$$(12) \quad z_n = y_n - \bar{\tau}_n \frac{f(y_n)}{f[x_n, w_n]},$$

where

$$(13) \quad \bar{\tau}_n = \frac{1}{1 - \hat{d}_n \theta_n}, \quad \hat{d}_n = \frac{2 + \gamma \phi_n}{1 + \gamma \phi_n}.$$

Thus, the first two sub-steps of (4) and (5) are the same. In passing, we obtain very simple calculation formula (12) for iteration method (5). It remains to compare the third sub-steps in (4) and (5). The third sub-steps in (4) and (5) can be rewritten as

$$x_{n+1} = z_n - \alpha_n \frac{f(z_n)}{f[x_n, w_n]},$$

where

$$(14) \quad \alpha_n = \frac{\phi_n}{f[z_n, y_n] + (z_n - y_n)f[z_n, y_n, x_n] + (z_n - y_n)F}$$

for iteration (4) and

$$(15) \quad \alpha_n = \frac{\phi_n}{H_1 + H_2 + H_3 - H_4},$$

for iteration (5). Using the following relations

$$(16) \quad \begin{aligned} f[z_n, y_n] &= \frac{\phi_n}{\bar{\tau}_n} (1 - v_n), \quad v_n = \frac{f(z_n)}{f(y_n)}, \quad f[z_n, x_n] = \frac{\phi_n}{\tau_n} (1 - \theta_n v_n), \\ f[z_n, y_n, x_n] &= \frac{\phi_n^2}{\tau_n f(x_n)} \frac{\bar{\tau}_n (1 - \theta_n) - (1 - v_n)}{\bar{\tau}_n}, \\ f[z_n, y_n, x_n, w_n] &= \frac{\phi_n^3}{f^2(x_n) (\tau_n + \gamma \phi_n)} \left(\frac{\theta_n}{1 + \gamma \phi_n} - \frac{\bar{\tau}_n - \tau_n + v_n}{\bar{\tau}_n \tau_n} \right), \end{aligned}$$

one can write (14) as:

$$(17) \quad \alpha_n = \frac{1}{\frac{\tau_n (\tau_n + \gamma \phi_n)}{(\tau_n - 1)(1 + \gamma \phi_n)} \theta_n + (1 - \tau_n) \frac{2\tau_n + \gamma \phi_n}{\tau_n (\tau_n + \gamma \phi_n)} - Q \theta_n v_n},$$

where $Q = \frac{\tau_n (3\tau_n - 2) + \gamma \phi_n (2\tau_n - 1)}{\tau_n (\tau_n - 1) (\tau_n + \gamma \phi_n)}$. In a similar way, using (16), the expression (15) can be easily rewritten as (17). Thus, the third-step of (4) and (5) also coincide with each other.

Therefore, the iterations (4) and (5) are identical. \square

So the methods (5) can be considered as rediscovered variant of Zheng et al. [14] ones. Now, we use the relations (16) in (9). After some algebraic manipulations we again arrive at (17). It means that the third sub-step of iterations (4) and (7) are the same.

Therefore, the iterations (4), (5) and (7) can be written in more convenient and unified form as:

$$(18) \quad \begin{aligned} y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \\ z_n &= y_n - \bar{\tau}_n \frac{f(y_n)}{f[x_n, w_n]}, \\ x_{n+1} &= z_n - \frac{f(z_n)}{f[z_n, y_n] + (z_n - y_n)f[z_n, y_n, x_n] + (z_n - y_n)F}, \end{aligned}$$

where $\bar{\tau}_n$ is given by (8) for (7) and is given by (13) for (4) and (5). When $c = 1$, $d = -\hat{d}_n$ and $\omega = b = 0$ in (8), $\bar{\tau}_n$ coincides with (12). In this case the iterations (4) and (5) and (18) are identical. So our iterations (7) contain the methods (4) and (5) as particular cases. In addition, the iterations (18) contain some well-known iterations as particular cases (see Table 1).

Later on, we denote the method (18) with $c = 1$, $d = -\hat{d}_n$, $b = -\frac{1}{1+\gamma\phi_n}$ and $\omega = 0$ by M1. These parameters are chosen to have a large region of convergence and a big basin of attraction for family (18). Moreover, the iteration

c	d	b	w	$\bar{\tau}_n$	methods
1	$-\hat{d}_n$	0	0	$\frac{1}{1-\hat{d}_n\theta_n}$	(Z8), (KS8)
1	$-\frac{1}{1+\gamma\phi_n}$	0	$\frac{a\hat{d}_n}{2}$	$\frac{1+\theta_n+a\hat{d}_n\frac{\theta_n^2}{2}}{1-\frac{\theta_n}{1+\gamma\phi_n}}$	Lotfi (L8) [6]
1	$\beta - 1 - \hat{d}_n$	$\frac{2-\beta}{1+\gamma\phi_n}$	β	$\frac{1+(\beta-1)\theta_k+\beta\theta_k^2}{1+(\beta-2-\frac{1}{1+\gamma\phi_k})\theta_k+\frac{\beta-2}{1+\gamma\phi_k}\theta_k^2}$	King's type (K8) [7]
1	$-\frac{1}{1+\gamma\phi_n}$	0	0	$\frac{1+\theta_n}{1-\frac{\theta_n}{1+\gamma\phi_n}}$	Sharma (S8)[9]
1	$-2\alpha - \frac{1}{1+\gamma\phi_n}$	$\frac{2\alpha}{1+\gamma\phi_n}$	$H(\theta_n)$	$\frac{1}{1-2\alpha\theta_n} \frac{H(\theta_n)}{(1-\frac{\theta_n}{1+\gamma\phi_n})}$	Chebyshev-Halley (CH8)[1]
1	$-\hat{d}_n$	$\frac{\hat{d}_n^2}{4}$	0	$\frac{1}{(1-\frac{\hat{d}_n}{2}\theta_n)^2}$	[4]
1	$-\hat{d}_n$	$\frac{1}{1+\gamma\phi_n}$	0	$\frac{1}{1-\hat{d}_n\theta_n+\frac{1}{1+\gamma\phi_n}\theta_n^2}$	Thukral (T8)[12]
					Kung-Traub (KT8)[17]
1	$-\hat{d}_n$	$\frac{1}{1-\phi_n}$	0	$\frac{1}{(1-\frac{\theta_n}{1-\phi_n})(1-\theta_n)}$	Soleymani (SS8) [10]
1	-1	0	$\frac{1}{(1+\gamma\phi_n)^2}$	$(1 + \frac{\theta_n}{(1+\gamma\phi_n)} + \frac{\theta_n^2}{(1+\gamma\phi_n)^2}) \frac{1}{1-\theta_n}$	Soleymani (SV8) [11]
1	$-\hat{d}_n$	$-\frac{1}{1+\gamma\phi_n}$	0	$\frac{1}{1-\hat{d}_n\theta_n-\frac{\theta_n^2}{1+\gamma\phi_n}}$	M1

Table 1. Choices of parameters for methods.

(18) can be rewritten as

$$\begin{aligned}
 (19) \quad y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \quad w_n = x_n + \gamma f(x_n), \quad \gamma \in \mathbb{R} \setminus \{0\} \\
 z_n &= \psi_4(x_n, y_n, z_n), \\
 x_{n+1} &= z_n - \frac{f(z_n)}{f[z_n, y_n] + (z_n - y_n)f[z_n, y_n, x_n] + (z_n - y_n)F},
 \end{aligned}$$

where ψ_4 is any optimal fourth order derivative-free method. From (19) we see that the each iteration of the family of derivative-free optimal three-point iterations (19) essentially depends on the choice ψ_4 or the choice of iteration parameter $\bar{\tau}_n$ in (18).

3. CONVERGENCE ANALYSIS

Generally, the convergence properties of family of iterations (18) essentially depend on the convergence of iterations consisting of the first two sub-steps

in (18) i.e.,

$$(20) \quad \begin{aligned} y_n &= x_n - \frac{f(x_n)}{f[x_n, w_n]}, \\ z_n &= y_n - \bar{\tau}_n \frac{f(y_n)}{f[x_n, w_n]}, \end{aligned}$$

where $\bar{\tau}_n$ is given by (8). It is easy to show that if the iterations (20) converge then its convergence order is four. Moreover, if the iterations (20) converge, so does (18) with convergence order eight. From this clear that in order to establish the convergence of (18) it suffice to establish the convergence of iterations (20). To this end we use Taylor expansion of function $f \in C^2(I)$ and another form of second-step in (20) as

$$(21) \quad z_n = x_n - \tau_n \frac{f(x_n)}{\phi_n}, \quad \tau_n = 1 + \bar{\tau}_n \theta_n.$$

As a result, we have

$$(22) \quad f(z_n) = \left(1 - \frac{f'(x_n)}{\phi_n} \tau_n + \frac{w_n f'(x_n)^2}{2 \phi_n^2} \tau_n^2\right) f(x_n).$$

where

$$(23) \quad w_n = \frac{f''(\xi_n) f(x_n)}{f'(x_n)^2}.$$

From (22) it follows

$$(24) \quad |f(z_n)| \leq \bar{q} |f(x_n)|,$$

where

$$(25) \quad \bar{q} = \left|1 - \eta_n + \frac{w_n}{2} \eta_n^2\right|, \quad \eta_n = \frac{f'(x_n)}{\phi_n} \tau_n.$$

From (24) we see that the convergence of iterations (20) is expected only when

$$(26) \quad \bar{q} < 1.$$

Thus, it suffice to find conditions for which (26) holds true. It is easy to prove that

LEMMA 1. *Let the $w_n \in (-2, 1)$. Then the inequality (26) holds true under conditions:*

$$(27a) \quad 0 < \eta_n < 2 \quad \text{when} \quad 0 < w_n < 1,$$

$$(27b) \quad 0 < \eta_n < 1 \quad \text{when} \quad -2 < w_n < 0.$$

THEOREM 2. *Let $1 + \gamma \phi_n > 0$ and $w_n \in (-2, 1)$. Then the two-point iterative methods (20) converge under condition*

$$(28) \quad |\theta_n| < 1 + \gamma \phi_n.$$

Proof. Using the following relations

$$\frac{f'(x_n)}{\phi_n} = 1 - \frac{\gamma\phi_n}{1 + \gamma\phi_n}\theta_n + \mathcal{O}(f_n^2),$$

and

$$\tau_n = 1 + \theta_n + \hat{d}_n\theta_n^2 + \dots,$$

in (25) we obtain

$$(29) \quad \eta_n = 1 + \frac{1}{1 + \gamma\phi_n}\theta_n + \mathcal{O}(f_n^2).$$

If we use (29) then the condition (27) can be written in term of θ_n as (28) within the accuracy $\mathcal{O}(f^2(x_n))$. In other words, (26) holds true under condition (28). \square

From (8) we obtain

$$(30) \quad \bar{\tau}_n - 1 = \frac{\theta_n(\hat{d}_n c + (\omega - b)\theta_n)}{c + d\theta_n + b\theta_n^2} = \frac{\theta_n\varphi(\theta_n)}{c + d\theta_n + b\theta_n^2},$$

where

$$\varphi(\theta_n) = \hat{d}_n c + (\omega - b)\theta_n.$$

Let $|\omega - b| < \hat{d}_n c$. Then $\varphi(\theta_n) > 0$ on $\theta_n \in [-1, 1]$. Then from (30) we deduce that the following relations

$$\bar{\tau}_n \rightarrow 1, \quad \theta_n \rightarrow 0,$$

are equivalent and the convergence of sequences $f(z_n)$ and θ_n to zero as $n \rightarrow \infty$ expected simultaneously with equal order four. On the other hand, the iteration (20) can be considered as damped Newton's method

$$y_n = x_n - \frac{f(x_n)}{f'(x_n)}\eta_n,$$

with damping parameter η_n given by (28). As is known that, the damped Newton's method converges [15] if

$$(31) \quad 0 < \eta_n < 2.$$

In term of θ_n the condition (31) gives the same result (29).

4. NUMERICAL EXPERIMENTS AND DYNAMICAL BEHAVIOR

In this section, we will give a numerical comparison of our method M1 with other well known optimal eighth order methods listed in Table 1. For this purpose, we consider several test functions given in Table 2. In particular, $f_2 = 0$ is Kepler's equation which relates the eccentric anomaly E , the mean anomaly M and the eccentricity ϵ in an elliptic orbit.

Additionally, we will make comparison of method M1 and other methods based on the dynamical behaviour.

Test functions	Roots
1. $f_1 = \exp(-x^2 + x + 2) + \sin(\pi x) \exp(x^2 + x \cos(x) - 1) + 1$, [6]	$x^* \approx 1.55$
2. $f_2 = M - E + \epsilon \sin(E)$, $0 < \epsilon < 1$, [1]	$x^* \approx 0.38$

Table 2. Nonlinear functions.

Further, we will use the abbreviated names for methods (see last column of Table 1). In Tables 3 to 5, we consider method (CH8) using the weight function $H(\theta_n) = 1 + (1 - 2\alpha)\theta_n$ with values of the parameter $\alpha = 0, \pm 1$ (see [1]), method (L8) for $(a = 0, \pm 1)$ and method (K8) for $(\beta = 0, \pm 1)$. In addition to compare family (18) with other methods we also consider some optimal methods, which third substeps are different from method (18). Namely, we used the following substeps:

Derivative-free Soleymani et al. [11] three-step method (SV8) has the following substep:

$$x_{n+1} = z_n - \frac{f(z_n)}{f[z_n, y_n]} \left(1 - \frac{1}{f[x_n, w_n] - 1} \left(\frac{f(y_n)}{f(x_n)} \right)^2 + (2 - f[z_n, y_n]) \frac{f(z_n)}{f(w_n)} \right).$$

Derivative-free Kung-Traub's [17] three-step method (KT8) has the following substep:

$$x_{n+1} = z_n - \frac{f(y_n)f(w_n)(y_n - x_n + f(x_n)/f[x_n, z_n])}{(f(y_n) - f(z_n))(f(w_n) - f(z_n))}.$$

Derivative-free Thukral's [12] three-step method (T8) has the following substep:

$$x_{n+1} = z_n - \left(1 - \frac{f(z_n)}{f(w_n)} \right)^{-1} \times \left(1 + \frac{2f(y_n)^3}{f(w_n)^2 f(x_n)} \right)^{-1} \left(\frac{f(z_n)}{f[z_n, y_n] - f[x_n, y_n] + f[z_n, x_n]} \right).$$

Derivative-free Soleymani et al. [10] three-step method (SS8) has the following substep:

$$x_{n+1} = z_n - \frac{f(z_n)f(w_n)}{(f(w_n) - f(y_n))f[x_n, y_n]} \times \left(1 + \frac{f(z_n)}{f(y_n)} \right) \left(1 + (2 - f[x_n, w_n]) \frac{f(z_n)}{f(w_n)} \right) \times \left(1 + \left(\frac{f(z_n)}{f(x_n)} \right)^2 \right) \left(1 + (1 - f[x_n, w_n]) \left(\frac{f(y_n)}{f(w_n)} \right)^2 \right).$$

All computations are carried out using Maple18 computer algebra system with 1000 digits. We use the following stopping criterion for the methods: $|x_n - x^*| \leq \epsilon$ where $\epsilon = 10^{-50}$ and x^* is the exact solution of the considered equation. In all examples, we consider that the parameter $\gamma = -0.01$.

To check the theoretical order of convergence of methods, we calculated the computational order of convergence ρ (see [19–21]) using formula

$$\rho \approx \frac{\ln(|x_n - x^*|/|x_{n-1} - x^*|)}{\ln(|x_{n-1} - x^*|/|x_{n-2} - x^*|)},$$

where x_n, x_{n-1}, x_{n-2} are last three consecutive approximations in the iteration process. In Tables 3 and 4, we use test functions f_1, f_2, f_3 and exhibit the iteration numbers n , the absolute value $|x_n - x^*|$ and the computational order of convergence ρ . When the iteration diverges for the considered initial guess x_0 , we denote it by '-'. From Tables 3 and 4 we see that the convergence order of all the methods in Table 1 confirmed by numerical experiments. From the result of Tables 3 and 4, we can observe that the region of convergence of methods M1 and Z8 are wider than that of other considered methods.

Methods	n	$ x_n - x^* $	ρ	n	$ x_n - x^* $	ρ
	$x^* = 1.55 \quad x_0 = 0.8$			$x^* = 1.55 \quad x_0 = 1$		
M1	3	0.5590e-58	7.94	3	0.3688e-69	7.98
Z8		—	—	3	0.8486e-64	7.93
L8	($a = 0$)	—	—	3	0.2124e-57	7.88
	($a = -1$)	—	—	3	0.4607e-55	7.86
	($a = 1$)	—	—	3	0.4097e-60	7.91
K8	($\beta = 0$)	—	—	3	0.2369e-64	7.93
	($\beta = -1$)	—	—	3	0.7934e-65	7.92
	($\beta = 1$)	—	—	3	0.4687e-64	7.94
S8		—	—	3	0.2124e-57	7.88
CH8	($\alpha = 0$)	—	—	3	0.2734e-60	7.90
	($\alpha = -1$)	—	—	3	0.1654e-54	7.85
	($\alpha = 1$)	—	—	3	0.1648e-70	7.97
[4]		—	—	3	0.2639e-60	7.90
SS8		—	—	4	0.8295e-191	7.99
T8		—	—	4	0.1898e-204	7.99
SV8		—	—	4	0.2008e-174	7.99
KT8		—	—	4	0.4856e-324	8.00

Table 3. Comparison of various iterative methods for $f_1(x)$

Generally, higher order convergence methods consist of multi-steps which may use more evaluations of functions than the original one. In this case, multi-point methods may have the extraneous fixed points (black points). In order to find the extraneous fixed points, we rewrite any three-point method as [4]:

$$x_{n+1} = x_n - \frac{f(x_n)}{f[x_n, w_n]} H_f(x_n),$$

where $H_f = 1 + \theta_n(\bar{\tau}_n + \alpha_n v_n)$. Clearly, the root x^* of $f(x)$ is a fixed point of the method. The points $\xi \neq x^*$ for which $H_f(\xi) = 0$ are also fixed points of

Methods		n	$ x_n - x^* $	ρ
$x^* = 0.38$ $x_0 = 1$				
M1		3	0.3388e-266	8.00
Z8		3	0.4157e-250	8.00
L8	($a = 0$)	3	0.1081e-232	8.00
	($a = -1$)	3	0.6929e-256	8.00
	($a = 1$)	3	0.2358e-222	8.00
K8	($\beta = 0$)	3	0.2496e-219	8.00
	($\beta = -1$)	3	0.1278e-252	7.99
	($\beta = 1$)	3	0.4075e-217	7.99
S8		3	0.1081e-232	8.00
CH8	($\alpha = 0$)	3	0.1081e-232	8.00
	($\alpha = -1$)	3	0.6152e-205	7.99
	($\alpha = 1$)	3	0.2496e-219	7.99
[4]		3	0.5045e-221	8.00
SS8		3	0.1295e-199	7.99
T8		3	0.2398e-206	7.99
SV8		3	0.2008e-174	7.99
KT8		3	0.4915e-151	7.99

Table 4. Comparison of various iterative methods for $f_2(x)$

the method. These fixed points are called extraneous fixed points. As we all know, a fixed point ξ is called:

- attractive if $|R'(\xi)| < 1$,
- repulsive if $|R'(\xi)| > 1$,
- parabolic if $|R'(\xi)| = 1$,

where $R(z) = z - \frac{f(z)}{f'[z,w]} H_f(z)$ is the iteration function.

In addition, if $|R'(\xi)| = 0$, the fixed point is superattracting. Now, we will discuss the extraneous fixed points of each method for comparison. To make it easier, we have taken the simple quadratic polynomial $p(z) = z^2 - 1$, whose roots are $z = \pm 1$.

In [Table 5](#), we have collected the extraneous fixed points of the methods Z8, KS8, M1. Next nine methods are analyzed and found that they are unable to compare with other methods. These methods have more than 20 extraneous fixed points. Therefore, we have not include those results in [Table 5](#). For methods Z8 and KS8, we found that the methods have same ten extraneous fixed points. All fixed points are repulsive.

The basin of attraction of iterative methods is another tool for comparing them. Thus, we compare our methods (18) with other methods by using the basins of attraction for polynomials $p(z) = z^3 - 1$.

To illustrate the behavior of the iterative methods, We take 600×600 equally spaced points in the square $[-3, 3] \times [-3, 3] \subset C$. In [Fig. 1](#), the basin of attraction for 12 methods are displayed. The red, green and blue colors are assigned for the attraction basin of the three zeros and the roots of function are marked with white points. Black color is shown lack of convergence to any of the roots. In this cases, the stopping criterion $\varepsilon = 10^{-4}$ and maximum of 25 iterations are used. These dynamical planes have been generated by using the Mathematica 11. From [Fig. 1](#) and [Table 5](#), we can also see that methods M1 and Z8 is much more stable than the others. It can be observed from the figures that the methods M1 along with the existing methods Z8 have wide attraction basins to corresponding zeros than other methods. Z8 also has the least number of black points.

5. CONCLUSION

We have shown that the well-known Khattri et al. [5] methods and Zheng et al. [14] methods are identical. For the Khattri methods, we propose a suitable calculation formula (18) instead of (5). Our proposed method (18) represents wide class of optimal derivative-free iterations. The method (18) contain some well known iterations as particular cases (see Table 1). The comparison of some eighth-order methods was made from the dynamic behavior of view. We observe that the methods M1 and Z8 are much more stable than the others. Note that the family of derivative-free methods (18) can be extended to the systems of nonlinear equations and this study is currently ongoing.

ACKNOWLEDGEMENTS. The authors wish to thank the editor and the anonymous referees for their valuable suggestions and comments, which improved paper. This work was supported by the Foundation of Science and Technology of Mongolian under grant SST_18/2018.

Methods	The extraneous fixed points ξ	Numbers of ξ
Z8	$-0.555220397255420 \pm 1.15928646739103i$ $-0.460115602837211 \pm 0.456390703516719i$ $-0.450000501793328 \pm 0.129063966758804i$ $1.89303155290658 \pm 0.233570409469479i$ $1.79931236664623, 2.67863086464586$	10
KS8	$-0.555220397255420 \pm 1.15928646739103i$ $-0.460115602837211 \pm 0.456390703516719i$ $-0.450000501793328 \pm 0.129063966758804i$ $1.89303155290658 \pm 0.233570409469479i$ $1.79931236664623, 2.67863086464586$	10
M1	$-0.676558832763406 \pm 1.36018262584118i$ $-0.624888463964184 \pm 0.20890104128772i$ $-0.493766364512498 \pm 0.607060501953625i$ $-0.461962845726289 \pm 0.221119195986523i$ $-0.204327487662501 \pm 0.86651046669376i$ $1.932083323 \pm 0.1163156841i$ $2.004864313 \pm 0.7365790432i$ $2.083325978 \pm 0.4554281653i$	16

^aTo save space, we do not include other points in Table 5.

Table 5. The extraneous fixed points.

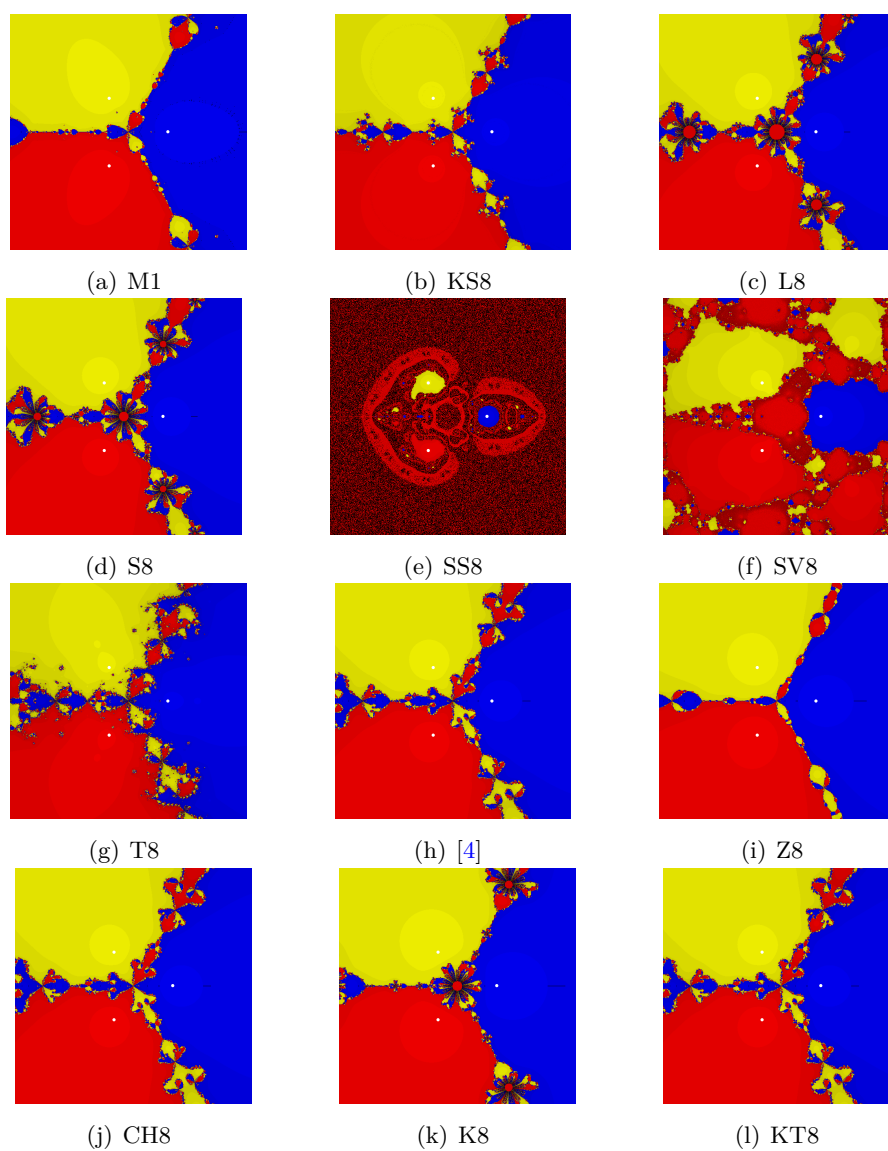



Fig. 1. (color online) Basins of attraction of different derivative-free three-point iterations on $z^3 - 1$.

REFERENCES

- [1] I.K. ARGYROS, M. KANSAL, V. KANWAR, S. BAJAJ, *Higher-order derivative-free families of Chebyshev-Halley type methods with or without memory for solving nonlinear equations*, Appl. Math. Comput., **315** (2017), pp. 224–245. [✉](#)
- [2] R. BEHL, D. GONZALEZ, P. MAROJU, S.S. MOTSA, *An optimal and efficient general eighth-order derivative-free scheme for simple roots*, J. Comput. Appl. Math., **330** (2018), pp. 666–675. [✉](#)
- [3] A. CORDERO, J.L. HUESO E. MARTINEZ, J.R. TORREGROSA, *A new technique to obtain derivative-free optimal iterative methods for solving nonlinear equations*, J. Comput. Appl. Math., **252** (2013), pp. 95–102. [✉](#)
- [4] C. CHUN, B. NETA, *Comparative study of eighth-Order methods for finding simple roots of nonlinear equations*, Numer. Algor., **74** (2017), pp. 1169–1201. [✉](#)
- [5] S.K. KHATTRI, T. STEIHAUG, *Algorithm for forming derivative-free optimal methods*, Numer. Algor., **65** (2014), pp. 809–824. [✉](#)
- [6] T. LOTFI, F. SOLEYMANI, M. GHORBANZADEH, P. ASSARI, *On the construction of some tri-parametric iterative methods with memory*, Numer. Algor., **70** (2015), pp. 835–845. [✉](#)
- [7] S. SHARIFI, S. SIEGMUND, M. SALIMI, *Solving nonlinear equations by a derivative-free form of the King's family with memory*, Calcolo, **53** (2016), pp. 201–215. [✉](#)
- [8] M. PETKOVIĆ, B. NETA, L. PETKOVIĆ J. DZUNIĆ, *Multipoint Methods for Solving Nonlinear Equations*, Elsevier, 2013.
- [9] J.R. SHARMA, R.K. GUHA, P. GUPTA, *Some efficient derivative free methods with memory for solving nonlinear equations*, Appl. Math. Comput., **219** (2012), pp. 699–707. [✉](#)
- [10] F. SOLEYMANI, S. SHATEYI, *Two optimal eighth-order derivative-free classes of iterative methods*, Abstr. Appl. Anal., **2012**, ID 318165, pp. 1–14. [✉](#)
- [11] F. SOLEYMANI, S.K. VANANI, *Optimal Steffensen-type methods with eighth order of convergence*, Comput. Math. Appl., **62** (2011), pp. 4619–4626. [✉](#)
- [12] R. THUKRAL, *Eighth-Order iterative Methods without derivatives for solving nonlinear equations*, ISRN. Appl. Math., **2011**, ID 693787, pp. 1–12. [✉](#)
- [13] T. ZHANLAV, O. CHULUUNBAATAR, KH. OTGONDORJ, *A derivative-free families of optimal two-and three-point iterative methods for solving nonlinear equations*, Comput. Math. Math. Phys., **59** (2019), pp. 920–936.
- [14] Q. ZHENG, J. LI, F. HUANG, *An optimal Steffensen-type family for solving nonlinear equations*, Appl. Math. Comput., **217** (2011), pp. 9592–9597. [✉](#)
- [15] T. ZHANLAV, O. CHULUUNBAATAR, G. ANKHBAYAR, *Relationship between inexact Newton method and the continuous analogy of Newton's method*, J. Numer. Anal. Approx. Theory, **40** (2011) no.2, pp. 182–189.
- [16] R.W. HAMMING, *Numerical methods for scientist and engineers*, McGraw-Hill, New-York, 1962.
- [17] H.T. KUNG, J.F. TRAUB, *Optimal order of one-point and multi-point iteration*, J. Assoc. Comput. Math., **21** (1974), pp. 643–651. [✉](#)
- [18] H. VEISEH, T. LOTFI, T. ALLAHVIRANLOO, *A study on the local convergence and dynamics of the two-step and derivative-free Kung-Traub's method*, Comp. Appl. Math., **37** (2018), pp. 2428–2444. [✉](#)

-
- [19] E. CĂȚINAȘ, *A survey on the high convergence orders and computational convergence orders of sequences*, Appl. Math. Comput., **343** (2019), pp. 1–20. 
- [20] F. A. POTRA, *Nondiscrete Induction and Iterative Processes*, Pitman, London, 1984.
- [21] J.M. ORTEGA, W.C. RHEINBOLDT, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

Received by the editors: March 14, 2019; accepted: March 4 2020; published online: August 11, 2020.

On the Optimal Choice of Parameters in Two-Point Iterative Methods for Solving Nonlinear Equations

T. Zhanlav^{a,*} and Kh. Otgondorj^{a,b,**}

^a Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulan-Bator, 13330 Mongolia

^b School of Applied Sciences, Mongolian University of Science and Technology, Ulan-Bator, 14191 Mongolia

*e-mail: tzhanlav@yahoo.com

**e-mail: otgondorj@gmail.com

Received November 5, 2019; revised July 7, 2020; accepted September 18, 2020

Abstract—A new optimal two-parameter class of derivative-free iterative methods with the application to the Hansen–Patrick type iterations is developed. Using self-accelerating parameters, new higher order methods with memory are obtained. Exact analytical formulas for the optimal values of the parameters are found for the first time. The convergence order is increased from four to seven without any additional computations. Thus, the proposed methods with memory have a high computational efficiency. Numerical examples and comparison with some other available methods confirm the theoretical results and high computational efficiency.

Keywords: nonlinear equations, two-point iterations, methods with memory, optimal methods

DOI: 10.1134/S0965542520120180

1. INTRODUCTION

In numerical analysis and engineering applications, it is often required to solve a nonlinear equation $f(x) = 0$, where $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ is a scalar function defined on an open interval D . The main methods for solving this equation are the Newton method given by (see [1] and references therein)

$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ ($n \geq 0$) and Steffensen's method [13] defined by

$$x_{n+1} = x_n - \frac{f(x_n)^2}{f(x_n + f(x_n)) - f(x_n)} \quad (n \geq 0).$$

In recent years, a large number of higher order iterative methods have been proposed [1–6] in which the concept of increasing the order of convergence was introduced. An advantage of these methods is that they rapidly converge to the solution. However, as the order of an iterative method increases, the number of function computations at each step also increases. Recently, a number of simple two-parameter two-step methods with and without memory have been proposed [2, 8, 13, 14]. The authors of these papers used symbolic computation for obtaining the order of convergence and the error equation. This makes computations considerably less tedious. Usually, the error equation includes the iteration parameters. A good choice of these parameters not only improves the order of convergence but also helps design new iterative methods with memory. The main purpose of this paper is to find the optimal parameters τ_n and γ , λ in two-point iterative methods. Analytical formulas for γ and λ are obtained without using symbolic computation.

In Section 2, optimal two-point derivative-free Hansen–Patrick iterations are obtained. In Section 3, we propose a family of two-point optimal iterations and prove a local convergence theorem. In Section 4, new two-point iterations with and without memory are proposed. In Section 5, we present numerical results that confirm the theoretical conclusion about the order of convergence and provide a comparison with other known methods of the same convergence order.

2. OPTIMAL TWO-POINT ITERATIONS

Consider two-point iterations

$$y_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x_{n+1} = x_n - \tau_n \frac{f(x_n)}{f'(x_n)}, \quad (2.1)$$

where τ_n is the iteration parameter. It is known that the optimal choice of the parameter extends the convergence region and improves the convergence rate of iterations (2.1). A sufficient condition for the fourth-order convergence [3] is

$$\tau_n = 1 + \theta_n + 2\theta_n^2 + O(f(x_n)^3), \quad (2.2)$$

where

$$\theta_n = \frac{f(y_n)}{f(x_n)}. \quad (2.3)$$

Condition (2.2) is often used not only for checking the convergence order of iterations (2.1) but also for designing new optimal methods. For clearness, we recall some definitions required for the following presentation. Multipoint methods with the convergence order 2^{n-1} , where n is the number of the function evaluations at each iteration step, are said to be optimal [10]. Another important characteristic of iterative methods is their efficiency index $EI = \rho^{1/n}$, where ρ is the convergence order. By way of example, consider the well-known family of Laguerre iterations (or Hansen–Patrick iterations) (2.1), which has cubic convergence; here the parameter τ_n is defined by

$$\tau_n = \frac{\alpha + 1}{\alpha + \operatorname{sgn}(\alpha) \sqrt{1 - (\alpha + 1) \frac{f''(x_n)f(x_n)}{f'(x_n)^2}}}, \quad \alpha \neq -1. \quad (2.4)$$

Using the expansion of the function $f(y_n)$ about the point x_n , it is easy to show that

$$\theta_n = \frac{f''(x_n)f(x_n)}{2f'(x_n)^2} + O(f(x_n)^2). \quad (2.5)$$

Then (2.4) yields

$$\tau_n = \frac{\alpha + 1}{\alpha \pm \sqrt{1 - 2(\alpha + 1)\theta_n + O(f(x_n)^2)}}, \quad \alpha \neq -1. \quad (2.6)$$

Neglect the small term $O(f(x_n)^2)$ in (2.6) to obtain

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}}, \quad \alpha \neq -1. \quad (2.7)$$

Kansal et al. in [2] considered iterations (2.1) with the parameters defined by (2.7). Using the known relation

$$(1 - x)^\alpha = 1 - \alpha x + \frac{\alpha(\alpha - 1)}{2} x^2 - \frac{\alpha(\alpha - 1)(\alpha - 2)}{6} x^3 + \dots, \quad |x| < 1, \quad (2.8)$$

it easy to verify that (2.7) has the asymptotics

$$\tau_n = 1 + \theta_n + \frac{\alpha + 3}{2} \theta_n^2 + O(\theta_n^3). \quad (2.9)$$

The comparison of (2.9) with (2.2) shows that iterations (2.1) with the parameter τ_n defined by (2.7) are not optimal. Indeed, three evaluations of the function $f(x_n)$, $f(y_n)$ and $f'(x_n)$ are required at each iteration step in this case, and the convergence order is three. The only exception is $\alpha = 1$, i.e.,

$$\tau_n = \frac{2}{1 + \sqrt{1 - 4\theta_n}}, \quad (2.10)$$

which satisfies condition (2.2). Note that using the accelerating procedure for τ_n proposed in [4] fourth-order iterations (2.1) with τ_n defined by (2.10) were obtained. For this reason, the value defined by (2.10) is called optimal. We also note that an attempt to find the optimal parameter α of the Laguerre family from the convergence viewpoint was made in [6]. As a rule, iterations (2.1) with the parameter τ_n defined by (2.7) have only the third convergence order. Using condition (2.2), one can find an optimal modification of the Hansen–Patrick family of order four. To this end, we seek τ_n in the form

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} H(\theta_n), \quad \alpha \neq -1, \tag{2.11}$$

where H is a real function satisfying the conditions

$$H(0) = 1, \quad H'(0) = a, \quad H''(0) = 2b. \tag{2.12}$$

Let us find a and b in (2.12) such that (2.11) satisfies condition (2.2). Using the Taylor expansion of the function $H(\theta_n)$ and (2.9), we obtain in (2.11)

$$\tau_n = 1 + (a + 1)\theta_n + \left(a + b + \frac{\alpha + 3}{2}\right)\theta_n^2 + \dots \tag{2.13}$$

The comparison of (2.13) with (2.2) gives

$$a = 0, \quad b = \frac{1 - \alpha}{2}.$$

Thus, we obtain an optimal version of the Hansen–Patrick family (2.1) with the parameter defined by

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} \left(1 + \frac{1 - \alpha}{2}\theta_n^2\right), \quad \alpha \neq -1. \tag{2.14}$$

If $\alpha = 1$, then (2.14) yields (2.10). Thus, we show that one can pass from any third-order iterations (2.1) to the optimal two-point iterations using condition (2.2). Similarly, it is easy to show that the Hansen–Patrick iterations have the optimal convergence order four if

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} + \frac{1 - \alpha}{2}\theta_n^2. \tag{2.15}$$

Note that the authors of [1] proposed a new optimal modification of the Hansen–Patrick family (2.1) of order four; the parameter for this modification is defined by the formula

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{\frac{1 - (\alpha + 3)\theta_n - (\alpha^2 - 1)\theta_n^2}{1 + (\alpha - 1)\theta_n}}}, \quad \alpha \neq -1. \tag{2.16}$$

Even though iterations (2.1) are optimal with the efficiency index $EI = 4^{1/3} \approx 1.587$, they require the first-order derivative to be evaluated at each iteration step; therefore, they cannot be applied to equations with nonsmooth functions. In [5], a rule for transforming iterations (2.1) into their derivative-free optimal version and conversely was proposed. According to this rule, it is easy to obtain a derivative-free version of (2.1) using (2.16). It has the form

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{\phi_n}, \\ x_{n+1} &= x_n - \tau_n \frac{f(x_n)}{\phi_n} \quad \text{or} \quad \left(x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n} \right), \end{aligned} \tag{2.17}$$

where

$$w_n = x_n + \mathcal{V}f(x_n), \quad \phi_n = f[x_n, w_n] = \frac{f(w_n) - f(x_n)}{w_n - x_n}$$

and

$$\tau_n = 1 + \bar{\tau}_n \theta_n, \tag{2.18}$$

$$\bar{\tau}_n = \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{\frac{1 - (\alpha + 3)\theta_n - (\alpha^2 - 1)\theta_n^2}{1 + (\alpha - 1)\theta_n}}} - 1 \right) + (\hat{d}_n - 2)\theta_n, \quad (2.19)$$

$$\hat{d}_n = \frac{2 + \gamma\phi_n}{1 + \gamma\phi_n}. \quad (2.20)$$

Similarly, using the sufficient condition for the fourth-order convergence (see [7])

$$\bar{\tau}_n = 1 + \hat{d}_n\theta_n + O(f(x_n)^2), \quad (2.21)$$

for (2.17), a derivative-free version of (2.1), (2.14) can be easily constructed. It can be written as (2.17) with the parameter

$$\bar{\tau}_n = \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} \left(1 + \left(\hat{d}_n - 2 - \frac{\alpha - 1}{2} \right) \theta_n^2 \right) - 1 \right), \quad \alpha \neq -1. \quad (2.22)$$

Thus, we have derivative-free families of the Hansen–Patrick iterations (2.17) with the parameters of two forms (2.19) and (2.22).

Remark 1. Generally, we can consider the weighting function

$$W(\theta_n, \alpha, m) = \frac{\alpha + 1}{\alpha + \sqrt[m]{1 - m(\alpha + 1)\theta_n}} = 1 + \theta_n + \left(1 - \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2 + \dots \quad (2.23)$$

in iteration (2.1). $W(\theta_n, \alpha, m)$ is called the generalized Hansen–Patrick type weighting function. The function $W(\theta_n, \alpha, 2)$ leads to (2.7). It is easy to show that the iterative methods (2.1) have the optimal fourth convergence order if τ_n satisfies one of the following conditions:

$$\tau_n = W(\theta_n, \alpha, m) + \left(1 + \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2, \quad \alpha \neq -1, \quad (2.24)$$

and

$$\tau_n = W(\theta_n, \alpha, m)H(\theta_n), \quad (2.25)$$

where H is a real function satisfying the conditions

$$H(0) = 1, \quad H'(0) = 0, \quad H''(0) = 2 \left(1 + \frac{1 - m}{2}(\alpha + 1) \right). \quad (2.26)$$

For example, the following functions can be used as H :

$$\begin{aligned} H_1 &= 1 + \left(1 + \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2, \\ H_2 &= \frac{1}{1 - \left(1 + \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2}, \\ H_3 &= \sqrt{1 + (2 + (1 - m)(\alpha + 1))\theta_n^2}. \end{aligned}$$

3. THE FAMILY OF DERIVATIVE-FREE TWO-PARAMETER ITERATIONS

Consider the derivative-free two-parameter iterations

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{\phi_n + \lambda f(w_n)}, \quad \lambda \in \mathbb{R}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda f(w_n)}, \end{aligned} \quad (3.1)$$

where $w_n = x_n + \gamma f(x_n)$, $\gamma \in R$, and $\phi_n = f[x_n, w_n] = \frac{f(w_n) - f(x_n)}{\gamma f_n}$. We want to find $\bar{\tau}_n$ in (3.1) such that iterations (3.1) have the optimal fourth convergence order. To this end, we first use the Taylor expansion of the function $f(w_n) = f(x_n)(1 + \gamma\phi_n)$ about the point x_n . Then, we obtain

$$\phi_n = f'(x_n) \left(1 + \frac{a_n}{2} f'(x_n) \gamma \right) + O(f_n^2), \quad (f_n) = f(x_n), \quad (3.2)$$

where

$$a_n = \frac{f''(x_n)f(x_n)}{f'(x_n)^2}. \quad (3.3)$$

Let $\eta_n = \frac{f'(x_n)}{\phi_n}$. Then, using (3.2), we obtain

$$\eta_n = \frac{1}{1 + \frac{a_n}{2} f'(x_n) \gamma + O(f_n^2)} = 1 - \frac{a_n}{2} f'(x_n) \gamma + O(f_n^2). \quad (3.4)$$

The Taylor expansion of $f(y_n)$ about the point x_n yields

$$f(y_n) = f(x_n) \left(1 - \eta_n \left(1 - \frac{\lambda f(w_n)}{\phi_n} \right) \right) + O(f_n^2) = f(x_n) \left(1 - \left(1 - \frac{a_n}{2} f'(x_n) \gamma \right) \left(1 - \frac{\lambda f(w_n)}{\phi_n} \right) \right) + O(f_n^2). \quad (3.5)$$

According to (3.3), we have $f(y_n) = O(f(x_n)^2)$. Similarly, the second step in (3.1) gives

$$f(x_{n+1}) = \left(1 - \bar{\tau}_n \frac{f'(y_n)}{\phi_n + \lambda f(w_n)} \right) f(y_n) + O(f(y_n)^2). \quad (3.6)$$

From (3.5) and (3.6), we obtain

$$f(x_{n+1}) = O(f_n^4) \quad (3.7)$$

if $\bar{\tau}_n$ is chosen such that

$$1 - \bar{\tau}_n \frac{f'(y_n)}{\phi_n + \lambda f(w_n)} = O(f_n^2)$$

or

$$\bar{\tau}_n = \frac{\phi_n + \lambda f(w_n)}{f'(y_n)} + O(f_n^2). \quad (3.8)$$

The Taylor expansion of $f'(y_n)$ about the point x_n gives

$$f'(y_n) = f'(x_n) \left(1 - \frac{f_n'' f_n}{f_n' \phi_n \left(1 + \frac{\lambda f(w_n)}{\phi_n} \right)} \right) + O(f_n^2), \quad f_n' = f'(x_n).$$

Using (3.3) and (3.4) in the last relation, we obtain

$$f'(y_n) = f_n'(1 - a_n) + O(f_n^2). \quad (3.9)$$

Substitute (3.2) and (3.9) in (3.8) to obtain

$$\begin{aligned} \bar{\tau}_n &= \frac{1 + \frac{a_n}{2} f_n' \gamma + \lambda \frac{(1 + \gamma \phi_n) f_n}{f_n'} + O(f_n^2)}{1 - a_n + O(f_n^2)} = \left(1 + \frac{a_n}{2} f_n' \gamma + \frac{\lambda(1 + \gamma \phi_n) f_n}{f_n'} \right) (1 + a_n) + O(f_n^2) \\ &= 1 + \left(1 + \frac{f_n' \gamma}{2} \right) a_n + \frac{\lambda(1 + \gamma \phi_n) f_n}{f_n'} + O(f_n^2). \end{aligned} \quad (3.10)$$

According to (3.3) and (3.4), we have $f'_n = \phi_n + O(f_n)$. Therefore, we may replace f'_n by ϕ_n in (3.10) without loss of accuracy. As a result, we have

$$\bar{\tau}_n = 1 + \frac{2 + \gamma\phi_n}{2} a_n + \frac{\lambda(1 + \gamma\phi_n)f_n}{\phi_n} + O(f_n^2). \quad (3.11)$$

Next, using the Taylor expansion of $f(y_n)$ and (3.4), we readily obtain

$$\begin{aligned} \theta_n &= \frac{a_n}{2} (1 + \gamma f'_n) + \frac{\lambda(1 + \gamma\phi_n)}{\phi_n} f(x_n) + O(f_n^2) = \frac{a_n}{2} (1 + \gamma\phi_n) + \frac{\lambda(1 + \gamma\phi_n)}{\phi_n} f(x_n) + O(f_n^2) \\ &= (1 + \gamma\phi_n) \left(\frac{a_n}{2} + \lambda \frac{f(x_n)}{\phi_n} \right) + O(f_n^2). \end{aligned} \quad (3.12)$$

Hence, we find

$$\frac{a_n}{2} = \frac{\theta_n}{1 + \gamma\phi_n} - \frac{\lambda f(x_n)}{\phi_n} + O(f_n^2). \quad (3.13)$$

Substitute (3.13) into (3.11) to obtain

$$\bar{\tau}_n = 1 + \hat{d}_n \theta_n - \frac{\lambda f(x_n)}{\phi_n} + O(f_n^2). \quad (3.14)$$

Thus, we may formulate the results in the form of the following theorem.

Theorem 1. *Assume that the function $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ is sufficiently smooth and has a simple zero $x^* \in D$. Suppose that the initial approximation x_0 is sufficiently close to x^* and the parameter $\bar{\tau}_n$ satisfies condition (3.14). Then, the iterative methods (3.1) have the optimal fourth order of convergence.*

Kansal et al. proposed in [2] a new derivative-free three-parameter optimal family of Hansen–Patrick iterations

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{\phi_n + \lambda f(w_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{f[y, w_n] + \lambda f(w_n)}, \end{aligned} \quad (3.15)$$

where

$$\bar{\tau}_n = \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} - 1 \right) H(\theta_n), \quad \alpha \neq -1. \quad (3.16)$$

Here, H is a real weight function satisfying the condition

$$H(0) = 1, \quad H'(0) = -\frac{\alpha + 1}{2}, \quad |H''(0)| < \infty. \quad (3.17)$$

Note that iterations (3.15) have a difference in the denominator at the second stage compared with (3.1). Using the easily verified relation

$$f[y, w_n] + \lambda f(w_n) = (\phi_n + \lambda f(w_n)) \left(1 - \frac{\phi_n \theta_n - \lambda f(w_n)}{(1 + \gamma\phi_n)(\phi_n + \lambda f(w_n))} \right) + O(f_n^2), \quad (3.18)$$

the second step in (3.15) can be written as

$$x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda f(w_n)},$$

where

$$\bar{\tau}_n = \left(1 + \frac{\phi_n \theta_n - \lambda f(w_n)}{(1 + \gamma\phi_n)(\phi_n + \lambda f(w_n))} \right) \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} - 1 \right) H(\theta_n), \quad \alpha \neq -1. \quad (3.19)$$

It is easy to show that $\bar{\tau}_n$ defined by (3.19) satisfies condition (3.14). That is, we prove that iterations (3.15)–(3.17) have the fourth-order convergence without using symbolic computation, which were employed in [2]. The two-parameter iteration (3.1) with $\bar{\tau}_n$ defined by (3.19) is a new derivative-free ver-

sion of the family of Hansen–Patrick iterations. Similarly, using formula (3.18), it is easy to show that the derivative-free two-parameter fourth-order methods described in [8, 10, 13] satisfy condition (3.14).

Let $\gamma = 0$ in (3.1). Then, (3.1) give the one-parameter iterations

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n) + \lambda f(x_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda f(x_n)}. \end{aligned} \quad (3.20)$$

By Theorem 1, iterations (3.20) have the optimal fourth-order convergence if $\bar{\tau}_n$ is defined by

$$\bar{\tau}_n = 1 + 2\theta_n - \frac{\lambda f(x_n)}{f'(x_n)} + O(f_n^2). \quad (3.21)$$

Iterations (3.20) require three calculations of the function $f(x_n)$, $f(y_n)$, and $f'(x_n)$. The efficiency index of these iterations is $EI = \sqrt[3]{4} \approx 1.587$. Now, let us try to find the optimal value of the free parameter λ . To this end, we first use the Taylor expansion of $f(y_n)$ about the point x_n and relation (2.8). As a result, we obtain

$$f(y_n) = f(x_n)^2 \left(\frac{\lambda}{f'_n} + \frac{f''_n}{2f_n'^2} - \frac{\lambda^2 f_n}{f_n'^2} - \frac{\lambda f_n f''_n}{f_n'^3} - \frac{f_n''' f_n}{6f_n'^3} \right) + O(f_n^4). \quad (3.22)$$

This implies that $f(y_n) = O(f_n^2)$ for every λ . If we choose

$$\lambda = \lambda_n = -\frac{f''_n}{2f'_n} \quad (3.23)$$

in (3.22), then we obtain $f(y_n) = O(f_n^3)$. The value λ_n defined by (3.23) will be called optimal in the sense that it increases the convergence order of the sequence y_n from two to three. If the parameter is defined by (3.23), expression (3.22) can be written as

$$f(y_n) = f(x_n) \left(\left(\frac{a_n}{2} \right)^2 - \frac{f_n''' f_n^2}{6f_n'^3} \right) + O(f_n^4). \quad (3.24)$$

Therefore,

$$\theta_n = \left(\frac{a_n}{2} \right)^2 - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3),$$

which means that

$$\frac{f_n''' f_n^2}{f_n'^3} = \frac{3}{2} a_n^2 - 6\theta_n + O(f_n^3). \quad (3.25)$$

Now, consider the Taylor expansion of $f(x_{n+1})$ about the point y_n :

$$f(x_{n+1}) = f(y_n) \left(1 - \frac{f'(y_n)}{f'(x_n)} \bar{\tau}_n \left(1 - \frac{\lambda f_n}{f'_n} \right) \right) + O(f(y_n)^2). \quad (3.26)$$

It was shown above that, if the parameter is defined by (3.23), then $f(y_n) = O(f_n^3)$. Therefore, (3.26) implies that

$$f(x_{n+1}) = O(f(x_n)^6) \quad (3.27)$$

if $\bar{\tau}_n$ is such that

$$1 - \frac{f'(y_n)}{f'(x_n)} \bar{\tau}_n \left(1 - \frac{\lambda f_n}{f'_n} \right) = O(f_n^3)$$

or

$$\bar{\tau}_n = \frac{1}{1 - \frac{\lambda f_n f'(y_n)}{f'_n}} \frac{f'(x_n)}{f'_n} + O(f_n^3). \quad (3.28)$$

Using the Taylor expansion of $f'(y_n)$ about the point x_n , it is easy to show that

$$f'(y_n) = f'(x_n) \left(1 - a_n - \frac{a_n^2}{2} + \frac{f_n''' f_n^2}{2 f_n'^3} \right) + O(f_n^3).$$

Taking into account (3.25), we obtain

$$\frac{f'(x_n)}{f'(y_n)} = \frac{1}{1 - a_n + a_n^2/4 - 3\theta_n + O(f_n^3)} = 1 + a_n + \frac{3a_n^2}{4} + 3\theta_n + O(f_n^3). \quad (3.29)$$

Substitute (3.23) and (3.29) into (3.28) to obtain

$$\bar{\tau}_n = 1 + \frac{a_n}{2} + \frac{a_n^2}{4} + 3\theta_n + O(f_n^3). \quad (3.30)$$

This implies that (3.27) is satisfied if the parameters are defined by (3.30) and (3.23). Therefore, we have the following result.

Theorem 2. *Assume that the function $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ is sufficiently smooth and has a simple zero $x^* \in D$. Suppose that the initial approximation x_0 is sufficiently close to x^* and the parameters λ_n and $\bar{\tau}_n$ satisfy conditions (3.23) and (3.30). Then, the iterative methods (3.20) have the sixth order of convergence.*

On the basis of the optimal choice of the parameters λ_n and $\bar{\tau}_n$, we can construct converging iterations of order six with memory:

x_0, λ_0 are given. Then

$$\begin{aligned} \lambda_n &= -\frac{\Delta_n}{2f'_n}, \quad n = 1, 2, \dots, \\ \bar{\tau}_n &= 1 - \frac{\lambda_n f_n}{f'_n} + \left(\frac{\lambda_n f_n}{f'_n} \right)^2 + 3\theta_n, \\ y_n &= x_n - \frac{f(x_n)}{f'(x_n) + \lambda_n f(x_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda_n f(x_n)}, \quad n = 0, 1, \dots, \end{aligned} \quad (3.31)$$

where

$$\Delta_n = \frac{f(x_n + \gamma f(x_n)) - 2f(x_n) + f(x_n - \gamma f(x_n))}{(\gamma f(x_n))^2}, \quad \gamma \in \mathbb{R} \setminus \{0\}.$$

It is clear that $f''(x_n) = \Delta_n + O(f_n^2)$.

Remark 2. The equation of error derived using symbolic computation plays an important role in creating new derivative-free methods with memory [1, 2, 8–13]. For example, if

$$\lambda = -c_2 = -\frac{f''(x^*)}{2f'(x^*)}, \quad \lim_{n \rightarrow \infty} \lambda_n = \lambda,$$

then the convergence order of the methods increases, and at each iteration step we have the exact analytical formula (3.23).

Similarly, it is easy to show that

$$f(x_{n+1}) = O(f(x_n)^5) \quad \text{if} \quad \bar{\tau}_n = 1 + \frac{a_n}{2} + O(f_n^2). \quad (3.32)$$

Note that methods similar to (3.20) were studied in [8], where

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n) + \lambda f(x_n)}, \\ x_{n+1} &= y_n - \frac{f(y_n)}{f'(x_n) + \gamma f(x_n)} G(\theta_n), \quad \lambda, \gamma \in R \end{aligned} \quad (3.33)$$

is considered and it is shown that (3.33) has the fourth convergence order if

$$\gamma = 2\lambda, \quad G(0) = 1, \quad G'(0) = 2, \quad |G''(0)| < \infty. \quad (3.34)$$

In [8], the self-accelerating parameter

$$\lambda_n = -\frac{H_m''(x_n)}{2H_m'(x_n)}, \quad m = 2, 3, 4, \quad (3.35)$$

was used in (3.33), and it was proved that the convergence order of the iterative methods (3.33) with the parameter (3.35) and memory is not lower than $(5 + \sqrt{17})/2 \approx 4.5616$, $(5 + \sqrt{21})/2 \approx 4.7913$ and 5, respectively. Here $H_m(x_n)$ is the Hermite interpolation polynomial of degree $m = 2, 3, 4$ satisfying the condition $H_m'(x_n) = f'(x_n)$. Iterations (3.33) and (3.34) can be written as (3.20) with $\bar{\tau}_n$ defined by

$$\bar{\tau}_n = \left(1 - \frac{\lambda f_n}{f'_n + \lambda f_n} + \dots \right) (1 + 2\theta_n + \dots) = 1 + \frac{a_n}{2} + O(f_n^2);$$

i.e., $\bar{\tau}_n$ satisfies condition (3.32). If we choose λ_n as in (3.31), then we obtain the following iterations with memory:

$$\begin{aligned} x_0, \lambda_0 \quad &\text{are given.} \quad \text{Then} \\ \lambda_n &= -\frac{\Delta_n}{2f'_n}, \quad a_n = -\frac{2\lambda_n f_n}{f'_n}, \quad \bar{\tau}_n = 1 + \frac{a_n}{2}, \\ y_n &= x_n - \frac{f(x_n)}{f'(x_n) + \lambda_n f(x_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda_n f(x_n)}, \end{aligned} \quad (3.36)$$

which have the fifth convergence order.

Remark 3. It has already been mentioned above that $\bar{\tau}_n$ for the Hansen–Patrick iterations is defined by (3.19).

4. NEW ITERATIVE METHODS WITH MEMORY

Now, we construct new iterative methods with memory on the basis of (3.1) using two self-accelerating parameters γ and λ . It is easy to verify that

$$w_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad (4.1)$$

and

$$f(w_n) = \frac{f_n'' f_n^2}{2f_n'^2} + O(f_n^3), \quad (4.2)$$

with the choice

$$\gamma = \gamma_n = -\frac{1}{f_n'} \quad (4.3)$$

Let $f(x_n) \in C^4(I)$. Using the Taylor expansion for $f(w_n)$ and (4.3), we obtain

$$\phi_n = f_n' \left(1 - \frac{a_n}{2} + \frac{f_n''' f_n^2}{6 f_n'^3} \right) + O(f_n'^3).$$

Therefore,

$$\eta_n = \frac{f_n'}{\phi_n} = 1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6 f_n'^3} + O(f_n'^3). \quad (4.4)$$

The Taylor expansion of $f(y_n)$ about the point x_n gives

$$f(y_n) = f(x_n) \left(1 - \frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} + \frac{a_n}{2} \left(\frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} \right)^2 - \frac{f_n''' f_n^2}{6 f_n'^3} \left(\frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} \right)^3 \right) + O(f_n'^4). \quad (4.5)$$

Due to (4.2) and (4.4), we have

$$\begin{aligned} \frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} &= \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6 f_n'^3} + \dots \right) \left(1 - \frac{\lambda f(w_n)}{\phi_n} + \dots \right) \\ &= \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6 f_n'^3} - \frac{\lambda f(w_n)}{\phi_n} + O(f_n'^3) \right). \end{aligned} \quad (4.6)$$

Using (4.6) in (4.5), we obtain

$$\begin{aligned} f(y_n) &= f(x_n) \left(-\frac{a_n}{2} - \frac{a_n^2}{4} + \frac{f_n''' f_n^2}{6 f_n'^3} + \frac{\lambda f(w_n)}{\phi_n} + \frac{a_n}{2} (1 + a_n) - \frac{f_n''' f_n^2}{6 f_n'^3} \right) + O(f_n'^4) \\ &= f(x_n) \left(\frac{a_n^2}{4} + \frac{\lambda f(w_n)}{\phi_n} \right) + O(f_n'^4). \end{aligned} \quad (4.7)$$

It is clear from (4.7) that

$$f(y_n) = O(f_n'^4) \quad (4.8)$$

if

$$\frac{a_n^2}{4} + \frac{\lambda f(w_n)}{\phi_n} = 0, \quad (4.9)$$

or

$$\lambda_n = -\frac{a_n^2 \phi_n}{4 f(w_n)}. \quad (4.10)$$

Using (4.2) and (4.4) in (4.10), we obtain

$$\lambda_n = -\frac{f_n'''}{2 f_n'}; \quad (4.11)$$

i.e., (4.8) holds with the choice (4.11). Furthermore, (3.6) and (4.8) imply that

$$f(x_{n+1}) = O(f_n'^7) \quad (4.12)$$

if

$$\bar{\tau}_n = -\frac{\phi_n + \lambda f(w_n)}{f'(y_n)} + O(f_n^3) \tag{4.13}$$

or

$$\bar{\tau}_n = \frac{\phi_n}{f'(y_n)} \left(1 - \frac{a_n^2}{4}\right) + O(f_n^3). \tag{4.14}$$

The Taylor expansion of $f'(y_n)$ about the point x_n gives

$$f'(y_n) = f'(x_n) \left(1 - a_n \left(\frac{\eta_n}{1 - \frac{a_n^2}{4}}\right) + \frac{f_n''' f_n^2}{2f_n'^3} \left(\frac{\eta_n}{1 - \frac{a_n^2}{4}}\right)^2\right) + O(f_n^3). \tag{4.15}$$

Since

$$\frac{\eta_n}{1 - \frac{a_n^2}{4}} = \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6f_n'^3} + \dots\right) \left(1 + \frac{a_n^2}{4} + \dots\right) = 1 + \frac{a_n}{2} + \frac{a_n^2}{2} - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3), \tag{4.16}$$

we conclude from (4.15) that

$$f'(y_n) = f'(x_n) \left(1 - a_n - \frac{a_n^2}{2} + \frac{f_n''' f_n^2}{2f_n'^3}\right) + O(f_n^3). \tag{4.17}$$

Using the last expression and (4.4) in (4.14), we obtain

$$\bar{\tau}_n = 1 - \frac{a_n}{2} + \frac{3}{4}a_n^2 + 2(1 + \gamma_n \phi_n) + O(f_n^3), \tag{4.18}$$

where the formula

$$1 + \gamma_n \phi_n = \frac{a_n}{2} - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3) \tag{4.19}$$

is used. Thus, the results obtained above can be formulated as the following theorem.

Theorem 3. *Assume that the function $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ is sufficiently smooth and has a simple zero $x^* \in D$. Suppose that the initial approximation x_0 is sufficiently close to x^* and the parameters γ and λ in (3.1) are chosen by*

$$\gamma = \gamma_n = -\frac{1}{f'(x_n)}, \quad \lambda = \lambda_n = -\frac{f''(x_n)}{2f'(x_n)}$$

and $\bar{\tau}_n$ is defined by formula (3.14) (or (4.18)). Then, the iterative methods (3.1) have the seventh order of convergence.

Thus, the optimal choice of parameters makes it possible to increase the convergence order from four to seven. However, $f'(x_n)$ and $f''(x_n)$ cannot be calculated in practice, and such an improvement of convergence cannot be implemented. However, we can find approximations of γ_n and λ_n . They can be found using the information available from the current and previous iteration steps. On the basis of versions (4.3) and (4.11), two-point derivative-free iterations with memory with the seventh convergence order can be constructed:

$$x_0, \lambda_0, \gamma_0 \quad \text{are given.} \quad \text{Then} \quad w_0 = x_0 + \gamma_0 f(x_0),$$

Table 1. $f_1 = e^{x^3-x} - \cos(x^2 - 1) + x^3 + 1, x_0 = -1.5, x^* = -1$ [14]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.1) ($\lambda = -0.1, \gamma = -0.01$)	4	(3.14)	0.1014e-217	4.00
(3.1) ($\alpha = 1, \gamma = -0.01$)	4	(3.19)	0.1544e-224	4.00
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.6919e-229	4.00
Dzunic [13] ($p = -0.1, \gamma = -0.01, g(\theta_n) = 1 + \theta_n$)	4		0.4682e-222	4.00
Wang-Zhang [8] ($t = 8, \lambda = -0.1, G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	4		0.1974e-192	4.00
Kung-Traub [11]	4		0.9297e-173	4.00
Chebyshev-Halley [11]	4		0.5980e-175	4.00

Table 2. $f_1 = e^{x^3-x} - \cos(x^2 - 1) + x^3 + 1, x_0 = -1.5, x^* = -1$ [14]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.20) ($\lambda_n = -f_n''/2f_n'$)	3	(3.21)	0.1735e-56	5.00
(3.20) ($\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1$)	3	(3.32)	0.7578e-99	5.00
(3.36) ($\lambda_n = -\Delta_n/2f_n', \lambda_0 = -0.1$)	3		0.4079e-85	5.02
(3.33)-(3.35) [8] ($\lambda_n = -H_4''/2f_n', \lambda_0 = -0.1$)	3		0.2404e-89	5.09
(3.20) ($\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1$)	3	(3.30)	0.6559e-176	6.00
(3.31) ($\lambda_n = -\Delta_n/2f_n', \lambda_0 = -0.1$)	3		0.4538e-125	6.00
(3.1) ($\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1$)	3	(4.18)	0.3128e-93	7.00
(4.20) ($\lambda_n = -N_4''(x_n)/2N_4'(x_n), \lambda_0 = -0.1, \gamma_0 = -0.01$)	3	(3.21)	0.4294e-162	7.06
Dzunic [13] ($p_0 = -0.1, \gamma_0 = -0.01, g(\theta_n) = 1 + \theta_n$)	3		0.1404e-157	7.06
Cordero [14] ($\lambda_0 = -0.1, \gamma_0 = -0.01$)	3		0.1114e-157	7.06
Kansal [2] ($\lambda_0 = -0.1, \gamma_0 = -0.01, \alpha = \beta = 1/2$)	3		0.1095e-100	7.08

Table 3. $f_2 = e^{x^3-3x} \sin x + \log(x^2 + 1), x_0 = 1, x^* = 0$ [12]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.1) ($\lambda = -0.1, \gamma = -0.01$)	4	(3.14)	0.1469e-82	4.00
(3.1) ($\alpha = 1, \gamma = -0.01$)	4	(3.19)	0.6589e-68	3.99
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.3650e-83	4.00
Dzunic [13] ($p = -0.1, \gamma = -0.01, g(\theta_n) = 1 + \theta_n$)	4		0.7008e-66	4.00
Wang-Zhang [8] ($t = 8, \lambda = -0.1, G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	5		0.7447e-204	4.00
Kung-Traub [11]	4		0.1469e-82	4.00
Chebyshev-Halley [11]	4		0.1975e-88	4.00

$$\gamma_n = -\frac{1}{N_3'(x_n)}, \quad w_n = x_n + \gamma_n f(x_n), \quad \lambda_n = -\frac{N_4''(x_n)}{2N_4'(x_n)}, \quad n = 1, 2, \dots, \tag{4.20}$$

$$y_n = x_n - \frac{f(x_n)}{\phi_n + \lambda_n f(w_n)},$$

$$x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda_n f(w_n)}, \quad n = 0, 1, \dots,$$

where $\bar{\tau}_n$ satisfies condition (3.14). Here $N_3(t, x_n, y_{n-1}, x_{n-1}, w_{n-1})$ and $N_4(t, w_n, x_n, w_{n-1}, y_{n-1}, x_{n-1})$ are the Newton interpolation polynomials of degrees three and four passing through the node points

Table 4. $f_2 = e^{x^3-3x} \sin x + \log(x^2 + 1)$, $x_0 = 1$, $x^* = 0$ [12]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.20) ($\lambda_n = -f_n''/2f_n'$)	4	(3.21)	0.2170e-217	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.32)	0.3916e-220	5.00
(3.36) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	4		0.2326e-136	5.00
(3.33)–(3.35) [8] ($\lambda_n = -H_4''/2f_n'$, $\lambda_0 = -0.1$)	4		0.2069e-122	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.30)	0.3111e-233	6.00
(3.31) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	4		0.2699e-291	6.00
(3.1) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(4.18)	0.1560e-119	7.00
(4.20) ($\lambda_n = -N_4''(x_n)/2N_4'(x_n)$, $\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	4	(3.21)	0.3134e-416	7.00
Dzunic [13] ($p_0 = -0.1$, $\gamma_0 = -0.01$), $g(\theta_n) = 1 + \theta_n$	4		0.3892e-330	6.99
Cordero [14] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	4		0.5524e-284	6.99
Kansal [2] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$), $\alpha = \beta = 1/2$	4		0.8391e-293	6.98

Table 5. $f_3 = (x^6 + x^{-6} + 4)(x - 1)\sin x^2$, $x_0 = 0.8$, $x^* = 1$ [13]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.1) ($\lambda = -0.1$, $\gamma = -0.01$)	4	(3.14)	0.3589e-140	4.00
(3.1) ($\alpha = 1$, $\gamma = -0.01$)	4	(3.19)	0.9036e-111	4.00
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.1007e-138	4.00
Dzunic [13] ($p = -0.1$, $\gamma = -0.01$, $g(\theta_n) = 1 + \theta_n$)	4		0.4671e-130	4.00
Wang-Zhang [8] ($t = 8$, $\lambda = -0.1$, $G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	4		0.1552e-96	4.00
Kung-Traub [11]	4		0.2972e-132	4.00
Chebyshev–Halley [11]	4		0.2847e-118	4.00

$(x_n, x_{n-1}, y_{n-1}, w_{n-1})$ and $(x_n, w_n, x_{n-1}, y_{n-1}, w_{n-1})$, respectively. Note that procedure (4.20) was obtained in [13] with the choice

$$\lambda_n = -\frac{N_4''(w_n)}{2N_4'(w_n)}.$$

5. NUMERICAL EXPERIMENTS

To illustrate the behavior of convergence and the efficiency of methods (3.1), (3.20), (3.36), and (4.20), we consider a few numerical examples and make comparisons with some other existing methods of the same order. The computations were performed in Maple 18 using the multi-precision arithmetic with 1000 digits. In the numerical computations, we used the following functions [12–14]:

$$\begin{aligned} f_1 &= e^{x^3-x} - \cos(x^2 - 1) + x^3 + 1, & x^* &= -1, \\ f_2 &= e^{x^3-3x} \sin x + \log(x^2 + 1), & x^* &= 0, \\ f_3 &= (x^6 + x^{-6} + 4)(x - 1)\sin x^2, & x^* &= 1, \end{aligned}$$

and the stopping rule $|x_n - x^*| < 10^{-60}$. The computation results are presented in Tables 1–6, where the number of iterations (n), the absolute error $|x_n - x^*|$, and the computational order of convergence (ρ) defined by the formula

$$\rho \approx \frac{\ln(|x_n - x^*|/|x_{n-1} - x^*|)}{\ln(|x_{n-1} - x^*|/|x_{n-2} - x^*|)}$$

Table 6. $f_3 = (x^6 + x^{-6} + 4)(x - 1) \sin x^2$, $x_0 = 0.8$, $x^* = 1$ [13]

Methods	n	$\bar{\tau}_n$	$ x_n - x^* $	ρ
(3.20) $(\lambda_n = -f_n''/2f_n')$	3	(3.21)	0.1344e-53	4.99
(3.20) $(\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1)$	4	(3.32)	0.2239e-250	5.00
(3.36) $(\lambda_n = -\Delta_n/2f_n', \lambda_0 = -0.1)$	4		0.5113e-183	5.00
(3.33)–(3.35) [8] $(\lambda_n = -H_4''/2f_n', \lambda_0 = -0.1)$	4		0.1142e-213	5.00
(3.20) $(\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1)$	4	(3.30)	0.2116e-259	6.00
(3.31) $(\lambda_n = -\Delta_n/2f_n', \lambda_0 = -0.1)$	3		0.1080e-60	5.96
(3.1) $(\lambda_n = -f_n''/2f_n', \lambda_0 = -0.1)$	4	(4.18)	0.1802e-315	7.00
(4.20) $(\lambda_n = -N_4''(x_n)/2N_4'(x_n), \lambda_0 = -0.1, \gamma_0 = -0.01)$	3	(3.21)	0.6532e-107	7.03
Dzunic [13] $(p_0 = -0.1, \gamma_0 = -0.01), g(\theta_n) = 1 + \theta_n$	3		0.1364e-87	7.05
Cordero [14] $(\lambda_0 = -0.1, \gamma_0 = -0.01)$	3		0.8409e-80	7.04
Kansal [2] $(\lambda_0 = -0.1, \gamma_0 = -0.01), \alpha = \beta = 1/2$	3		0.9084e-72	7.02

are shown. It is seen from Tables 1–6 that the computation results completely confirm the theoretical order of convergence obtained in the preceding sections.

6. CONCLUSIONS

A new class of optimal derivative-free methods with two free parameters is proposed. Analytical formulas for the optimal values of these parameters are found, which makes it possible to improve the convergence order. On the basis of this fact, new iterative methods of a high convergence order with and without memory are proposed.

FUNDING

This work was supported by the Foundation of Science and Technology of Mongolia, project no. SST_18 /2018.

REFERENCES

1. M. Kansal, V. Kanwar, and S. Bhatia, “New modifications of Hansen–Patrick’s family with optimal fourth and eighth orders of convergence,” *Appl. Math. Comput.* **269**, 507–519 (2015).
2. M. Kansal, V. Kanwar, and S. Bhatia, “Efficient derivative-free variants of Hansen–Patrick’s family with memory for solving nonlinear equations,” *Numer. Algor.* **73**, 1017–1036 (2016).
3. T. Zhanlav, V. Ulziibayar, and O. Chuluunbaatar, “Necessary and sufficient conditions for the convergence of two- and three-point Newton-type iterations,” *Comput. Math. Math. Phys.* **57**, 1090–1100 (2017).
4. T. Zhanlav, O. Chuluunbaatar, and V. Ulziibayar, “Accelerating the convergence of Newton-type iterations,” *J. Numer. Anal. Approx. Theory* **46**, 162–180 (2017).
5. T. Zhanlav, R. Mijiddorj, and Kh. Otgondorj, “Constructive theory of designing optimal eighth-order derivative-free methods for solving nonlinear equations,” *Am. J. Comput. Math.* **10**, 100–117 (2020).
6. L. D. Petković, M. S. Petković, and B. Neta, “On optimal parameter of Laguerre’s family of zero-finding methods,” *Int. J. Comput. Math.* **95**, 692–707 (2018).
7. T. Zhanlav, O. Chuluunbaatar, and Kh. Otgondorj, “A derivative-free families of optimal two-and three-point iterative methods for solving nonlinear equations,” *Comput. Math. Math. Phys.* **50**, 920–936 (2019).
8. X. Wang and T. Zhang, “A new family of Newton-type iterative methods with and without memory for solving nonlinear equations,” *Calcolo* **51**, 1–15 (2014).
9. X. Wang, “A new accelerating technique applied to a variant of Cordero–Torregrosa method,” *J. Comput. Appl. Math.* **330**, 695–709 (2018).
10. M. S. Petković, S. Ilic, and J. Dzunić, “Derivative-free two-point methods with and without memory for solving nonlinear equations,” *Appl. Math. Comput.* **217**, 1887–1895 (2010).
11. I. K. Argyros, M. Kansal, V. Kanwar, and S. Bajaj, “Higher-order derivative-free families of Chebyshev–Halley type methods with or without memory for solving nonlinear equations,” *Appl. Math. Comput.* **315**, 224–245 (2017).
12. T. Lotfi, F. Soleymani, M. Ghorbanzadeh, and P. Assari, “On the construction of some tri-parametric iterative methods with memory,” *Numer. Algorithms* **70**, 835–845 (2015).
13. J. Dzunić, “On efficient two-parameter methods for solving nonlinear equations,” *Numer. Algor.* **63**, 549–569 (2013).
14. A. Cordero, T. Lotfi, P. Bakhtiari, and J. R. Torregrosa, “An efficient two-parametric family with memory for nonlinear equations,” *Numer. Algor.* **68**, 323–335 (2015).

Translated by A. Klimontovich

Constructive Theory of Designing Optimal Eighth-Order Derivative-Free Methods for Solving Nonlinear Equations

Tugal Zhanlav¹, Khuder Otgondorj², Renchin-Ochir Mijiddorj^{1,3}

¹Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia

²School of Applied Sciences, Mongolian University of Science and Technology, Ulaanbaatar, Mongolia

³Department of Informatics, Mongolian National University of Education, Ulaanbaatar, Mongolia

Email: tzhanlav@yahoo.com, otgondorj@gmail.com, mijiddorj@msue.edu.mn

How to cite this paper: Zhanlav, T., Otgondorj, Kh. and Mijiddorj, R.-O. (2020) Constructive Theory of Designing Optimal Eighth-Order Derivative-Free Methods for Solving Nonlinear Equations. *American Journal of Computational Mathematics*, 10, 100-117.

<https://doi.org/10.4236/ajcm.2020.101007>

Received: February 17, 2020

Accepted: March 14, 2020

Published: March 17, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This paper stresses the theoretical nature of constructing the optimal derivative-free iterations. We give necessary and sufficient conditions for derivative-free three-point iterations with the eighth-order of convergence. We also establish the connection of derivative-free and derivative presence three-point iterations. The use of the sufficient convergence conditions allows us to design wide class of optimal derivative-free iterations. The proposed family of iterations includes not only existing methods but also new methods with a higher order of convergence.

Keywords

Multipoint Methods, Derivative-Free Methods, Order of Convergence

1. Introduction

At present, there are a lot of iterative methods for solving nonlinear equations and systems of equations (see [1] [2] [3] and reference therein). In particular, the derivative-free methods are necessary when the derivative of the function f is unavailable or expensive to obtain. In the last decade, the derivative-free two and three-point methods with better convergence properties were developed (see [4]-[19] and references therein). It should be pointed out that most of these methods were proposed mainly for the concrete choice of parameters (see **Table 1**). Evidently, a systematic theory or an approach for constructing derivative-free methods is still needed. It is therefore of interest and necessity to develop a global theory. The aim of this paper is to fill up the above mentioned gap

Table 1. The derivative-free three-point iterative methods.

Methods	$\tilde{\tau}_n$	$H(\theta_n) = \frac{c + (\tilde{d}_n c + d)\theta_n + \omega\theta_n^2}{c + d\theta_n + b\theta_n^2}$
M_1, M_2, M_3 , Thukral [24]		
Kung-Traub's method	$\frac{1 + \gamma\phi_n}{(1 + \gamma\phi_n - \theta_n)(1 - \theta_n)}$	$c = 1, d = -\tilde{d}_n, b = \frac{1}{1 + \gamma\phi_n}, \omega = 0$
Thukral [7]		
$P1$ Thukral [24]	$1 + \tilde{d}_n\theta_n$	$c = 1, d = b = \omega = 0$
Soleymani, Khattri [5]		
$P2$ Thukral [24]	$\frac{1 + \theta_n}{1 - \frac{\theta_n}{1 + \gamma\phi_n}}$	$c = 1, d = -\frac{1}{1 + \gamma\phi_n}, b = \omega = 0,$
Sharma [14]		
$M2$ [24]	$\left(1 + \frac{\theta_n}{1 + \gamma\phi_n} + \frac{\theta_n^2}{(1 + \gamma\phi_n)^2}\right) \cdot \frac{1}{1 - \theta_n}$	$b = 0, c = 1, d = -1, \omega = \frac{1}{(1 + \gamma\phi_n)^2}$
method in [3]	$h(\theta_n, s_n) = 1 + \theta_n + s_n + \frac{a}{2}\theta_n^2 + b\theta_n s_n + \frac{c}{2}s_n^2$	$c = 1, d = b = 0, \omega = \left(\frac{a+c}{2} + \frac{a}{1 + \gamma\phi}\right)$
Soleymani [23]	$\frac{1}{1 - \tilde{d}_n\theta_n}$	$c = 1, d = -\tilde{d}_n, b = \omega = 0$
Zheng et al. [12]		
Soleymani [8]	$1 + \frac{\theta_n}{1 + \gamma\phi_n} + \frac{\theta_n^2}{(1 + \gamma\phi_n)^2}$	$c = 1, d = b = 0, \omega = \frac{1}{(1 + \gamma\phi_n)^2}$
Cordero et al. [17]	$\psi_4(x_n, y_n, \omega_n), (\gamma = 0)$	
Sharifi et al. [16]	$\frac{1 + \beta\theta_n}{1 + (\beta - 2)\theta_n} \frac{1}{1 - \frac{\theta_n}{1 + \gamma\phi_n}} G(\theta_n)$	$c = 1, b = \frac{2 - \beta}{1 + \gamma\phi_n}, \omega = -\beta, d = \beta - \tilde{d}_n - 1$
Chebyshev-Halley type method [4]	$\frac{1}{1 - 2\alpha\theta_n} \frac{1}{1 - \frac{\theta_n}{1 + \gamma\phi_n}} H(\theta_n)$ $H(0) = 1, H'(0) = 1 - 2\alpha$	$c = 1, d = -\left(2\alpha + \frac{1}{1 + \gamma\phi_n}\right), b = \frac{2\alpha}{1 + \gamma\phi_n}, \omega = H(\theta_n)$
Lotfi et al. [15]	$\frac{1 + \theta_n + a\tilde{d}_n \frac{\theta_n^2}{2}}{1 - \frac{\theta_n}{1 + \gamma\phi_n}}$	$c = 1, b = 0, d = -\frac{1}{1 + \gamma\phi_n}, \omega = \frac{a\tilde{d}_n}{2}$
Behl. [18]	$\psi_4(x_n, y_n, \omega_n)$	

and to obtain the wide class of optimal derivative-free three-point methods. The paper is organized as follows. In Section 2, we give the necessary and sufficient conditions for derivative-free three-point iterations to be optimal order eight. We also establish the connection between derivative presence and derivative-free three-point methods. In Section 3, we apply the sufficient convergence conditions to obtain the optimal derivative-free methods which are dependent on parameters in the third-step of considered iterations. We obtain families of optimal derivative-free three-point methods. They include many existing methods as

particular cases as well as new methods with the higher order of convergence. In last section, we present the results of numerical experiments that confirm the theoretical conclusion about the convergence order and make comparison with other known methods of the same order of convergence. Finally, numerical results show that new iterative methods can be significant by its high precision and practical use.

2. The Optimal Derivative-Free Three-Point Iterations

Typically, the optimal three-point iterative methods have a form [9]

$$y_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad z_n = y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n)}, \quad x_{n+1} = z_n - \alpha_n \frac{f(z_n)}{f'(x_n)}, \quad (1)$$

in which the parameters $\bar{\tau}_n$ and α_n are given by

$$\bar{\tau}_n = 1 + 2\theta_n + \tilde{\beta}\theta_n^2 + \tilde{\gamma}\theta_n^3 + \dots, \quad (2)$$

and

$$\alpha_n = 1 + 2\theta_n + (\tilde{\beta} + 1)\theta_n^2 + (2\tilde{\beta} + \tilde{\gamma} - 4)\theta_n^3 + (1 + 4\theta_n) \frac{f(z_n)}{f(y_n)} + O(f(x_n)^4), \quad (3)$$

where $\tilde{\beta}, \tilde{\gamma} \in R$, and $\theta_n = \frac{f(y_n)}{f(x_n)}$. In [9] was proven the following theorem.

Theorem 1. *Let the function $f(x)$ be sufficiently smooth and have a simple root $x^* \in I$. Furthermore, let the initial approximation x_0 be sufficiently close to x^* . Then, the convergence order of the iterative method (1) is eight if and only if the parameters $\bar{\tau}_n$ and α_n satisfy conditions (2) and (3), respectively.*

Remark. *The second sub-step in (1) can be rewritten as any two-point optimal fourth-order method*

$$z_n = \psi_4(x_n, y_n),$$

where $\psi_4(x_n, y_n)$ is a real function using the evaluation of $f(x_n), f'(x_n)$ and $f(y_n)$. Each method in ψ_4 has a parameter $\bar{\tau}_n$ given by (2) with own $\tilde{\beta}$ and $\tilde{\gamma}$.

Now we consider the derivative-free variant of (1)

$$\begin{aligned} y_n &= \psi_2(x_n, y_n) = x_n - \frac{f(x_n)}{\phi_n}, \\ z_n &= \psi_4(x_n, y_n) = y_n - \tilde{\tau}_n \frac{f(y_n)}{\phi_n}, \\ x_{n+1} &= z_n - \tilde{\alpha}_n \frac{f(z_n)}{\phi_n}, \end{aligned} \quad (4)$$

where

$$\begin{aligned} w_n &= x_n + \gamma f(x_n), \\ \phi_n &= \frac{f(w_n) - f(x_n)}{w_n - x_n} = \frac{1}{\gamma} \left(\frac{f(w_n)}{f(x_n)} - 1 \right) \approx f'(x_n), \gamma \in R. \end{aligned} \quad (5)$$

Here $\psi_2(x_n, y_n)$ is any second-order method. Actually, in Formula (4), the fundamental quantities are

$$\theta_n = \frac{f(y_n)}{f(x_n)}, \sigma_n = \frac{f(y_n)}{f(w_n)}, \text{ and } \nu_n = \frac{f(z_n)}{f(y_n)}.$$

Then $\theta_n = O(f(x_n))$, $\sigma_n = O(f(x_n))$ for $x_n \rightarrow x^*$, where x^* is a simple root of $f(x)$. If $\psi_4(x_n, y_n)$ is any two-point optimal fourth-order method then $f(z_n) = O(f(x_n)^4)$, therefore $\nu_n = O(f(x)^2)$. The iteration (4) obtained from (1) replacing $f'(x_n)$ by ϕ_n . Due to change (5), the parameters in (4) does not remain as before and we denote them by $\tilde{\tau}_n$ and $\tilde{\alpha}_n$. We call the iterations (1) and (4) the derivative presence (DP) and derivative-free (DF) variants respectively. If we use the notations

$$\tilde{c}_n = \frac{1}{1 + \gamma\phi_n}, \quad \tilde{d}_n = 1 + \tilde{c}_n,$$

then we have

$$\sigma_n = \tilde{c}_n\theta_n, \theta_n + \sigma_n = \tilde{d}_n\theta_n. \tag{6}$$

DP can be derived from DF by substituting $\sigma_n = \theta_n$. The following is the main result of our work [11].

Theorem 2. *Let the assumptions of Theorem 1 be fulfilled. Then, the convergence order of the iteration (4) is eight if and only if the parameters $\tilde{\tau}_n$ and $\tilde{\alpha}_n$ in (4) are given by formulas*

$$\tilde{\tau}_n = 1 + \tilde{d}_n\theta_n + \tilde{\beta}\theta_n^2 + \tilde{\gamma}\theta_n^3 + \dots, \tag{7}$$

and

$$\begin{aligned} \tilde{\alpha}_n = & 1 + \tilde{d}_n\theta_n + (\tilde{\beta} + \tilde{c}_n)\theta_n^2 + (\tilde{\gamma} + \tilde{d}_n(\tilde{\beta} - 1 - \tilde{c}_n^2))\theta_n^3 \\ & + (1 + 2\tilde{d}_n\theta_n)\nu_n + O(f(x_n)^4). \end{aligned} \tag{8}$$

The proposed method (4) with parameters given by (7) and (8) is three-point derivative free and optimal in the sense of Kung and Traub. Kung-Traub conjecture [20] states that the multi-point iterative methods, based on k evaluations, could achieve optimal convergence order 2^{k-1} . Our proposed method is in concurrence with the conjecture as it needs only four function evaluation per iteration *i.e.*, $k = 4$. Moreover, using ideas in [3] [10] one can propose more general construction for $\tilde{\tau}_n$ and $\tilde{\alpha}_n$ as following:

Define $\tilde{\tau}_n = h(\theta_n, \sigma_n)$, $\tilde{\alpha}_n = g(\theta_n, \sigma_n, \nu_n)$ as sufficiently smooth functions of $\theta_n, \sigma_n, \nu_n$. It is easy to show that $f(z_n) = O(f(x_n)^4)$ if and only if $h_{00} = h_{10} = h_{01} = 1$, where $h_{ij} = h^{(i,j)}(0,0)$, ($i \geq 0, j \geq 0$). Hence, under the restriction $h_{11} = h_{02} = h_{21} = h_{12} = h_{03} = 1$, (4) is optimal if and only if

$$\begin{aligned} g_{000} &= 1, \\ g_{100} &= g_{010} = g_{001} = 1, \\ g_{101} &= g_{011} = 2, \\ g_{200} &= h_{20}, \quad g_{110} = 1, \quad g_{020} = 0, \end{aligned}$$

$$g_{300} = h_{20} + h_{30} - 1, \quad g_{210} = h_{20} - 1, \quad g_{120} = g_{030} = -1.$$

Those can be easily checked with using (6). For the optimal formula, the remainder term is $O(f(x_n)^4)$ in (8) because $v_n = O(f(x)^2)$. In this sense, we can say that (4) is **optimal** if and only if $\tilde{\tau}_n, \tilde{\alpha}_n$ can be written as (7) and (8).

When $\gamma \rightarrow 0$ the Formula (7) leads to (2) and the Formula (8) leads to (3). A query may arise that there exists an optimal (DF) variant (4) for each optimal (DP) variant (1) and vice versa. If yes, how to find its (DF) variant? To respond this we use the connection of formulas (3) and (8). Actually, from (3) and (8) we deduce that

$$\tilde{\alpha}_n = \alpha_n(x_n, y_n, \phi_n) + (\tilde{d}_n - 2)(1 + \theta_n + (\tilde{d}_n + \tilde{\beta})\theta_n^2 + 2v_n)\theta_n, \tag{9}$$

where $\alpha_n(x_n, y_n, \phi_n)$ is obtained replacing $f'(x_n)$ by ϕ_n in $\alpha_n(x_n, y_n, f'(x_n))$ in (1). From (9) we find that

$$\alpha_n(x_n, y_n, f'(x_n)) = \tilde{\alpha}_n(x_n, y_n, \phi_n)|_{\phi_n=f'(x_n)} = \tilde{\alpha}_n(x_n, y_n, f'(x_n)). \tag{10}$$

These relations (9) and (10) give the rule of mutual transition of (DP) and (DF) variants. There exists the one optimal (DP) variant (4) for each optimal (DF) variant (1). The converse does not true. Namely there are several (DF) variants of (DP).

3. Application of Sufficient Convergence Condition to Derive New DF Iterations

Now we give the application of Theorem 2 to construct new iterations. The sufficient convergence conditions (7) and (8) allow us to design new derivative-free optimal methods. Depending on the form of $\tilde{\alpha}_n$ we can obtain different iterations. We consider some special cases.

1) Let $\tilde{\alpha}_n$ in (4) be a form

$$\tilde{\alpha}_n = \varphi(\theta_n) + \psi(v_n) + \mu\left(\frac{f(z_n)}{f(x_n)}\right), \tag{11}$$

where φ, ψ , and μ are smooth enough functions. As regarding the iteration (4) with $\tilde{\alpha}_n$ given by (11) we give the following result.

Theorem 3. *The iteration (4) with $\tilde{\tau}_n$ given by (7) and with $\tilde{\alpha}_n$ given by (11) have the order of convergence eight, if the following conditions hold:*

$$\begin{aligned} \varphi(0) = 1, \quad \varphi'(0) = \tilde{d}_n, \quad \varphi''(0) = 2(\tilde{\beta} + \tilde{c}_n), \\ \varphi'''(0) = 6(\tilde{\gamma} + \tilde{d}_n(\tilde{\beta} - 1 - \tilde{c}_n^2)), \\ \psi(0) = 0, \quad \psi'(0) = 1, \\ \mu(0) = 0, \quad \mu'(0) = 2\tilde{d}_n. \end{aligned} \tag{12}$$

Proof. Using the Taylor expansion of smooth enough functions $\varphi(\theta_n), \psi(v_n)$ and $\mu(\theta_n, v_n)$ we obtain an expression for (11). The comparison of this expression with sufficient condition (8) gives conditions (12).

When $\tilde{\beta} = \tilde{\gamma} = 0$ in (7) the Theorem 3 leads to a theorem in [5]. That is to say, the similar theorem was proved in [5] only for special case of $\tilde{\tau}_n$:

$$\tilde{\tau}_n = 1 + \tilde{d}_n \theta_n. \quad (13)$$

Therefore, Theorem 3 is more general, than that of [5]. Note that, in [5] are proposed four variants of $\tilde{\alpha}_n$ that include redundant terms like ν_n^2 and $\left(\frac{f(z_n)}{f(\omega_n)}\right)^2$. By neglecting these terms, $\tilde{\alpha}_n$ can be simplified essentially without loss of the order of convergence. When $\gamma = 0$ the condition (12) reduced to

$$\begin{aligned} \varphi(0) = 1, \quad \varphi'(0) = 2, \quad \varphi''(0) = 2(\tilde{\beta} + 1), \quad \varphi'''(0) = 6(\tilde{\gamma} + 2\tilde{\beta} - 4), \\ \psi(0) = 0, \quad \psi'(0) = 1, \quad \mu(0) = 0, \quad \mu'(0) = 4. \end{aligned} \quad (14)$$

It means that the derivative presence variant (1) with parameters given by (7) and (11) has a convergence order eight under conditions (14).

Thukral and Petković considered in [1] the particular case of (1) with $\tilde{\alpha}_n$ given by (11) and with

$$\tilde{\tau}_n = \frac{1 + b\theta_n}{1 + (b-2)\theta_n} = 1 + 2\theta_n + 2(2-b)\theta_n^2 + 2(2-b)^2\theta_n^3.$$

In this case $\tilde{\beta} = 2(2-b)$ and $\tilde{\gamma} = 2(2-b)^2$ and the condition (14) coincides with that of [1]. They also considered another particular case of (1) with $\tilde{\alpha}_n$ given by (11) and

$$\tilde{\tau}_n = \theta_n + \frac{1}{1-\theta_n} = 1 + 2\theta_n + \theta_n^2 + \theta_n^3 + \dots.$$

In this case $\tilde{\beta} = \tilde{\gamma} = 1$ and the condition (14) leads to that of [1]. The function $\varphi(\theta_n)$ in (11) can be written as

$$\varphi(\theta_n) = \tilde{\tau}_n + \tilde{c}_n \theta_n^2 + \tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2) \theta_n^3. \quad (15)$$

Due to generating function method [10] instead of $\tilde{\tau}_n$ we can take any function H

$$\tilde{\tau}_n = H(\theta_n) = \frac{c + (H'(0)c + d)\theta_n + \omega\theta_n^2}{c + d\theta_n + b\theta_n^2}, \quad b, c, d, \omega \in R, \quad (16)$$

satisfying conditions

$$H(0) = 1, \quad H'(0) = \tilde{d}_n, \quad H''(0) = 2\tilde{\beta}, \quad H'''(0) = 6\tilde{\gamma}.$$

As a result, we have a family of optimal derivative-free three-point methods (4) with (11), (15), and (16). The constants $\tilde{\beta}$ and $\tilde{\gamma}$ can be expressed through b, c, d and ω as:

$$\tilde{\beta} = \frac{\omega - b - dH'(0)}{c}, \quad \tilde{\gamma} = \frac{d(2b - \omega)}{c^2} + (H'(0)c + d) \left(\frac{d^2}{c^3} - \frac{b}{c^2} \right) - \frac{d^3}{c^3}.$$

That is we have the iterations (4) with $\tilde{\tau}_n$ is given by (16) and $\tilde{\alpha}_n$ is given by

$$\tilde{\alpha}_n = \tilde{\tau}_n + \tilde{c}_n \theta_n^2 + \tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2) \theta_n^3 + (1 + 2\tilde{d}_n \theta_n) \nu_n. \tag{17}$$

Note that the choice of parameter $\tilde{\tau}_n$ defined by (16) includes almost all the choices listed in **Table 1** as particular cases. Thus the family of iterations (4) with (16) and (17) represents a wide class of optimal derivative-free three-point iterations.

2) Let $\tilde{\alpha}_n$ in (4) be a form

$$\tilde{\alpha}_n = \tilde{\tau}_n + K(\theta_n, \nu_n), \tag{18}$$

where $\tilde{\tau}_n$ is given by (7) and $K(\theta_n, \nu_n)$ is sufficient smooth function of θ_n and ν_n .

Theorem 4. *The iteration (4) with $\tilde{\tau}_n$ given by (7) and $\tilde{\alpha}_n$ given by (18) has the order of convergence eight, if the following conditions hold:*

$$\begin{aligned} K(0,0) = K'_\theta(0,0) = 0, \quad K'_\nu(0,0) = 1, \quad K''_{\theta\nu}(0,0) = 2\tilde{d}_n, \\ K''_{\theta\theta}(0,0) = 2\tilde{c}_n, \quad K'''_{\theta\theta\theta}(0,0) = 6\tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2). \end{aligned} \tag{19}$$

Proof. From (7) and (8) it is clear that

$$K(\theta_n, \nu_n) = \tilde{c}_n \theta_n^2 + \tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2) \theta_n^3 + (1 + 2\tilde{d}_n \theta_n) \nu_n + O(f(x)^4) \tag{20}$$

which holds under conditions (19).

The (DP) variant of this iteration is obtained from (4), (7), and (18) when $\gamma \rightarrow 0$. Note that the similar scheme was considered in [2].

In some cases, the form

$$\tilde{\alpha}_n = \tilde{\tau}_n \left(1 + \frac{K(\theta_n, \nu_n)}{\tilde{\tau}_n} \right) = \tilde{\tau}_n W(\theta_n, \nu_n) \tag{21}$$

obtained from (18) is useful. Using (20) we obtain

$$\begin{aligned} W(\theta_n, \nu_n) = 1 + \tilde{c}_n \theta_n^2 + (-\tilde{c}_n \tilde{d}_n + \tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2)) \theta_n^3 \\ + (1 + \tilde{d}_n \theta_n) \nu_n + O(f(x_n)^4). \end{aligned} \tag{22}$$

For the iteration (4) with (7) and (21) we can formulate the following:

Theorem 5. *The iteration (4) with (7) and (21) has the order of convergence eight, if the following conditions hold:*

$$\begin{aligned} W(0,0) = 1, \quad W'_\theta(0,0) = 0, \quad W''_{\theta\theta}(0,0) = 2\tilde{c}_n, \\ W'''_{\theta\theta\theta}(0,0) = 6(-\tilde{c}_n \tilde{d}_n + \tilde{d}_n (\tilde{\beta} - 1 - \tilde{c}_n^2)), \\ W'_\nu(0,0) = 1, \quad W''_{\nu\theta}(0,0) = \tilde{d}_n. \end{aligned} \tag{23}$$

Proof. If we take (22) into account in the Taylor expansion of function $W(\theta_n, \nu_n)$ we arrive at (23).

When $\gamma = 0$ the conditions (23) take a form

$$\begin{aligned} W(0,0) = 1, \quad W'_\theta(0,0) = 0, \quad W''_{\theta\theta}(0,0) = 2, \\ W'''_{\theta\theta\theta}(0,0) = 12(\tilde{\beta} - 3), \quad W'_\nu(0,0) = 1, \quad W''_{\nu\theta}(0,0) = 2. \end{aligned} \tag{24}$$

Remark. Obviously, as for $\tilde{\tau}_n$ one can take any function H given by (16) in the formulas (17) and (21).

Note that in [3] were obtained some conditions that guarantee order eight of the method (4) with (7) and (21) i.e.,

$$\begin{aligned}\tilde{\tau}_n &= h(\theta_n, s_n) = 1 + \theta_n + s_n + \frac{a}{2}\theta_n^2 + b\theta_n s_n + \frac{c}{2}s_n^2 + \dots, \\ s_n &= \frac{\tilde{c}_n f(y_n)}{f(x_n)}, \\ \tilde{\alpha}_n &= h(\theta_n, s_n) \cdot \mu(\theta_n, s_n, v_n), \\ \mu(\theta_n, s_n, v_n) &= 1 + v_n + \frac{d}{2}v_n^2 + \theta_n s_n + \theta_n v_n + s_n v_n + \frac{a-2}{2}\theta_n^3 \\ &\quad + \frac{c-2}{2}s_n^3 + \frac{m}{6}v_n^3 + \frac{a+2b-4}{2}\theta_n s_n^2 + \frac{2b+c-4}{2}\theta_n s_n^2 \\ &= 1 + \tilde{c}_n \theta_n^2 + (1 + \tilde{d}_n \theta_n)v_n + (\tilde{d}_n \tilde{\beta} - (2 + \tilde{c}_n^3 + 2\tilde{c}_n)\tilde{d}_n)\theta_n^3 \\ &\quad + O(f(x_n)^4)\end{aligned}\tag{25}$$

that does not coincide with (22). Moreover, the terms $\frac{d}{2}v_n^2$ and $\frac{m}{6}v_n^3$ seem to be redundant, because it suffices to determine $\tilde{\alpha}_n$ with accuracy $O(f(x_n)^4)$.

Note that (DP) methods with (7) and (21) are often used. For example, Kung-Traub's eighth-order method [21] has a form (1) with

$$\begin{aligned}\tilde{\tau}_n &= \frac{1}{(1-\theta_n)^2}, \\ W(\theta_n, v_n) &= \frac{f(y_n)(f^2(x_n) + f(y_n)(f(y_n) - f(z_n)))}{(f(x_n) - f(z_n))^2 (f(y_n) - f(z_n))} = \frac{1 + \theta_n(\theta_n - \theta_n v_n)}{(1 - \theta_n v_n)^2 (1 - v_n)}.\end{aligned}\tag{26}$$

The Bi-Wu-Ren's optimal eighth-order method [22] has a form (1) with

$$\begin{aligned}\tilde{\tau}_n &= h(\theta_n), \\ \alpha_n &= \frac{f'_n(x_n)(f(x_n) + \beta f(z_n))}{f(x_n) + (\beta - 2)f(z_n)} \frac{1}{f[z_n, y_n] + f[z_n, x_n, x_n](z_n - y_n)},\end{aligned}\tag{27}$$

where

$$f[z_n, y_n] = \frac{f(y_n) - f(z_n)}{y_n - z_n}, \quad f[z_n, x_n, x_n] = \frac{f'(x_n) - f[z_n, x_n]}{x_n - z_n}.$$

But (27) is not the example for (21).

The Sharma and Arora's optimal eighth-order method [21] has a form (1) with

$$\begin{aligned}\tilde{\tau}_n &= \frac{1}{1-2\theta_n}, \\ \tilde{\alpha}_n &= \tilde{\tau}_n \frac{4\theta_n^2(1-\theta_n)^2(1-v_n)}{(-1+\theta_n v_n)((1-2\theta_n)^2 + (3-4\theta_n)\theta_n v_n)}.\end{aligned}\tag{28}$$

Moreover, we suggest that more general theory for $\tilde{\tau}_n$ as

$$\tilde{\tau}_n = 1 + \theta_n + \sigma_n + h_{20}\theta_n^2 + h_{11}\theta_n\sigma_n + h_{02}\sigma_n^2 + h_{30}\theta_n^3 + h_{21}\theta_n^2\sigma_n + h_{12}\theta_n\sigma_n^2 + h_{03}\sigma_n^3 + \dots$$

3) Let $\tilde{\alpha}_n$ in (4) be a form

$$\tilde{\alpha}_n = \frac{\phi_n}{f[z_n, y_n] + (z_n - y_n)f[z_n, y_n, x_n] + (z_n - y_n)(z_n - x_n)f[z_n, y_n, x_n, \omega_n]} \tag{29}$$

that often used in practice, see [4] [12] [14] [15] [16]. Of course, $\tilde{\tau}_n$ and $\tilde{\alpha}_n$ given by (7) and (29) satisfy the sufficient conditions (7) and (8). The (DP) variant of (4) with (7) and (29) has a form (1)

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ z_n &= \psi_4(x_n, y_n), \\ x_{n+1} &= z_n - \alpha_n \frac{f(z_n)}{f'(x_n)}, \end{aligned} \tag{30}$$

where $\alpha_n = \frac{f'(x_n)}{f[z_n, y_n] + (z_n - y_n)f[y_n, x_n, x_n]}$.

In [6] is proposed the eighth-order iteration (1) with (29) ($\gamma \rightarrow 0$) and special $\bar{\tau}_n$

$$\bar{\tau}_n = 1 + \frac{1}{\theta_n} + \left(1 + \frac{1}{1 + \frac{f(x_n)}{f'(x_n)}} \right) \theta_n. \tag{31}$$

Our iteration (1) with (2) and (30) is more general than that of [6].

4) Let $\tilde{\alpha}_n$ in (4) be a form

$$\tilde{\alpha}_n = \frac{\phi_n(1 + A\theta_n + B\theta_n^2 + C\theta_n^3 + (\omega + \Delta\theta_n)\nu_n)}{af[x_n, z_n] + bf[z_n, y_n] + cf[x_n, y_n]}, \quad a + b + c = 1, \tag{32}$$

where $a, b, c \in R$.

We shall find the coefficients A, B, C and ω, Δ such that the iteration (4) with (7) and (32) has the order of convergence eight and state the following:

Theorem 6. *The iteration (4) with (7) and (32) has the order of convergence eight, if the following conditions hold:*

$$\begin{aligned} A &= (1-b)(\tilde{d}_n - 1), \quad B = (\tilde{\beta} - \tilde{d}_n)(1-b) + \tilde{c}_n(1-a), \\ C &= a - 1 + (b-a)\beta + (1-b)\gamma + (a + \beta - 2)\tilde{c}_n + (c-2)\tilde{c}_n^2 - \tilde{c}_n^3, \\ \omega &= 1 - b, \quad \Delta = -a + b - 1 + \tilde{d}_n(2-b). \end{aligned} \tag{33}$$

Proof. Using the following relations

$$f[x_n, y_n] = \phi_n(1 - \theta_n), \quad f[z_n, y_n] = \frac{\phi_n}{\tilde{\tau}_n}(1 - \nu_n),$$

$$f[x_n, z_n] = \frac{\phi_n}{1 + \tilde{\tau}_n \theta_n} (1 - \theta_n \nu_n),$$

$$\frac{1}{\tilde{\tau}_n} = 1 - \tilde{d}_n \theta_n + (\tilde{d}_n^2 - \tilde{\beta}_n) \theta_n^2 + (2\tilde{\beta}_n \tilde{d}_n - \tilde{\gamma} - \tilde{d}_n^3) \theta_n^3 + \dots, \quad (34)$$

$$\frac{1}{1 + \tilde{\tau}_n \theta_n} = 1 - \theta_n + (1 - \tilde{d}_n) \theta_n^2 + (2\tilde{d}_n - \tilde{\beta}_n - 1) \theta_n^3 + \dots,$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots, \quad |x| < 1,$$

we get

$$\frac{\phi_n}{af[x_n, z_n] + bf[z_n, y_n] + cf[x_n, y_n]}$$

$$= 1 + (a + c + b\tilde{d}_n) \theta_n + (a(\tilde{d}_n - 1) + b(\tilde{\beta} - \tilde{d}_n^2) + (a + c + b\tilde{d}_n)^2) \theta^2$$

$$+ (a(\tilde{\beta} + 1 - 2\tilde{d}_n) + b(\tilde{\gamma} + \tilde{d}_n^3 - 2\tilde{\beta}\tilde{d}_n)) \theta^3$$

$$+ 2(a + c + b\tilde{d}_n)(a(\tilde{d}_n - 1) + b(\tilde{\beta} - \tilde{d}_n^2)) + (a + c + b\tilde{d}_n)^3 \theta^3$$

$$+ (b + (a + 2b(a + c) + (2b - 1)b\tilde{d}_n) \theta_n) \nu_n + O(f(x_n^4)). \quad (35)$$

Substituting (35) into (32) and using the sufficient convergence condition (8) we arrive at (33).

Thus, we have a family of optimal three-point (DF) the iteration (4) with (7) and (32) that contains three parameters a , b and c . Now, we consider some particular cases of the iteration (4) with (7) and (32). Let $a = b = 1$ and $c = -1$. Then from (33) we find that

$$A = B = 0, \quad C = \frac{(1 + \gamma\phi_n)\tilde{\beta} - \tilde{d}_n - 3 - \gamma\phi_n}{(1 + \gamma\phi_n)^2}, \quad \omega = 0, \quad \Delta = \tilde{c}_n.$$

Hence we obtain

$$\tilde{\alpha}_n = \frac{(1 + ((1 + \gamma\phi_n)\tilde{\beta} - \tilde{d}_n - 3 - \gamma\phi_n)\tilde{c}_n^2 \theta_n^3 + \tilde{c}_n \theta_n \nu_n) \phi_n}{f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n]}$$

$$\approx \frac{\left(1 + \frac{f(z_n)}{f(\omega_n)}\right) \left(1 + ((1 + \gamma\phi_n)\tilde{\beta} - \tilde{d}_n - 3 - \gamma\phi_n) \frac{f^3(y_n)}{f^2(\omega_n) f(x_n)}\right) \phi_n}{f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n]}, \quad (36)$$

or

$$\tilde{\alpha}_n \approx \frac{\phi_n}{\left(1 - \frac{f(z_n)}{f(\omega_n)}\right) \left(1 - \hat{C} \frac{f^3(y_n)}{f^2(\omega_n) f(x_n)}\right) (f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n])}, \quad (37)$$

where $\hat{C} = (1 + \gamma\phi_n)\tilde{\beta} - \tilde{d}_n - 3 - \gamma\phi_n$. The sign \approx in (37) indicates that it holds with accuracy $O(f(x_n^4))$. Now, we consider concrete choice of $\tilde{\tau}_n$:

$$\tilde{\tau}_n = \frac{1}{1 - \tilde{d}_n \theta_n + p\tilde{c}_n \theta_n^2}, \quad p \in N. \quad (38)$$

For the choice (38) we have

$$\tilde{\beta} = \tilde{d}_n^2 - p\tilde{c}_n$$

and

$$\hat{C} = -(p+1).$$

The iteration (4) with (38) and (36) (or (37)) is converted to one given by Soleymani in [23] for $p = 0$ and one given by Thukral in [7] for $p = 1$. For the choice $p = -1$ the parameter $\tilde{\alpha}_n$ is simplified as

$$\tilde{\alpha}_n = \frac{\left(1 + \frac{f(z_n)}{f(\omega_n)}\right)\phi_n}{f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n]}, \tag{39}$$

or using $1 + \frac{f(z_n)}{f(\omega_n)} = \frac{1}{1 - \frac{f(z_n)}{f(\omega_n)}} + O(f(x_n)^6)$ we have

$$\tilde{\alpha}_n = \frac{\phi_n}{\left(1 - \frac{f(z_n)}{f(\omega_n)}\right)(f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n])}. \tag{40}$$

Let

$$\tilde{v}_n = \frac{1}{1 - \tilde{d}_n\theta_n} = 1 + \tilde{d}_n\theta_n + \tilde{\beta}\theta_n^2 + \tilde{\gamma}\theta_n^3 + \dots, \tag{41}$$

with

$$\tilde{\beta} = \tilde{d}_n^2, \quad \tilde{\gamma} = \tilde{d}_n^3.$$

Then $C = -\tilde{c}_n^2$ and we have

$$\tilde{\alpha}_n = \frac{(1 - \tilde{c}_n^2\theta_n^3 + \tilde{c}_n\theta_n\nu_n)\phi_n}{f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n]}. \tag{42}$$

The iteration (4) with (41) and (42) coincides with one given by Soleymani in [23] with

$$\tilde{\alpha}_n = \frac{\left(1 + \theta_n^4 - (1 + \gamma\phi_n)\frac{f^3(y_n)}{(1 + \gamma\phi_n)^3 f^3(x_n)} - \nu_n^2 + \tilde{c}_n\sigma_n\nu_n + (\theta_n\nu_n)^2\right)\phi_n}{f[x_n, z_n] + f[z_n, y_n] - f[x_n, y_n]},$$

here we can neglect the redundant terms $\theta_n^4 - \nu_n^2 + \theta_n^2\nu_n^2$. Let $a = 1$, and $b = c = 0$. Then from (33) we find that

$$A = \tilde{d}_n - 1, \quad B = \tilde{\beta} - \tilde{d}_n, \quad C = \tilde{\beta}(\tilde{d}_n - 2) + \tilde{\gamma} - \tilde{c}_n\tilde{d}_n^2, \\ \omega = 1, \quad \Delta = 2(\tilde{d}_n - 1).$$

The Formula (32) is converted to

$$\tilde{\alpha}_n = \frac{\left(1 + (\tilde{d}_n - 1)\theta_n + (\tilde{\beta} - \tilde{d}_n)\theta_n^2 + (\tilde{\beta}(\tilde{d}_n - 2) + \tilde{\gamma} - \tilde{c}_n\tilde{d}_n^2)\theta_n^3 + (1 + 2(\tilde{d}_n - 1)\theta_n)\nu_n\right)\phi_n}{f[x_n, z_n]}. \tag{43}$$

On the other hand, the direct calculation using relations (34) gives

$$\begin{aligned} & \left(1 - \frac{f(z_n)}{f(\omega_n)}\right)^{-1} \left(1 - \eta \frac{f^3(y_n)}{f(\omega_n)f^2(x_n)}\right) \times \frac{f[x_n, y_n]}{f[z_n, y_n]} \\ &= 1 + (\tilde{d}_n - 1)\theta_n + (\tilde{\beta} - \tilde{d}_n)\theta_n^2 + (\tilde{\gamma} - \tilde{\beta} - \eta\tilde{c}_n)\theta_n^3 \\ & \quad + (1 + 2\tilde{c}_n\theta_n)v_n + O(f(x_n)^4), \eta \in R. \end{aligned} \quad (44)$$

We choose parameter η in (44) such that the expression (44) coincides with the numerator of (43) within accuracy $O(f(x_n)^4)$. That is to say, that

$$\eta = -\tilde{\beta} + \tilde{d}_n^2.$$

As a result, (43) can be rewritten as

$$\tilde{\alpha}_n = \left(1 - \frac{f(z_n)}{f(\omega_n)}\right)^{-1} \left(1 - (\tilde{d}_n^2 - \tilde{\beta}) \frac{f^3(y_n)}{f(\omega_n)f^2(x_n)}\right) \frac{f[x_n, y_n]\phi_n}{f[x_n, z_n] \cdot f[z_n, y_n]}. \quad (45)$$

Thus, we find a family of optimal (DF) iteration (4) with (7) and (45), that contains some existing iterations as particular cases. Thukral in [24] proposed eighth-order derivative-free iterations (called M_1, M_2, M_3) for some special \tilde{c}_n :

$$\tilde{c}_n = \frac{1}{1 - \tilde{d}_n\theta_n + \tilde{c}_n\theta_n^2} = 1 + \tilde{d}_n\theta_n + (\tilde{d}_n^2 - \tilde{c}_n)\theta_n^2 + \dots$$

In this case $\tilde{\beta} = \tilde{d}_n^2 - \tilde{c}_n$ and hence $\eta = \tilde{c}_n$, the $\tilde{\alpha}_n$ given by (45) leads to that of M_1 and M_3 in [24]. So, the Thukral's method (M_1, M_2, M_3) are included in our family of (4) with (7) and (45). Thukral in [24] proposed also Petković type methods ($P1, P2$). For $P1$ we get $\tilde{c}_n = 1 + \tilde{d}_n\theta_n$, i.e. $\tilde{\beta} = 0$. In this case $\eta = \tilde{d}_n^2$ in (45) and our family of method (4) with (45) converted to $P1$. For $P2$ we get

$$\tilde{c}_n = \frac{1 + \theta_n}{1 - \tilde{c}_n\theta_n}. \quad (46)$$

In this case $\tilde{\beta} = \tilde{c}_n\tilde{d}_n$ and $\eta = \tilde{d}_n$. Thus, our family of method (4) with (45) converted to $P2$. It means that the ($P1, P2$) methods are also included in our family of (4) with (7) and (45). As stated above for the choice of (41) we have $\tilde{\beta} = \tilde{d}_n^2$, so (45) is simplified as

$$\tilde{\alpha}_n = \left(1 - \frac{f(z_n)}{f(\omega_n)}\right)^{-1} \frac{f[x_n, y_n]\phi_n}{f[x_n, z_n]f[z_n, y_n]}. \quad (47)$$

Thus, we have optimal (DF) methods

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{\phi_n}, \\ z_n &= y_n - \tilde{c}_n \frac{f(y_n)}{\phi_n}, \quad \tilde{c}_n = \frac{1}{1 - \tilde{d}_n\theta_n}, \\ x_{n+1} &= z_n - \tilde{\alpha}_n \frac{f(z_n)}{\phi_n}, \end{aligned} \quad (48)$$

where $\tilde{\alpha}_n$ is defined by (47). This is (DF) variant of Sharma and Sharma’s optimal methods given in [19] [21] within accuracy $O(f(x_n)^4)$. It means that we develop (DF) variant of Sharma and Sharma’s method.

4. Numerical Experiments

In this section, we make some numerical experiments to show the convergence behavior of the presented derivative-free method (4) with parameters $\tilde{\tau}_n$ and $\tilde{\alpha}_n$. We also compare them with the ones developed by Soleymani [23], Thukral [7] [24] and Sharma *et al.* [19]. For this purpose, we consider smooth and non-smooth nonlinear functions, which are given as follows:

$$f_1(x) = e^{-x} + \frac{x}{5} - 1, x^* = 4.9651142$$

$$f_2(x) = e^{x^3-x} - \cos(x^2 - 1) + x^3 + 1, x^* = -1$$

$$f_3(x) = \sin x + e^{x^2} - 1, x^* = 0$$

$$f_4(x) = \frac{1}{x} - |x|, x^* = 1.$$

All computations are performed using the programming package Maple18 with multiple-precision arithmetic and 2500 significant digits. The test functions have been used with stopping criterion $|x_n - x^*| < 10^{-250}$, where x^* is a root of $f(x)$ and the approximation x_n to x^* . In all examples, we consider that the parameter $\gamma = -0.01$ and that $\alpha = -2$ in Chebyshev-Halley’s method.

Nowadays, high order methods are important due to scientific computations in many areas of science and engineering use. For instance, planetary orbit calculation, radiation calculations and many real life problems demand higher precision for desired results [4] [13]. The first example addresses this situation and we apply the presented methods to solve one such physical problem. In [4] have considered one of the famous classical physics problem which is known as Planck’s radiation law problem. First nonlinear function f_1 arises from this problem.

$f_1(x) = 0$ has two zeros. Obviously, one of the roots $x = 0$ is not taken for discussion. Another root is $x^* \approx 4.965114231744276303699$. Now, we give some numerical experiments and compare new methods with some well-known methods for the smooth function f_1 using the initial guess $x_0 = 6$. In **Table 2** and **Table 3**, we exhibit computational order of convergence (COC) and absolute error $|x_n - x^*|$ as well as iteration numbers n are displayed. For presented methods and test functions, by using (see, e.g., [4] [11] [16])

$$COC \approx \frac{\ln\left(\frac{|x_n - x^*|}{|x_{n-1} - x^*|}\right)}{\ln\left(\frac{|x_{n-1} - x^*|}{|x_{n-2} - x^*|}\right)},$$

we have computed the order of convergence.

From **Table 2**, we can observe that computed results completely support the

Table 2. Convergence behavior of scheme (4) for $f_1(x), x_0 = 6$.

methods	$\tilde{\alpha}_n$	$\tilde{\tau}_n$	n	$ x^* - x_n $	COC
(4)	(32), $(a = b = 1, c = -1)$	(38), $(p = -1)$	3	0.3130e-674	8.00
	(32), $(b = 1, a = 1, c = -1)$	(13)	3	0.3422e-670	8.00
	(32), $(b = 1, a = c = 0)$	(13)	3	0.1346e-667	8.00
	(32), $(b = 1, a = c = 0)$	(38), $(p = 0)$	3	0.1078e-670	8.00
	(32), $(b = 1, a = -1, c = 1)$	(13)	3	0.3285e-665	8.00
	(32), $(b = 1, a = -1, c = 1)$	(38), $(p = 0)$	3	0.3378e-668	8.00
Soleymani [23]	(32), $(a = b = 1, c = -1)$	(38), $(p = 0)$	3	0.2023e-673	8.00
Thukral [7]	(32), $(a = b = 1, c = -1)$	(38), $(p = 1)$	3	0.1239e-672	8.00
M_1, M_2, M_3 [24]	(45), $\left(\eta = \frac{1}{1 + \gamma\phi_n}\right)$	(38), $(p = 1)$	3	0.4813e-670	8.00
$P1$ Thukral [24]	(45), $(\beta = 0)$	(13)	3	0.1271e-667	8.00
$P2$ Thukral [24]	(45), $\left(\beta = \frac{\tilde{d}_n}{1 + \gamma\phi_n}\right)$	(46)	3	0.3112e-669	8.00
Sharma et al. [19]	(45), $(\eta = 0)$	(38), $(p = 0)$	3	0.7836e-671	8.00

Table 3. Some particular cases of (4) with $\tilde{\tau}_n$ (16) and $\tilde{\alpha}_n$ (29).

methods	$\tilde{\tau}_n = H(\theta_n)$	n	$ x^* - x_n $	COC
	choices of parameters			
Lotfi [15]	$c = 1, d = -\frac{1}{1 + \gamma\phi_n}, b = 0, \omega = \frac{\tilde{d}_n}{2}$	3	0.2785e-672	8.00
King's type [16]	$c = \omega = 1, d = \beta - 1 - \tilde{d}_n, b = \frac{2 - \beta}{1 + \gamma\phi_n}, (\beta = 2)$	3	0.1004e-674	8.00
Zheng [12]	$c = 1, d = -\tilde{d}_n, b = \omega = 0$	3	0.9462e-674	8.00
Sharma [14]	$c = 1, d = -\frac{1}{1 + \gamma\phi_n}, b = \omega = 0$	3	0.4414e-673	8.00
Chebyshev-Halley [4]	$c = 1, d = -\left(2\alpha + \frac{1}{1 + \gamma\phi_n}\right), b = \frac{2\alpha}{1 + \gamma\phi_n}, \omega = H(\theta_n)$	3	0.7466e-671	8.00

theory of convergence discussed in previous section. In addition to the comparison of new methods with other methods we include some special cases of proposed family (4) in **Table 3**.

Table 4 illustrates the number of iterations needed to achieve approximate solution and absolute residual error of the corresponding function $|f(x_n)|$ using the stopping criterion $|x_n - x^*| < 10^{-250}$.

As $\tilde{\tau}_n$ in **Table 2** is used same in each method, it is shown in **Table 4**.

Table 4. Comparisons between different methods.

$\tilde{\alpha}_n$	$f(x)$	$f_2(x)$	$f_3(x)$	Methods		
	x_0	-0.6	-1.5			
	a, b, c	n	$ f(x_n) $	n	$ f(x_n) $	
(32)	$a = 1$					
	$b = 1$	4	1.60 (-691)	4	1.37 (-349)	
	$c = -1$					
	$a = 1$					
	$b = 1$	5	2.44 (-395)	4	2.20 (-260)	
	$c = -1$					
	$a = 0$					
	$b = 1$		-	5	6.80 (-1233)	(4)
	$c = 0$					
	$a = 0$					
	$b = 1$	4	9.48 (-541)	4	9.00 (-300)	
	$c = 0$					
	$a = -1$					
	$b = 1$		-	5	8.55 (-988)	
	$c = 1$					
	$a = -1$					
$b = 1$	4	7.55 (-316)	5	2.25 (-1777)		
$c = 1$						
$a = 1$						
$b = 1$	4	2.72 (-484)	-	-	Soleymani [23]	
$c = -1$						
$a = 1$						
$b = 1$	4	1.75 (-449)	4	7.99 (-702)	Thukral [7]	
$c = -1$						
(45)	$\eta = \frac{1}{1 + \gamma\phi_n}$	4	1.75 (-449)	4	9.49 (-267)	M_1, M_3 [24]
	$\beta = 0$	5	1.05 (-875)	5	4.59 (-1301)	$P1$ Thukral [24]
	$\beta = \frac{\tilde{d}_n}{1 + \gamma\phi_n}$	5	1.09 (-707)	5	1.14 (-1709)	$P2$ Thukral [24]
	$\eta = 0$	5	2.98 (-1069)	4	2.33 (-373)	Sharma et al. [19]

Furthermore, when the iteration diverges for the considered initial guess x_0 , we denote it by “-”.

From **Table 4**, we see that the convergence behavior of the presented families with different parameters and the iteration number n are the same as for all considered methods.

The result of **Table 5** demonstrates that new methods iteration numbers are

Table 5. Comparison of various iterative methods for $f_4(x), x_0 = -2$.

methods	$\tilde{\alpha}_n$	$\tilde{\tau}_n$	n	COC
	(32), $(a = b = 1, c = -1)$	(38), $(p = -1)$	6	8.00
	(32), $(b = 1, a = 1, c = -1)$	(13)	5	8.00
(4)	(32), $(b = 1, a = c = 0)$	(13)	6	8.00
	(32), $(b = 1, a = c = 0)$	(38), $(p = 0)$	6	8.00
	(32), $(b = 1, a = -1, c = 1)$	(13)	5	8.00
	(32), $(b = 1, a = -1, c = 1)$	(38), $(p = 0)$	10	8.00
Soleymani [23]	(32), $(a = b = 1, c = -1)$	(38), $(p = 0)$	6	8.00
Thukral [7]	(32), $(a = b = 1, c = -1)$	(38), $(p = 1)$	8	8.00
$P1$ Thukral [24]	(45), $(\beta = 0)$	(13)	14	8.00
$P2$ Thukral [24]	(45), $\left(\beta = \frac{\tilde{d}_n}{1 + \gamma\phi_n}\right)$	(46)	8	8.00
Sharma et al. [19]	(45), $(\eta = 0)$	(38), $(p = 0)$	21	8.00

used lesser than other existing methods under condition $|x_n - x^*| < 10^{-250}$. However, the dynamic behavior of iterations may depend on the choices of parameters and problems under consideration. In sum, numerical results show that new iterative methods can be significant by its high precision and practical use.

5. Conclusion

We derive the necessary and sufficient conditions for derivative-free three-point iterations with the optimal order. The use of these conditions allows us to derive the families of optimal derivative-free iterations. We propose the families of optimal derivative-free iterations (4) with $\tilde{\tau}_n$ given by (16) and $\tilde{\alpha}_n$ given by (17), (29), (32), and (45). Our families include many existing iterations as particular cases, as well as new effective iterations. We reveal redundant terms in well-known methods given in [3] [5] [23]. Dropping these terms allows us to simplify their algorithms and save computation time.

Acknowledgements

The authors wish to thank the editor and anonymous referees for their valuable suggestions on the first version of this paper. This work was supported by the Foundation of Science and Technology of Mongolian under grant SST_18/2018.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Thukral, R. and Petković, M.S. (2010) A Family of Three-Point Methods of Optimal Order for Solving Nonlinear Equations. *The Journal of Computational and Applied Mathematics*, **233**, 2278-2284. <https://doi.org/10.1016/j.cam.2009.10.012>
- [2] Rhee, M.S., Kim, Y.I. and Neta, B. (2018) An Optimal Eighth-Order Class of Three-Step Weighted Newton's Methods and Their Dynamics behind the Purely Imaginary Extraneous Fixed Points. *International Journal of Computer Mathematics*, **95**, 2174-2211. <https://doi.org/10.1080/00207160.2017.1367387>
- [3] Petković, M.S., Neta, B., Petković, L.D. and Dzunic, J. (2014) Multipoint Methods for Solving Nonlinear Equations. *Applied Mathematics and Computation*, **226**, 635-660. <https://doi.org/10.1016/j.amc.2013.10.072>
- [4] Argyros, I.K., Kansal, M., Kanwar, V. and Bajaj, S. (2017) Higher-Order Derivative-Free Families of Chebyshev-Halley Type Methods with or without Memory for Solving Nonlinear Equations. *Applied Mathematics and Computation*, **315**, 224-245. <https://doi.org/10.1016/j.amc.2017.07.051>
- [5] Soleymani, F. and Khattri, S.K. (2012) Finding Simple Roots by Seventh- and Eighth-Order Derivative-Free Methods. *International Journal of Mathematical Models and Methods in Applied Sciences*, **6**, 45-52. <https://doi.org/10.1155/2012/932420>
- [6] Matthies, G., Salimi, M., Sharifi, S. and Varona, J.L. (2016) An Optimal Three-Point Eighth-Order Iterative Method without Memory for Solving Nonlinear Equations with Its Dynamics. *Japan Journal of Industrial and Applied Mathematics*, **33**, 751-766. <https://doi.org/10.1007/s13160-016-0229-5>
- [7] Thukral, R. (2011) Eighth-Order Iterative Methods without Derivatives for Solving Nonlinear Equations. *International Scholarly Research Network ISRN Applied Mathematics*, **2011**, Article ID: 693787. <https://doi.org/10.5402/2011/693787>
<https://www.hindawi.com/journals/isrn/2011/693787>
- [8] Soleymani, F. and Vanani, S.K. (2011) Optimal Steffensen-Type Methods with Eighth Order of Convergence. *Computers & Mathematics with Applications*, **62**, 4619-4626. <https://doi.org/10.1016/j.camwa.2011.10.047>
- [9] Zhanlav, T., Ulziibayar, V. and Chuluunbaatar, O. (2017) Necessary and Sufficient Conditions for the Convergence of Two and Three-Point Newton-Type Iterations. *Computational Mathematics and Mathematical Physics*, **57**, 1090-1100. <https://doi.org/10.1134/S0965542517070120>
- [10] Zhanlav, T., Chuluunbaatar, O. and Ulziibayar, V. (2017) Generating Function Method for Constructing New Iterations. *Applied Mathematics and Computation*, **315**, 414-423. <https://doi.org/10.1016/j.amc.2017.07.078>
- [11] Zhanlav, T., Chuluunbaatar, O. and Otgondorj, Kh. (2019) A Derivative-Free Families of Optimal Two- and Three-Point Iterative Methods for Solving Nonlinear Equations. *Computational Mathematics and Mathematical Physics*, **59**, 920-936. <https://doi.org/10.1134/S0965542519060149>
- [12] Zheng, Q., Li, J. and Huang, F. (2011) An Optimal Steffensen-Type Family for Solving Nonlinear Equations. *Applied Mathematics and Computation*, **217**, 9592-9597. <https://doi.org/10.1016/j.amc.2011.04.035>
- [13] Khattri, S.K. and Steihaug, T. (2014) Algorithm for Forming Derivative-Free Optimal Methods. *Numerical Algorithms*, **65**, 809-842. <https://doi.org/10.1007/s11075-013-9715-x>
- [14] Sharma, J.R., Guha, R.K. and Gupta, P. (2012) Some Efficient Derivative Free Me-

- thods with Memory for Solving Nonlinear Equations. *Applied Mathematics and Computation*, **219**, 699-707. <https://doi.org/10.1016/j.amc.2012.06.062>
- [15] Lotfi, T., Soleymani, F., Ghorbanzadeh, M. and Assari, P. (2015) On the Construction of Some Tri-Parametric Iterative Methods with Memory. *Numerical Algorithms*, **70**, 835-845. <https://doi.org/10.1007/s11075-015-9976-7>
- [16] Sharifi, S., Siegmund, S. and Salimi, M. (2016) Solving Nonlinear Equations by a Derivative-Free Form of the King's Family with Memory. *Calcolo*, **53**, 201-215. <https://doi.org/10.1007/s10092-015-0144-1>
- [17] Cordero, A., Hueso, J.L., Martinez, E. and Torregrosa, J.R. (2013) A New Technique to Obtain Derivative-Free Optimal Iterative Methods for Solving Nonlinear Equations. *The Journal of Computational and Applied Mathematics*, **252**, 95-102. <https://doi.org/10.1016/j.cam.2012.03.030>
- [18] Behl, R., Gonzalez, D., Maraju, P. and Motsa, S.S. (2018) An Optimal and Efficient General Eighth-Order Derivative-Free Scheme for Simple Roots. *The Journal of Computational and Applied Mathematics*, **330**, 666-675. <https://doi.org/10.1016/j.cam.2017.07.036>
- [19] Sharma, J.R. and Sharma, R. (2010) A New Family of Modified Ostrowskis Method with Accelerated Eighth-Order Convergence. *Numerical Algorithms*, **54**, 445-458. <https://doi.org/10.1007/s11075-009-9345-5>
- [20] Kung, H.T. and Traub, J.F. (1974) Optimal Order of One-Point and Multi-Point Iteration. *Journal of Applied and Computational Mathematics*, **21**, 643-651. <https://doi.org/10.1145/321850.321860>
- [21] Chun, C. and Neta, B. (2017) Comparative Study of Eighth-Order Methods for Finding Simple Roots of Nonlinear Equations. *Numerical Algorithms*, **74**, 1169-1201. <https://doi.org/10.1007/s11075-016-0191-y>
- [22] Bi, W., Wu, Q. and Ren, H. (2009) A New Family of Eighth-Order Iterative Methods for Solving Nonlinear Equations. *Applied Mathematics and Computation*, **214**, 236-245. <https://doi.org/10.1016/j.amc.2009.03.077>
- [23] Soleymani, F. (2011) On a Bi-Parametric Class of Optimal Eighth-Order Derivative-Free Methods. *International Journal of Pure and Applied Mathematics*, **72**, 27-37.
- [24] Thukral, R. (2012) A Family of Three-Point Derivative-Free Methods of Eighth-Order for Solving Nonlinear Equations. *Journal of Modern Methods in Numerical Mathematics*, **3**, 11-21. <https://doi.org/10.20454/jmnm.2012.281>

A comparative analysis of local cubic splines

T. Zhanlav¹ · R. Mijiddorj²

Received: 29 September 2016 / Revised: 23 May 2017 / Accepted: 15 May 2018
© SBMAC - Sociedade Brasileira de Matemática Aplicada e Computacional 2018

Abstract In this note, we develop a local construction of cubic splines and make a comparative analysis of local integro cubic splines. We also derive explicit formulae for a local integro cubic spline and its first two derivatives. These formulae are short and four-point ones that require less computational cost compared to an integro cubic spline quasi-interpolant.

Keywords Cubic splines · Approximation properties · End conditions

Mathematics Subject Classification 65D05 · 65D07

1 Introduction

In last years, the local construction of splines has attracted a lot of attention from researchers. For example, the local integro cubic spline was constructed in Zhanlav and Mijiddorj (2010). Quite recently, the integro cubic spline quasi-interpolant was developed in Boujraf et al. (2015) and some comparisons were made with the results obtained in Zhanlav and Mijiddorj (2010). As seen from Boujraf et al. (2015) and Zhanlav and Mijiddorj (2010), the approximation order of these two integro splines is equal and is $O(h^4)$. The difference between these two splines is the construction of algorithms.

It should be mentioned that numerical results given in the work (Zhanlav and Mijiddorj 2010) show that the maximum errors are obtained near the end points as a consequence of using recurrence relations unsuitable for finding boundary coefficients in B -spline representation. In this note, we develop a new approach to completely construct explicit formulae

Communicated by Luz de Teresa.

R. Mijiddorj
mijiddorj@msue.edu.mn

¹ Institute of Mathematics, National University of Mongolia, Ulaanbaatar, Mongolia

² Department of Informatics, Mongolian National University of Education, Ulaanbaatar, Mongolia

for coefficients, which eliminates above-mentioned errors and improves the computational efficiency of the algorithm to construct the local integro cubic spline. For the sake of completeness, in Sect. 2, we develop some constructions of local cubic splines, and we show that these splines are equivalent. In Sect. 3, we derive explicit formulae for the local integro cubic spline and its derivatives and make comparative analysis of two existing cubic splines. For numerical tests, several examples are given in Sect. 4 to show the efficiency of the approach and to illustrate the theoretical results.

2 Local construction of cubic splines

We will use its B -representation of cubic splines of class $C^2[a, b]$ on the nonuniform partition:

$$\Delta_N = \{a = x_0 < x_1 < \cdots < x_N = b\}, \quad h_i = x_{i+1} - x_i, \quad i = 0(1), N - 1.$$

To this end, we extend the partition Δ_N by knots:

$$x_{-3} \leq x_{-2} \leq x_{-1} \leq x_0 < x_1 < \cdots < x_N \leq x_{N+1} \leq x_{N+2} \leq x_{N+3}.$$

We set

$$h_{-3} = h_{-2} = h_{-1} = wh_0, \quad h_N = h_{N+1} = h_{N+2} = wh_{N-1}, \quad w \geq 0.$$

Then, any cubic C^2 spline $S(x)$ is represented as

$$S(x) = \sum_{j=-1}^{N+1} \mu_j B_j(x), \quad (2.1)$$

where $B_j(x)$ are normalized cubic B -splines (Zhanlav and Mijiddorj 2010) that form a basis in space of cubic splines $S \in C^2[a, b]$. The support of B_j is $\text{supp}(B_j(x)) = [x_{j-2}, x_{j+2}]$. The coefficients in (2.1) are defined as (see Zhanlav 1981)

$$\begin{aligned} \mu_{-1} &= S_0 - h_0 w m_0 + \frac{h_0^2 w^2}{3} \mathcal{M}_0, \\ \mu_i &= S_i + \frac{h_i - h_{i-1}}{3} m_i - \frac{h_i h_{i-1}}{6} \mathcal{M}_i, \quad i = 0(1)N, \\ \mu_{N+1} &= S_N - h_{N-1} w m_N + \frac{h_{N-1}^2 w^2}{3} \mathcal{M}_N, \end{aligned} \quad (2.2)$$

where $S_i = S(x_i)$, $m_i = S'(x_i)$, and $\mathcal{M}_i = S''(x_i)$. We consider the quasi-interpolatory cubic spline operator:

$$\mathcal{Q}f := \sum_{j=-1}^{N+1} \mu_j(f) B_j, \quad (2.3)$$

where the coefficients are given by the formulae (see Zhanlav 1981):

$$\begin{aligned} \mu_{-1}(f) &= f_0 - h_0 w f'_0 + \frac{h_0^2 w^2}{3} f''_0, \\ \mu_i(f) &= f_i + \frac{h_i - h_{i-1}}{3} f'_i - \frac{h_i h_{i-1}}{6} f''_i, \quad i = 0(1)N, \\ \mu_{N+1}(f) &= f_N - h_{N-1} w f'_N + \frac{h_{N-1}^2 w^2}{3} f''_N. \end{aligned} \quad (2.4)$$

We are able to construct local cubic splines based on (2.4). Now let us consider the uniform partition case. Using approximate formulae

$$\begin{aligned}
 f'_0 &= \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3) + O(h^3), \\
 f''_0 &= \frac{1}{h^2}(2f_0 - 5f_1 + 4f_2 - f_3) + O(h^2), \\
 f''_i &= \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} + O(h^2), \quad i = 1(1)N - 1, \\
 f''_N &= \frac{1}{h^2}(2f_N - 5f_{N-1} + 4f_{N-2} - f_{N-3}) + O(h^2), \\
 f'_N &= \frac{1}{6h}(11f_N - 18f_{N-1} + 9f_{N-2} - 2f_{N-3}) + O(h^3),
 \end{aligned}
 \tag{2.5}$$

in (2.4), we obtain fully local cubic splines (2.3) with coefficients given by

$$\begin{aligned}
 \mu_{-1}(f) &= \frac{1}{6}[(6 + 11w + 4w^2)f_0 - w(18 + 10w)f_1 + w(9 + 8w)f_2 - 2w(1 + w)f_3], \\
 \mu_0(f) &= \frac{1}{18}[(7 + 5w)f_0 + (18 - 3w)f_1 - (9 + 3w)f_2 + (2 + w)f_3], \\
 \mu_i(f) &= \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1(1)N - 1, \\
 \mu_N(f) &= \frac{1}{18}[(7 + 5w)f_N + (18 - 3w)f_{N-1} - (9 + 3w)f_{N-2} + (2 + w)f_{N-3}], \\
 \mu_{N+1}(f) &= \frac{1}{6}[(6 + 11w + 4w^2)f_N - w(18 + 10w)f_{N-1} \\
 &\quad + w(9 + 8w)f_{N-2} - 2w(1 + w)f_{N-3}].
 \end{aligned}
 \tag{2.6}$$

Of course, these local cubic splines possess the same approximation properties as the interpolating ones. We consider some particular cases. Let $w = 0$ that corresponds to the multiple knots at the end points. Then, (2.6) leads to

$$\begin{aligned}
 \mu_{-1}(f) &= f_0, \quad \mu_0(f) = \frac{1}{18}[7f_0 + 18f_1 - 9f_2 + 2f_3], \\
 \mu_i(f) &= \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1(1)N - 1, \\
 \mu_N(f) &= \frac{1}{18}[7f_N + 18f_{N-1} - 9f_{N-2} + 2f_{N-3}], \quad \mu_{N+1}(f) = f_N.
 \end{aligned}
 \tag{2.7}$$

Thus, we obtain the cubic discrete spline quasi-interpolant $\mathcal{Q}_3 f$ with coefficients (2.7) given in Boujraf et al. (2015), Sablonnière (2005). Now, we will find $\mathcal{Q}_3 f(x_i)$, $\mathcal{Q}'_3 f(x_i)$, and $\mathcal{Q}''_3 f(x_i)$ for $i = 0, 1, N - 1$, and N . To this end, we use the following relations (Zhanlav 1981):

$$\begin{aligned}
 \mu_{k-1} &= S_k - \frac{2h_{k-1} + h_{k-2}}{3}m_k + \frac{h_{k-1}(h_{k-1} + h_{k-2})}{6}\mathcal{M}_k, \\
 \mu_k &= S_k + \frac{h_k - h_{k-1}}{3}m_k - \frac{h_k h_{k-1}}{6}\mathcal{M}_k, \\
 \mu_{k+1} &= S_k + \frac{2h_k + h_{k+1}}{3}m_k + \frac{h_k(h_k + h_{k+1})}{6}\mathcal{M}_k,
 \end{aligned}
 \tag{2.8}$$

which are valid for $k = 0(1)N$ and any cubic spline of class $C^2[a, b]$, for instance, for the cubic quasi-interpolant $\mathcal{Q}_3 f$ ($\mathcal{S}_i = \mathcal{Q}_3 f(x_i)$, $m_i = \mathcal{Q}'_3 f(x_i)$, $\mathcal{M}_i = \mathcal{Q}''_3 f(x_i)$). Setting $k = 0$ and $w = 0$ in (2.8), we have

$$\begin{aligned}\mu_{-1}(f) &= \mathcal{Q}_3 f(x_0), \quad \mu_0(f) = \mathcal{Q}_3 f(x_0) + \frac{h}{3} \mathcal{Q}'_3 f(x_0), \\ \mu_1(f) &= \mathcal{Q}_3 f(x_0) + h \mathcal{Q}'_3 f(x_0) + \frac{h^2}{3} \mathcal{Q}''_3 f(x_0).\end{aligned}$$

From this, we find that

$$\begin{aligned}\mathcal{Q}'_3 f(x_0) &= \frac{3}{h}(\mu_0(f) - \mu_{-1}(f)), \\ \mathcal{Q}_3 f(x_0) &= \mu_0(f) - \frac{h}{3} \mathcal{Q}'_3 f(x_0), \\ \mathcal{Q}''_3 f(x_0) &= \frac{3}{h^2}(\mu_1(f) - \mathcal{Q}_3 f(x_0) - h \mathcal{Q}'_3 f(x_0)).\end{aligned}\tag{2.9}$$

Substituting $\mu_{-1}(f)$, $\mu_0(f)$, and $\mu_1(f)$ from (2.7) into (2.9), we get

$$\begin{aligned}\mathcal{Q}_3 f(x_0) &= f_0, \\ \mathcal{Q}'_3 f(x_0) &= \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3), \\ \mathcal{Q}''_3 f(x_0) &= \frac{1}{h^2}(2f_0 - 5f_1 + 4f_2 - f_3).\end{aligned}\tag{2.10}$$

Analogously, setting $k = 1$ and $w = 0$ in (2.8), and after some calculations, we find that

$$\begin{aligned}\mathcal{Q}_3 f(x_1) &= \frac{3\mu_0(f) + 7\mu_1(f) + 2\mu_2(f)}{12}, \\ \mathcal{Q}'_3 f(x_1) &= \frac{-3\mu_0 + \mu_1 + 2\mu_2(f)}{4h}, \\ \mathcal{Q}''_3 f(x_1) &= \frac{6}{h^2}(\mathcal{Q}_3 f(x_1) - \mu_1(f)).\end{aligned}\tag{2.11}$$

Substituting $\mu_0(f)$, $\mu_1(f)$, and $\mu_2(f)$ from (2.7) into (2.11), we get

$$\begin{aligned}\mathcal{Q}_3 f(x_1) &= f_1, \\ \mathcal{Q}'_3 f(x_1) &= \frac{1}{6h}(-2f_0 - 3f_1 + 6f_2 - f_3), \\ \mathcal{Q}''_3 f(x_1) &= \frac{1}{h^2}(f_0 - 2f_1 + f_2).\end{aligned}\tag{2.12}$$

Similarly, we get

$$\begin{aligned}\mathcal{Q}_3 f(x_{N-1}) &= f_{N-1}, \quad \mathcal{Q}_3 f(x_N) = f_N, \\ \mathcal{Q}'_3 f(x_{N-1}) &= \frac{2f_N + 3f_{N-1} - 6f_{N-2} + f_{N-3}}{6h}, \\ \mathcal{Q}'_3 f(x_N) &= \frac{11f_N - 18f_{N-1} + 9f_{N-2} - 2f_{N-3}}{6h}, \\ \mathcal{Q}''_3 f(x_{N-1}) &= \frac{f_N - 2f_{N-1} + f_{N-2}}{h^2}, \\ \mathcal{Q}''_3 f(x_N) &= \frac{2f_N - 5f_{N-1} + 4f_{N-2} - f_{N-3}}{h^2}.\end{aligned}\tag{2.13}$$

Let $w = 1$ in (2.6). Then from (2.6) we obtain another local cubic spline $\tilde{Q}_3(x)$ with coefficients given by

$$\begin{aligned} \tilde{\mu}_{-1}(f) &= \frac{1}{6}[21f_0 - 28f_1 + 17f_2 - 4f_3], \\ \tilde{\mu}_0(f) &= \frac{1}{6}[4f_0 + 5f_1 - 4f_2 + f_3], \\ \tilde{\mu}_i(f) &= \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1(1)N - 1, \\ \tilde{\mu}_N(f) &= \frac{1}{6}[4f_N + 5f_{N-1} - 4f_{N-2} + f_{N-3}], \\ \tilde{\mu}_{N+1}(f) &= \frac{1}{6}[21f_N - 28f_{N-1} + 17f_{N-2} - 4f_{N-3}]. \end{aligned} \tag{2.14}$$

Hence, we have two local cubic splines. They differ from each other only by the bordered coefficients in (2.7) and (2.14). Note that both these local cubic splines are exact on the space of polynomials of degree at most 3. Now, we show that they coincide with each other, although their representations are different. Using the properties of B -spline, the direct calculations give us

$$\tilde{Q}_3 f(x_i) = f_i, \quad i = 0, 1, N - 1, N. \tag{2.15}$$

From (2.7), (2.10), and (2.12)–(2.15), we conclude that

$$Q_3 f(x_i) = \tilde{Q}_3 f(x_i), \quad Q'_3 f(x_i) = \tilde{Q}'_3 f(x_i), \quad Q''_3 f(x_i) = \tilde{Q}''_3 f(x_i), \quad i = 0(1)N.$$

Thus, we prove that local cubic splines (2.3), (2.7), and (2.3), (2.14) coincide. Another approach to construct local cubic splines is to use a well-known not-a-knot end conditions (Behforooz 2006; Zhanlav 1984):

$$S'''_{i+0} = S'''_{i-0}, \quad i = 1, N - 1. \tag{2.16}$$

In Zhanlav (1984), it was shown that the conditions (2.16) are equivalent to

$$\mu_i = \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1, N - 1. \tag{2.17}$$

On the other hand, according to $w = 1$, and (2.2) and (2.16), we have

$$\mu_i = S_i - \frac{h^2}{6} S''_i = \frac{8S_i - S_{i-1} - S_{i+1}}{6}, \quad i = 1, N - 1. \tag{2.18}$$

From (2.17) and (2.18), it follows that

$$S_{i-1} - f_{i-1} - 8(S_i - f_i) + S_{i+1} - f_{i+1} = 0, \quad i = 1, N - 1. \tag{2.19}$$

Moreover, from (2.18), we get

$$\mu_{i-2} - 4\mu_{i-1} + 6\mu_i - 4\mu_{i+1} + \mu_{i+2} = 0, \quad i = 1, N - 1. \tag{2.20}$$

From (2.14) and (2.20), we find that

$$\begin{aligned}\hat{\mu}_{-1}(f) &= \frac{1}{6}[22f_0 - 32f_1 + 23f_2 - 8f_3 + f_4], \\ \hat{\mu}_0(f) &= \frac{1}{6}[4f_0 + 5f_1 - 4f_2 + f_3], \\ \hat{\mu}_i(f) &= \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1(1)N-1, \\ \hat{\mu}_N(f) &= \frac{1}{6}[4f_N + 5f_{N-1} - 4f_{N-2} + f_{N-3}], \\ \hat{\mu}_{N+1}(f) &= \frac{1}{6}[22f_N - 32f_{N-1} + 23f_{N-2} - 8f_{N-3} + f_{N-4}].\end{aligned}\tag{2.21}$$

The local cubic spline satisfying the end conditions (2.16) holds (2.19). If we use relations $f_{i-2} - 4f_{i-1} + 6f_i - 4f_{i+1} + f_{i+2} = 0$ valid for $f(x) \in C^4$, $i = 2(1)N-2$, then the coefficients in (2.21) lead to (2.14) within the accuracy of $O(h^4)$. This means that the last two local splines, $\tilde{Q}_3(x)$ and $\hat{Q}_3(x)$ with coefficients (2.21), are almost identical. Moreover, using (2.14) and $f \in C^4$, we easily show that

$$\tilde{Q}'_3 f(x_i) - f'_i = O(h^4), \quad i = 0(1)N.\tag{2.22}$$

Note that this super convergence property holds only for uniform partitions.

3 The local integro cubic spline

The construction of local cubic splines considered in the previous section is based on the local spline approximations. We similarly construct a local integro cubic spline and introduce a uniform partition on $[a, b]$, $x_i = a + ih$ for $i = 0, 1, \dots, N$ with $h = (b - a)/N$. Let $S(x)$ be a local integro cubic spline belonging to $C^2[a, b]$ and satisfying the following conditions (Behforooz 2006; Zhanlav and Mijiddorj 2010):

$$\int_{x_{i-1}}^{x_i} S(x)dx = \int_{x_{i-1}}^{x_i} u(x)dx = I_i, \quad i = 1(1)N.\tag{3.1}$$

We will use B -spline representation of $S(x)$:

$$S(x) = \sum_{j=-1}^{N+1} \alpha_j B_j(x),\tag{3.2}$$

where $B_j(x)$ are normalized cubic B -splines that form a basis for cubic splines from $C^2[a, b]$. According to the properties of B -spline, we have (Zhanlav 1981)

$$S_i = \frac{\alpha_{i+1} + 4\alpha_i + \alpha_{i-1}}{6},\tag{3.3a}$$

$$m_i = \frac{\alpha_{i+1} - \alpha_{i-1}}{2h}, \quad i = 0(1)N,\tag{3.3b}$$

$$M_i = \frac{\alpha_{i+1} - 2\alpha_i + \alpha_{i-1}}{h^2},\tag{3.3c}$$

where $S_i = S(x_i)$, $m_i = S'(x_i)$, and $M_i = S''(x_i)$. The integro cubic spline (3.2) satisfying the conditions (3.1) holds the following relations (Zhanlav and Mijiddorj 2010):

$$\alpha_{i-2} + 12\alpha_{i-1} + 22\alpha_i + 12\alpha_{i+1} + \alpha_{i+2} = \frac{24}{h}(I_i + I_{i+1}), \quad i = 1(1)N - 1. \quad (3.4)$$

In Zhanlav and Mijiddorj (2010), we obtained the explicit and approximate formulae:

$$\alpha_i = \frac{1}{6h}(-I_{i-1} + 4I_i + 4I_{i+1} - I_{i+2}), \quad i = 2(1)N - 2, \quad (3.5)$$

and

$$\alpha_{i-1} + \alpha_i + \alpha_{i+1} = \frac{3}{2h}(I_i + I_{i+1}), \quad i = 1(1)N - 1, \quad (3.6)$$

with accuracy of $O(h^4)$. The remainder coefficients in (3.2) are determined from (3.4)–(3.6) explicitly and we present the final results:

$$\begin{aligned} \alpha_{-1} &= \frac{1}{6h}(26I_1 - 23I_2 - 14I_3 + 26I_4 - 9I_5), \\ \alpha_0 &= \frac{1}{6h}(9I_1 - I_2 - 5I_3 + 4I_4 - I_5), \\ \alpha_1 &= \frac{1}{6h}(I_1 + 6I_2 + I_3 - 3I_4 + I_5), \end{aligned} \quad (3.7)$$

and

$$\begin{aligned} \alpha_{N-1} &= \frac{1}{6h}(I_N + 6I_{N-1} + I_{N-2} - 3I_{N-3} + I_{N-4}), \\ \alpha_N &= \frac{1}{6h}(9I_N - I_{N-1} - 5I_{N-2} + 4I_{N-3} - I_{N-4}), \\ \alpha_{N+1} &= \frac{1}{6h}(26I_N - 23I_{N-1} - 14I_{N-2} + 26I_{N-3} - 9I_{N-4}). \end{aligned} \quad (3.8)$$

Thus, we have the local integro cubic spline (3.2), all the coefficients of which are completely determined by the explicit formulae (3.5), (3.7), and (3.8).

For the sake of comparison, we also present here the integro cubic spline quasi-interpolant, a modified quasi-interpolant of the $Q_3 f$, given in Boujraf et al. (2015)

$$\tilde{Q}_3 f = \sum_{i=1}^{N+3} v_i(f) B_i^3, \quad (3.9)$$

where the coefficients $v_i(f)$ are defined as follows:

$$\begin{aligned} v_1(f) &= \frac{1}{12h}(25I_1 - 23I_2 + 13I_3 - 3I_4), \\ v_2(f) &= \frac{1}{108h}(119I_1 + 4I_2 - 24I_3 + 10I_4 - I_5), \\ v_3(f) &= \frac{1}{6h}(10I_2 - 5I_3 + I_4), \\ v_4(f) &= \frac{1}{72h}(-11I_1 + 44I_2 + 54I_3 - 16I_4 + I_5), \\ v_i(f) &= \frac{1}{72h}(I_{i-4} - 15I_{i-3} + 50I_{i-2} + 50I_{i-1} - 15I_i + I_{i+1}), \\ &5 \leq i \leq N - 1, \end{aligned}$$

Table 1 Comparison of number of arithmetic operations for S_i and $\tilde{Q}_3 f_i$

Splines	Total oper.	*	+
$S_i, i = 0(1)N$	$5N+15$	$2N+12$	$3N+3$
$\tilde{Q}_3 f_i, i = 0(1)N$	$12N+26$	$5N+23$	$7N+3$

$$\begin{aligned}
 v_N(f) &= \frac{1}{72h}(I_{N-4} - 16I_{N-3} + 54I_{N-2} + 44I_{N-1} - 11I_N), \\
 v_{N+1}(f) &= \frac{1}{6h}(I_{N-3} - 5I_{N-2} + 10I_{N-1}), \\
 v_{N+2}(f) &= \frac{1}{108h}(-I_{N-4} + 10I_{N-3} - 24I_{N-2} + 4I_{N-1} + 119I_N), \\
 v_{N+3}(f) &= \frac{1}{12h}(-3I_{N-3} + 13I_{N-2} - 23I_{N-1} + 25I_N).
 \end{aligned}
 \tag{3.10}$$

Note that the support of B_i^3 is $supp(B_i^3) = [x_{i-4}, x_i]$, whereas $supp(B_j) = [x_{i-2}, x_{i+2}]$. Certainly, we can use (3.10) to approximate the boundary values $\tilde{Q}_3 f(x_i)$ and its first and second derivatives, $i = 0, 1, N - 1, N$. In particular, we have

$$\begin{aligned}
 \tilde{Q}'_3 f(x_0) &= \frac{-35I_1 + 69I_2 - 45I_3 + 11I_4}{12h^2}, \\
 \tilde{Q}'_3 f(x_1) &= \frac{-11I_1 + 9I_2 + 3I_3 - I_4}{12h^2}, \\
 \tilde{Q}'_3 f(x_{N-1}) &= \frac{11I_N - 9I_{N-1} - 3I_{N-2} + I_{N-3}}{12h^2}, \\
 \tilde{Q}'_3 f(x_N) &= \frac{35I_N - 69I_{N-1} + 45I_{N-2} - 11I_{N-3}}{12h^2}.
 \end{aligned}
 \tag{3.11}$$

Let us introduce the definition.

Definition 1 The local integro splines are called m -point ones if the values of S_i are expressed by linear combination of I_j at m -adjacent knots in a neighborhood of x_i .

According to this definition and to formulae (3.5) and (3.10), the local integro spline (3.2) is 6-point one, whereas $\tilde{Q}_3 f$ is 8-point one. That is, $S(x)$ is more compact than $\tilde{Q}_3 f$. As a consequence, S_i requires less computational cost than that of $\tilde{Q}_3 f$ (see Table 1). Moreover, formulae (3.7) and (3.8) can be further simplified using the following formulae:

$$I_{i-2} - 4I_{i-1} + 6I_i - 4I_{i+1} + I_{i+2} = O(h^5), \quad i = 3(1)N - 2,
 \tag{3.12}$$

that are valid for any function $u \in C^4$. Thus, we have

$$\begin{aligned}
 \alpha_{-1} &= \frac{1}{6h}(35I_1 - 59I_2 + 40I_3 - 10I_4), \\
 \alpha_i &= \frac{1}{6h}(10I_{i+1} - 5I_{i+2} + I_{i+3}), \quad i = 0, 1, \\
 \alpha_i &= \frac{1}{6h}(-I_{i-1} + 4I_i + 4I_{i+1} - I_{i+2}), \quad i = 2(1)N - 2, \\
 \alpha_i &= \frac{1}{6h}(10I_i - 5I_{i-1} + I_{i-2}), \quad i = N - 1, N, \\
 \alpha_{N+1} &= \frac{1}{6h}(35I_N - 59I_{N-1} + 40I_{N-2} - 10I_{N-3}).
 \end{aligned}
 \tag{3.13}$$

The formulae (3.13) are more compact and uniform compared to (3.10). The values of S_i , m_i , and M_i are defined by (3.3) using (3.13). Moreover, by virtue of (3.12) within the accuracy of $O(h^4)$ the values of S_i can be determined explicitly as

$$\begin{aligned}
 S_0 &= \frac{1}{12h}(25I_1 - 23I_2 + 13I_3 - 3I_4), \\
 S_1 &= \frac{1}{12h}(3I_1 + 13I_2 - 5I_3 + I_4), \\
 S_i &= \frac{1}{12h}(-I_{i-1} + 7I_i + 7I_{i+1} - I_{i+2}), \quad i = 2(1)N - 2, \\
 S_{N-1} &= \frac{1}{12h}(3I_N + 13I_{N-1} - 5I_{N-2} + I_{N-3}), \\
 S_N &= \frac{1}{12h}(25I_N - 23I_{N-1} + 13I_{N-2} - 3I_{N-3}).
 \end{aligned} \tag{3.14}$$

From (3.14), we see that $S_i = \tilde{f}_i$, $i = 0(1)N$, where \tilde{f}_i are given by (4)–(8) in Boujraf et al. (2015). In a similar way, we obtain

$$\begin{aligned}
 m_0 &= \frac{1}{12h^2}(-45I_1 + 109I_2 - 105I_3 + 51I_4 - 10I_5), \\
 m_1 &= \frac{1}{12h^2}(-10I_1 + 5I_2 + 9I_3 - 5I_4 + I_5), \\
 m_i &= \frac{1}{12h^2}(I_{i-1} + 15(I_{i+1} - I_i) - I_{i+2}), \quad i = 2(1)N - 2, \\
 m_{N-1} &= \frac{1}{12h^2}(10I_N - 5I_{N-1} - 9I_{N-2} + 5I_{N-3} - I_{N-4}), \\
 m_N &= \frac{1}{12h^2}(45I_N - 109I_{N-1} + 105I_{N-2} - 51I_{N-3} + 10I_{N-4}),
 \end{aligned} \tag{3.15}$$

and

$$\begin{aligned}
 M_0 &= \frac{1}{2h^3}(5I_1 - 13I_2 + 11I_3 - 3I_4), \\
 M_1 &= \frac{1}{2h^3}(3I_1 - 7I_2 + 5I_3 - I_4), \\
 M_i &= \frac{1}{2h^3}(I_{i-1} - I_i - I_{i+1} + I_{i+2}), \quad i = 2(1)N - 2, \\
 M_{N-1} &= \frac{1}{2h^3}(3I_N - 7I_{N-1} + 5I_{N-2} - I_{N-3}), \\
 M_N &= \frac{1}{2h^3}(5I_N - 13I_{N-1} + 11I_{N-2} - 3I_{N-3}).
 \end{aligned} \tag{3.16}$$

Note that formulae (3.15) and (3.16) are valid within accuracy of $O(h^4)$ and $O(h^2)$, respectively. The formulae (3.14)–(3.16) show that, in fact, we proceed from 6-point approximation to 4-point one except for m_i for $i = 0, 1, N - 1, N$ without loss of accuracy. It is easy to show that for $f \in C^4$, the error order of the first derivative of the local integro cubic spline defined by (3.2) and (3.13) equals to $O(h^4)$ as (2.22), whereas this error order does not hold for (3.9) and (3.10), because (3.11) have accuracy of $O(h^3)$. This is another advantage of our construction.

Table 2 Results obtained for the functions f and g

N	$f(x)$				$g(x)$			
	$\ S_i - f_i\ _{\infty, N}$	NCO	$\ \tilde{Q}_3 f_i - f_i\ _{\infty, N}$	NCO	$\ S_i - g_i\ _{\infty, N}$	NCO	$\ \tilde{Q}_3 g_i - g_i\ _{\infty, N}$	NCO
8	1.080×10^{-04}	–	1.080×10^{-04}	–	3.465×10^{-05}	–	3.465×10^{-05}	–
16	7.479×10^{-06}	3.85	7.479×10^{-06}	3.85	2.381×10^{-06}	3.86	2.381×10^{-06}	3.86
32	4.922×10^{-07}	3.92	4.922×10^{-07}	3.92	1.548×10^{-07}	3.94	1.548×10^{-07}	3.94
64	3.157×10^{-08}	3.96	3.157×10^{-08}	3.96	9.859×10^{-09}	3.97	9.859×10^{-09}	3.97
128	1.999×10^{-09}	3.98	1.999×10^{-09}	3.98	6.216×10^{-10}	3.98	6.216×10^{-10}	3.98
256	1.257×10^{-10}	3.99	1.256×10^{-10}	3.99	3.901×10^{-11}	3.99	3.897×10^{-11}	3.99

Table 3 Results obtained for the functions f' and g'

N	$f'(x)$				$g'(x)$			
	$\ S'_i - f'_i\ _{\infty, N}$	NCO	$\ \tilde{Q}'_3 f_i - f'_i\ _{\infty, N}$	NCO	$\ S'_i - g'_i\ _{\infty, N}$	NCO	$\ \tilde{Q}'_3 g_i - g'_i\ _{\infty, N}$	NCO
8	3.907×10^{-04}	–	3.529×10^{-03}	–	1.789×10^{-04}	–	1.131×10^{-03}	–
16	2.774×10^{-05}	3.81	4.853×10^{-04}	2.86	1.150×10^{-05}	3.95	1.543×10^{-04}	2.87
32	1.849×10^{-06}	3.90	6.366×10^{-05}	2.93	7.241×10^{-07}	3.98	2.001×10^{-05}	2.94
64	1.193×10^{-07}	3.95	8.152×10^{-06}	2.96	4.533×10^{-08}	3.99	2.544×10^{-06}	2.97
128	7.584×10^{-09}	3.97	1.031×10^{-06}	2.98	2.834×10^{-09}	3.99	3.206×10^{-07}	2.98
256	4.795×10^{-10}	3.98	1.296×10^{-07}	2.99	1.772×10^{-10}	3.99	4.020×10^{-08}	2.99

4 Numerical examples

The number of the required arithmetic operations to calculate S_i , $i = 0(1)N$ is 2.4 times less than that of $\tilde{Q}_3 f_i = \tilde{Q}_3 f(x_i)$ (see Table 1). We construct the integro cubic splines for functions $f(x) = \exp(x)$, and $g(x) = \sin(x)$, $x \in [0, 1]$, and calculate the maximum norm for different values of $N = 8, 16, 32, \dots, 256$ as in Boujraf et al. (2015). The numerical results for S_i and $\tilde{Q}_3 f_i$ obtained using both integrosplines appear to agree amazingly well (see Table 2).

Finally, we compare the S' with the quasi-interpolant $\tilde{Q}'_3 f$ constructed in Boujraf et al. (2015). The obtained results are listed in Table 3, where $\tilde{Q}'_3 f_i = \tilde{Q}'_3 f(x_i)$ and $\tilde{Q}''_3 f_i = \tilde{Q}''_3 f(x_i)$. In Table 2, 3 and 4, numerical convergence order is denoted by NCO, maximum absolute errors by $R(N) = \|\cdot\|_{\infty, N} = \max_{0 \leq i \leq N} |\cdot|$, where $\text{NCO} = \log_2 \left| \frac{R(N)}{R(2N)} \right|$. From the Table 3, we can see that the results obtained by our approach are better than those obtained by the quasi-interpolant introduced in Boujraf et al. (2015).

Conclusion

We completely obtain local cubic and integrocubic splines and make some comparisons. The algorithm to construct the local integro cubic spline is easy to implement and requires less computational cost than that of $\tilde{Q}_3 f$ (see Table 1). The approximation orders of the local integro cubic and quasi-interpolant splines as well as their second derivatives equal to four

Table 4 Results obtained for the functions f'' and g''

N	$f''(x)$				$g''(x)$			
	$\ S_i'' - f_i''\ _{\infty, N}$	NCO	$\ \tilde{Q}_3'' f_i - f_i''\ _{\infty, N}$	NCO	$\ S_i'' - g_i''\ _{\infty, N}$	NCO	$\ \tilde{Q}_3'' g_i - g_i''\ _{\infty, N}$	NCO
8	6.23×10^{-02}	–	3.10×10^{-02}	–	1.99×10^{-02}	–	5.823×10^{-03}	–
16	1.70×10^{-02}	1.87	7.09×10^{-03}	2.12	5.39×10^{-03}	1.88	1.495×10^{-03}	1.96
32	4.44×10^{-03}	1.93	1.69×10^{-03}	2.06	1.39×10^{-03}	1.95	3.786×10^{-04}	1.98
64	1.13×10^{-03}	1.96	4.15×10^{-04}	2.03	3.54×10^{-04}	1.97	9.526×10^{-05}	1.99
128	2.87×10^{-04}	1.98	1.02×10^{-04}	2.01	8.92×10^{-05}	1.98	2.389×10^{-05}	1.99
256	7.21×10^{-05}	1.99	2.55×10^{-05}	2.00	2.23×10^{-05}	1.99	5.986×10^{-06}	1.99

and two, respectively, whereas the first derivative approximation order of our local integro cubic spline is better than that of $\tilde{Q}_3 f$ (see Table 3). Hence, the usage of the local integro cubic spline defined by formula (3.2) and (3.13) is more suitable for applications. If we need to use $\tilde{Q}_3 f$, then (3.10) should be simplified by (3.12).

Acknowledgements The work was partially supported by Foundation of Science and Technology of Mongolia (no. SST_007/2015). The authors are grateful to the reviewer for the valuable comments and useful suggestions that greatly improved the quality of the manuscript.

References

Behforooz H (2006) Approximation by integro cubic splines. *Appl Math Comput* 175:8–15

Boujraf A, Sbibi D, Tahrichi M, Tijini A (2015) A simple method for constructing integro spline quasi-interpolants. *Math Comput Simul* 111:36–47

Sablonnière P (2005) Univariate spline quasi-interpolants and applications to numerical analysis. *Rend Sem Mat Univ Pol Torino* 63:107–118

Zhanlav T (1981) B-representation of interpolatory cubic splines. *Vychislitel'nye Syst Novosibirsk* 87:3–10

Zhanlav T (1984) End conditions for interpolatory cubic splines. *Vychislitel'nye Syst Novosibirsk* 106:25–28

Zhanlav T, Mijidorj R (2010) The local integro cubic splines and their approximation properties. *Appl Math Comput* 216:2215–2219

On the approximation of inverse of some band matrices and their applications in local splines

T. Zhanlav¹, R. Mijiddorj^{2*}, H. Behforooz³

(1) *Institute of Mathematics, National University of Mongolia, Ulaanbaatar, Mongolia.*

(2) *Department of Informatics, Mongolian National University of Education, Ulaanbaatar, Mongolia.*

(3) *Department of Mathematics, Utica College, Utica, NY 13502, USA.*

Copyright 2018 © T. Zhanlav, R. Mijiddorj and H. Behforooz. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

In this paper, we obtain approximate inverses of popular tri-diagonal and penta-diagonal matrices which are used to construct local (or a discrete quasi-interpolant) interpolatory and integro splines.

Keywords: Tri-diagonal matrices; Penta-diagonal matrices; Inverses; Local construction; Splines.

1 Introduction

The band matrices often arise in a range of science and engineering applications such as numerical solutions of ordinary and partial differential equations, spline approximation, image and signal processing, and parallel computing, see [1, 5, 6, 11] and references therein. In many of these areas, inversion of the tri-diagonal matrix is required. In particular, in [11] Yamamoto obtained explicit formulas for the entries of the inverse of nonsingular tri-diagonal matrices. In [7], Jia and Li derived the numerical or symbolic algorithms for the inverses of k -diagonal matrices. Moreover, in [9] Smolarski discussed a particular type of banded matrix, namely a diagonally striped matrix, and the structure of its inverse. Bickel and Lindner in [3] proved that if an infinite matrix \mathbf{A} , which is invertible as a bounded operator on l^2 , can be uniformly approximated by banded matrices then so can the inverse of \mathbf{A} .

Although there are explicit formulas for entries of the inverse of band matrices but most of the time, practically, they are not suitable for simple and hand calculations. In some cases, it suffices to find only approximate inverse of these matrices. On the other hand, for band matrices, it is well established that the entries of its inverse decay exponentially away from the main diagonal; see for example [4]. Therefore, we only need to find approximate entries α_{ij} of the main and its few adjacent diagonals i.e. we need

$$\alpha_{ij}, \quad \text{for } |i-j| \leq k, \quad \text{for } k = 1, 2, 3. \quad (1.1)$$

For example, in constructing interpolatory splines and integro-splines with small degrees, it is often required to solve a system of linear equations

$$\mathbf{Ax} = \mathbf{f}, \quad (1.2)$$

where \mathbf{A} is a band matrix. In particular, we consider the following cases:

*Corresponding author. Email address: mijiddorj@msue.edu.mn; Tel.: +97699010363

1. $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$,
2. $\mathbf{A} = \text{Tri-diag}\{1, 10, 1\}$,
3. $\mathbf{A} = \text{Tri-diag}\{1, d, 1\}$, $|d| > 2$,
4. $\mathbf{A} = \text{Penta-diag}\{1, 26, 66, 26, 1\}$,
5. $\mathbf{A} = \text{Penta-diag}\{1, 56, 246, 56, 1\}$.

If we have approximate inverse $\mathbf{A}^{-1} = (\alpha_{ij})$, $|i - j| \leq k$ then we obtain the approximate solution of (1.2) as follows:

$$\tilde{x}_i = \sum_{|i-j| \leq k} \alpha_{ij} f_j. \quad (1.3)$$

The error of approximate solution given by (1.3) is estimated as

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \max_{|i-j| > k} |\alpha_{ij}| \|\mathbf{f}\|. \quad (1.4)$$

From (1.4) it is clear that it is better to restrict k by small values, because of the exponential decay of entries of inverse of band matrices [4]. To find approximate inverse $\mathbf{A}^{-1} = (\alpha_{ij})$ we use the approximate solution of system (1.2) known in some cases. First we consider the system (1.2) with matrix $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$. Such system arises in constructing interpolatory cubic spline on the uniform partition $[a, b]$ with knots $x_i = a + ih$, $i = 0(1)n$, $h = \frac{b-a}{n}$.

2 Approximate inverse of special tri-diagonal matrices and its applications

Let $S_3(x)$ be a cubic C^2 spline satisfying the interpolation conditions

$$S_3(x_i) = f_i, \quad f_i = f(x_i), \quad i = 0(1)n. \quad (2.5)$$

By using the B -spline representation of $S_3(x) \in C^2$, we have:

$$S_3(x) = \sum_{i=-1}^{n+1} c_i B_i(x), \quad (2.6)$$

where $B_i(x)$ are normalized cubic B -splines that constitute basis for $S_3 \in C^2[a, b]$ cubic splines space, see [12]. Then the interpolatory conditions (2.5) implies

$$c_{i-1} + 4c_i + c_{i+1} = 6f_i, \quad i = 0(1)n, \quad (2.7a)$$

or

$$\mathbf{c} = 6\mathbf{A}^{-1}\mathbf{f}. \quad (2.7b)$$

Now we will find the entries of near of the main diagonal, using the approximate explicit formula given in [8, 12]

$$c_i = \frac{8f_i - f_{i-1} - f_{i+1}}{6}, \quad i = 1(1)n-1, \quad (2.8)$$

with accuracy $O(h^4)$. Using (2.8), we write (2.7b) as

$$8f_i - f_{i-1} - f_{i+1} + O(h^4) = 36\{\dots + \alpha_{i,i-2}f_{i-2} + \alpha_{i,i-1}f_{i-1} + \alpha_{ii}f_i + \alpha_{i,i+1}f_{i+1} + \alpha_{i,i+2}f_{i+2}\dots\}.$$

If we use explicit formulas given in [11] for matrix $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$ then it is easy to show that

$$\alpha_{ij} = \alpha_{ji}, \quad \text{and } \alpha_{i,i+j} = \alpha_{i,i-j} \text{ for } j = 1, 2, \dots. \quad (2.9)$$

Therefore, the last expression can be rewritten as

$$8f_i - f_{i-1} - f_{i+1} + O(h^4) = 36\{\dots + \alpha_{i,i-2}(f_{i-2} + f_{i+2}) + \alpha_{i,i-1}(f_{i-1} + f_{i+1}) + \alpha_{ii}f_i\}. \quad (2.10)$$

If we take into account the formula

$$f_{i-2} - 4f_{i-1} + 6f_i - 4f_{i+1} + f_{i+2} = O(h^4),$$

which holds for $f(x) \in C^4$, then the expression (2.10) becomes

$$8f_i - f_{i-1} - f_{i+1} = 36\{\dots + (4\alpha_{i,i-2} + \alpha_{i,i-1})(f_{i-1} + f_{i+1}) + (\alpha_{ii} - 6\alpha_{i,i-2})f_i\} + O(h^4). \quad (2.11)$$

It follows from (2.11) that

$$\begin{aligned} 36(4\alpha_{i,i-2} + \alpha_{i,i-1}) &= -1, \\ 36(\alpha_{ii} - 6\alpha_{i,i-2}) &= 8, \end{aligned} \quad (2.12)$$

where $\alpha_{ij} = O(h^4)$, $|i - j| \geq 3$. From (2.12) the unknowns $\alpha_{i,i-1}$ and α_{ii} are expressed by $\alpha_{i,i-2}$ as

$$\alpha_{i,i-1} = -\frac{1}{36} - 4\alpha_{i,i-2}, \quad \alpha_{ii} = \frac{2}{9} + 6\alpha_{i,i-2}. \quad (2.13)$$

We know that

$$(\mathbf{A}\mathbf{A}^{-1})_{ii} = 1, \quad (2.14)$$

which leads to

$$\alpha_{i,i-2} = \frac{1}{96}. \quad (2.15a)$$

Hence, from (2.13) we find

$$\alpha_{i,i-1} = -\frac{5}{72}, \quad \alpha_{ii} = \frac{41}{144}, \quad (2.15b)$$

and

$$\alpha_{ij} = O(h^4), \quad |i - j| \geq 3. \quad (2.15c)$$

Thus, the entries of i -th row of \mathbf{A}^{-1} are given by explicit formula (2.15). Further, if we use the notation $M_i = S_3''(x_i)$ then we have the following system of equations [12]

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(f_{i-1} - 2f_i + f_{i+1}), \quad i = 1(1)n - 1. \quad (2.16)$$

The matrix of system (2.16) is, as preceding case, $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$. Then according to (2.13) we have

$$\begin{aligned} M_i &= \frac{6}{h^2}\{\dots + \alpha_{i,i-2}(f_{i-3} - 2f_{i-2} + f_{i-1} + f_{i+1} - 2f_{i+2} + f_{i+3}) \\ &\quad + (\frac{1}{36} - 4\alpha_{i,i-2})(f_{i-2} - 2f_{i-1} + 2f_i - 2f_{i+1} + f_{i+2}) + (\frac{2}{9} + 6\alpha_{i,i-2})(f_{i-1} - 2f_i + f_{i+1})\} \\ &= \frac{1}{6h^2}\{-f_{i-2} + 10f_{i-1} - 18f_i + 10f_{i+1} - f_{i+2}\} \\ &\quad + \frac{6}{h^2}\alpha_{i,i-2}\{f_{i-3} - 6f_{i-2} + 15f_{i-1} - 20f_i + 15f_{i+1} - 6f_{i+2} + f_{i+3}\} + O(h^2). \end{aligned}$$

Let $f \in C^6[a, b]$. The Taylor expansions of $f(x_i + kh)$ give us

$$f_{i+k} = f_i + khf'_i + \frac{(kh)^2}{2}f''_i + \frac{(kh)^3}{6}f^{(3)}_i + \frac{(kh)^4}{24}f^{(4)}_i + \frac{(kh)^5}{120}f^{(5)}_i + O(h^6), \quad k = \pm 1, \pm 2, \pm 3,$$

from these we have

$$f_{i-3} - 6f_{i-2} + 15f_{i-1} - 20f_i + 15f_{i+1} - 6f_{i+2} + f_{i+3} = O(h^6).$$

Using the last formulas and (2.15a), we have

$$M_i = \frac{1}{6h^2} \{-f_{i-2} + 10f_{i-1} - 18f_i + 10f_{i+1} - f_{i+2}\} + O(h^2) = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2} + O(h^2).$$

Thus, we find the solution of (2.16) with accuracy $O(h^2)$ without solving it. One can write system for $m_i = S_3'(x_i)$

$$m_{i-1} + 4m_i + m_{i+1} = \frac{3}{h}(f_{i+1} - f_{i-1}), \quad i = 1(1)n - 1,$$

which has the same matrix \mathbf{A} as (2.16). Consequently, using the same technique as above, we find

$$m_i = \frac{1}{12h} \{f_{i-2} - 8f_{i-1} + 8f_{i+1} - f_{i+2}\} + O(h^4). \tag{2.17}$$

Note that the system (1.2) with matrix $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$ arises also in constructing integro splines.

3 Application of approximate inverse matrices on constructing integro splines

In a uniform partition case the integro quadratic spline $S_2(x)$ satisfies relations [10]

$$S_2(x_{i-1}) + 4S_2(x_i) + S_2(x_{i+1}) = \frac{3}{h}(I_i + I_{i-1}), \quad i = 1(1)n - 1, \tag{3.18}$$

where

$$\int_{x_i}^{x_{i+1}} S_2(x) dx = \int_{x_i}^{x_{i+1}} y(x) dx = I_i, \quad i = 0(1)n - 1, \tag{3.19}$$

i.e. the integral values I_i of function $y(x)$ are known on the subintervals $[x_i, x_{i+1}]$, $h = (b - a)/n$. Obviously, one can use the B -spline representation of $S_2(x)$:

$$S_2(x) = \sum_{i=-1}^n b_i B_i(x), \tag{3.20}$$

where $B_i(x)$ are a normalized quadratic B -splines that forms a basis for C^1 quadratic splines space. For convenience, we present here B_i as:

$$B_i(x) = \frac{1}{2h^2} \begin{cases} (x - x_{i-1})^2, & [x_{i-1}, x_i], \\ (x - x_{i-1})^2 - 3(x - x_i)^2, & [x_i, x_{i+1}], \\ (x - x_{i+2})^2, & [x_{i+1}, x_{i+2}], \\ 0, & \text{else.} \end{cases} \tag{3.21}$$

The values of $B_i(x)$ and $B_i'(x)$ at the knots are given in Table 1.

Table 1

$B(x)$	x_{i-1}	x_i	x_{i+1}	x_{i+2}
$B_i(x)$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
$B_i'(x)$	0	$\frac{1}{h}$	$-\frac{1}{h}$	0

From (3.20) and using the properties of B -spline in Table 1, we obtain

$$S_2(x_i) = \frac{b_{i-1} + b_i}{2}, \quad S_2'(x_i) = \frac{b_i - b_{i-1}}{h}, \quad i = 0(1)n. \tag{3.22}$$

Taking into account (3.22), the relations (3.18) can be written in term of coefficients b_i as:

$$b_{i-2} + 5b_{i-1} + 5b_i + b_{i+1} = \frac{6}{h}(I_i + I_{i-1}), \quad i = 1(1)n - 1, \quad (3.23a)$$

or

$$z_{i-1} + z_i = \frac{6}{h}(I_i + I_{i-1}), \quad i = 1(1)n - 1, \quad (3.23b)$$

where

$$z_i = b_{i-1} + 4b_i + b_{i+1}. \quad (3.24)$$

From (3.23b) we deduce

$$z_i = \frac{6}{h}I_i,$$

or

$$b_{i-1} + 4b_i + b_{i+1} = \frac{6}{h}I_i, \quad i = 0(1)n - 1. \quad (3.25)$$

Analogously, using (3.22) and the relations

$$b_i = S_2(x_i) + \frac{h}{2}S_2'(x_i), \quad (3.26)$$

one can obtain

$$S_2'(x_{i-1}) + 4S_2'(x_i) + S_2'(x_{i+1}) = \frac{6}{h^2}(I_i - I_{i-1}), \quad i = 1(1)n - 1. \quad (3.27)$$

Thus, we have the systems (3.18), (3.25), and (3.27) with the same matrix but different right-hand sides. Since the matrix of these system is $\mathbf{A} = \text{Tri-diag}\{1, 4, 1\}$, we can use the above computed approximate inverse of this matrix. Using (2.15), from (3.18) we find

$$S_2(x_i) = \frac{3}{h} \left\{ \frac{1}{96}(I_{i-3} + I_{i-2} + I_{i+1} + I_{i+2}) - \frac{5}{72}(I_{i-2} + I_{i-1} + I_i + I_{i+1}) + \frac{41}{144}(I_i + I_{i-1}) \right\} + O(h^4) = \frac{1}{96h} \{ 3I_{i-3} - 17I_{i-2} + 62I_{i-1} + 62I_i - 17I_{i+1} + 3I_{i+2} \} + O(h^4). \quad (3.28)$$

For the values of I_i and $y \in C^4$, the following property holds

$$I_{i-2} - 4I_{i-1} + 6I_i - 4I_{i+1} + I_{i+2} = O(h^5). \quad (3.29)$$

We can simplify (3.28) by using (3.29). As a result we have

$$S_2(x_i) = \frac{1}{12h} \{ -I_{i-2} + 7I_{i-1} + 7I_i - I_{i+1} \} + O(h^4), \quad i = 2(1)n - 2. \quad (3.30)$$

Using the same technique, as preceding case, in (3.25) and (3.27) we obtain

$$b_i = \frac{8I_i - I_{i-1} - I_{i+1}}{6h} + O(h^4), \quad i = 1(1)n - 2, \quad (3.31)$$

and

$$S_2'(x_i) = \frac{1}{6h^2} \{ I_{i-2} - 9I_{i-1} + 9I_i - I_{i+1} \} + O(h^3), \quad i = 2(1)n - 2. \quad (3.32)$$

Thus, we first obtain approximate explicit formulas for $S_2(x_i)$, b_i , and $S_2'(x_i)$. In [10], we have the following estimation

$$S_2(x_i) = y_i + O(h^4),$$

but no estimation for the first derivative is given. Due to the explicit formula (3.32) one can obtain

$$S_2'(x_i) = y_i' + O(h^2), \quad i = 2(1)n - 3. \quad (3.33)$$

Another application of approximate inverse of matrix $\mathbf{A}=\text{Tri-diag}\{1, 4, 1\}$ is the well-known relations in [15]:

$$n_{i-1} + 4n_i + n_{i+1} = \frac{6}{h^4}(-I_{i-1} + 3I_i - 3I_{i+1} + I_{i+2}) + O(h^3), \quad i = 1(1)n - 3, \quad (3.34)$$

where $n_i = S_5'''(x_i)$. Such system appears in constructing quintic integro spline. As above, from (3.34) it follows that

$$\begin{aligned} n_i = & \frac{6}{h^4} \{ \dots + \alpha_{i,i-2}(-I_{i-3} + 3I_{i-2} - 3I_{i-1} + I_i) + \alpha_{i,i-1}(-I_{i-2} + 3I_{i-1} - 3I_i + I_{i+1}) \\ & + \alpha_{ii}(-I_{i-1} + 3I_i - 3I_{i+1} + I_{i+2}) + \alpha_{i,i+1}(-I_i + 3I_{i+1} - 3I_{i+2} + I_{i+3}) \\ & + \alpha_{i,i+2}(-I_{i+1} + 3I_{i+2} - 3I_{i+3} + I_{i+4}) + \dots \} + O(h^3). \end{aligned}$$

Using (2.15) and

$$I_{i-3} - 6I_{i-2} + 15I_{i-1} - 20I_i + 15I_{i+1} - 6I_{i+2} + I_{i+3} = O(h^7) \quad (3.35)$$

into the last formula, we obtain

$$n_i = \frac{1}{6h^4}(I_{i-2} - 11I_{i-1} + 28I_i - 28I_{i+1} + 11I_{i+2} - I_{i+3}) + O(h^3). \quad (3.36)$$

Analogously, we can find approximate inverse of matrix $\mathbf{A}=\text{Tri-diag}\{1, 10, 1\}$. Let $S_4(x)$ be an integro-quartic spline satisfying the conditions (3.19) and $m_i = S_4'(x_i)$, $T_i = S_4'''(x_i)$. Then we have the following relations [14]

$$m_{i-1} + 10m_i + m_{i+1} = \frac{12}{h^2}(I_i - I_{i-1}) + O(h^4), \quad (3.37)$$

and

$$T_{i-1} + 10T_i + T_{i+1} = \frac{12}{h^4}(-I_{i-2} + 3I_{i-1} - 3I_i + I_{i+1}) + O(h^2), \quad i = 2(1)n - 2. \quad (3.38)$$

In [14], we obtained the approximate formula

$$m_i = \frac{1}{12h^2}(I_{i-2} - 15I_{i-1} + 15I_i - I_{i+1}) + O(h^4), \quad i = 2(1)n - 2. \quad (3.39)$$

As above, we denote the entries of inverse matrix \mathbf{A}^{-1} by α_{ij} . Then from (3.37) and (3.39) we get

$$\begin{aligned} \frac{1}{144}(I_{i-2} - 15I_{i-1} + 15I_i - I_{i+1}) = & \dots + \alpha_{i,i-2}(I_{i-2} - I_{i-3}) + \alpha_{i,i-1}(I_{i-1} - I_{i-2}) \\ & + \alpha_{ii}(I_i - I_{i-1}) + \alpha_{i,i+1}(I_{i+1} - I_i) + \alpha_{i,i+2}(I_{i+2} - I_{i+1}) + \dots + O(h^4). \end{aligned} \quad (3.40)$$

Using symmetry of \mathbf{A}^{-1} and matching the coefficients of I_i on both sides (3.40) we obtain

$$\begin{aligned} 4\alpha_{i,i-2} + \alpha_{i,i-1} &= -\frac{1}{144}, \\ 10\alpha_{i,i-2} + \alpha_{i,i-1} - \alpha_{ii} &= -\frac{15}{144}, \\ \alpha_{ij} &= O(h^4), \quad |i - j| \geq 3. \end{aligned} \quad (3.41)$$

To derive the last formulas, we have used the relation (3.29). In addition to (3.41) we require that

$$2\alpha_{i,i-1} + 10\alpha_{ii} = 1, \quad (3.42)$$

which follows from (2.14). From (3.41), (3.42) we find that

$$\alpha_{ii} = \frac{191}{1872}, \quad \alpha_{i,i-1} = \alpha_{i,i+1} = -\frac{19}{1872}, \quad \alpha_{i,i-2} = \alpha_{i,i+2} = \frac{1}{1248}, \quad \alpha_{ij} = O(h^4), \quad |i - j| \geq 3. \quad (3.43)$$

Thus we find the entries of i -th row of \mathbf{A}^{-1} by formulas (3.43). Now we can use (3.43) to determine T_i from (3.38). As above, we get

$$T_i = \frac{12}{h^4} \{ \dots + \alpha_{i,i-2}(-I_{i-3} + 3I_{i-2} - 3I_{i-1} + I_i) + \alpha_{i,i-1}(-I_{i-2} + 3I_{i-1} - 3I_i + I_{i+1}) \\ + \alpha_{ii}(-I_{i-1} + 3I_i - 3I_{i+1} + I_{i+2}) + \alpha_{i,i+1}(-I_i + 3I_{i+1} - 3I_{i+2} + I_{i+3}) \\ + \alpha_{i,i+2}(-I_{i+1} + 3I_{i+2} - 3I_{i+3} + I_{i+4}) + \dots \}.$$

Using (3.29) and (3.43) into the last formula, we obtain the well-known explicit formula that was derived first in [14]

$$T_i = \frac{1}{h^4} (-I_{i-2} + 3I_{i-1} - 3I_i + I_{i+1}) + O(h^2). \tag{3.44}$$

Now we consider the matrix $\mathbf{A} = \text{Tri-diag}\{1, d, 1\}$ with $|d| > 2$. Obviously, the above two cases are particular cases of this matrix with $d = 4$ and $d = 10$. From Theorem 2.1 in [7], we get the following explicit formula for \mathbf{A}^{-1} of $\mathbf{A} = \text{Tri-diag}\{1, d, 1\}$,

$$\alpha_{ij} = (-1)^{i-j} \frac{1}{p_j - q_j} \prod_{s=i}^{j-1} \frac{1}{p_s}, \quad i = 1, 2, \dots, \left\lceil \frac{n}{2} \right\rceil; \quad j = i, i+1, \dots, n-i+1, \tag{3.45}$$

where

$$p_1 = d, \quad p_i = d - \frac{1}{p_{i-1}}, \quad i = 2, 3, \dots, n, \\ q_n = 0, \quad q_i = \frac{1}{d - q_{i+1}}, \quad i = n-1, n-2, \dots, 1, \quad \text{and } \alpha_{ij} = \alpha_{ji}.$$

From (2.15) and (3.43) we obtain the approximate inverse of $\mathbf{A} = \text{Tri-diag}\{1, d, 1\}$ with $|d| > 2$. Indeed from (3.45) it follows that

$$\alpha_{ii} = \frac{1}{d - \frac{1}{\frac{d}{2} + \frac{1}{3d+8}}} = \frac{3d^2 + 8d + 2}{3d^3 + 8d^2 - 4d - 16}, \quad \alpha_{i,i\pm 1} = \frac{1 - d\alpha_{ii}}{2}, \tag{3.46} \\ \alpha_{i,i\pm 2} = \frac{3}{3d^3 + 8d^2 - 4d - 16}, \quad \alpha_{ij} = O(h^4), \quad |i - j| \geq 3.$$

Using the general formulas (3.46) one can easily get (2.15) and (3.43).

4 Approximate inverse of penta-diagonal matrices

Let $S_4(x)$ be an integro-quartic spline satisfying conditions (3.19) with its B -spline representation

$$S_4(x) = \sum_{i=-2}^{n+1} c_i B_i(x). \tag{4.47}$$

Here $B_i(x)$ are quartic B -spline which forms basis for $S_4 \in C^3[a, b]$ spaces. Then with the uniform partition, the conditions (3.19) imply

$$c_{i-2} + 26c_{i-1} + 66c_i + 26c_{i+1} + c_{i+2} = \frac{120}{h} I_i, \quad i = 0(1)n-1. \tag{4.48}$$

In [14] we obtained approximate and explicit formula

$$c_i = \frac{13I_{i-3} - 39I_{i-2} - 94I_{i-1} + 746I_i - 159I_{i+1} + 13I_{i+2}}{480h} + O(h^5), \quad i = 3(1)n-3.$$

It is easy to show that by using expression (3.29), the c_i can be rewritten in more symmetric form

$$c_i = \frac{13I_{i-2} - 112I_{i-1} + 438I_i - 112I_{i+1} + 13I_{i+2}}{240h} + O(h^5). \quad (4.49)$$

As before, we denote the entries of inverse of matrix $\mathbf{A} = \text{Penta-diag}\{1, 26, 66, 26, 1\}$ of system (4.48) by α_{ij} . Then, from (4.48) and (4.49) we get

$$\begin{aligned} & \frac{1}{28\,800} (13(I_{i-2} + I_{i+2}) - 112(I_{i-1} + I_{i+1}) + 438I_i) + O(h^6) = \\ & \dots + \alpha_{i,i-3}(I_{i-3} + I_{i+3}) + \alpha_{i,i-2}(I_{i-2} + I_{i+2}) + \alpha_{i,i-1}(I_{i-1} + I_{i+1}) + \alpha_{ii}I_i \end{aligned} \quad (4.50)$$

in which we have used symmetry of α_{ij} and $\alpha_{i,i-j} = \alpha_{i,i+j}$, $j = 1, 2, 3$.

Using (3.35), from (4.50) we have

$$\begin{aligned} & \frac{1}{28\,800} (13(I_{i-2} + I_{i+2}) - 112(I_{i-1} + I_{i+1}) + 438I_i) + O(h^6) = (20\alpha_{i,i-3} + \alpha_{ii})I_i \\ & + (-15\alpha_{i,i-3} + \alpha_{i,i-1})(I_{i-1} + I_{i+1}) + (6\alpha_{i,i-3} + \alpha_{i,i-2})(I_{i-2} + I_{i+2}) + \dots \end{aligned} \quad (4.51)$$

Equating the coefficients of $I_{i-j} + I_{i+j}$ for $j = 0, 1, 2$ in both sides of last expression we get

$$\begin{aligned} 6\alpha_{i,i-3} + \alpha_{i,i-2} &= \frac{13}{28\,800}, \\ -15\alpha_{i,i-3} + \alpha_{i,i-1} &= -\frac{112}{28\,800}, \\ 20\alpha_{i,i-3} + \alpha_{ii} &= \frac{438}{28\,800}, \\ \alpha_{i,i-j} &= O(h^5), \quad |i-j| \geq 4. \end{aligned} \quad (4.52)$$

In addition to (4.52), we require that

$$2\alpha_{i,i-2} + 52\alpha_{i,i-1} + 66\alpha_{ii} = 1 \quad (4.53)$$

which follows from (2.14). The solutions of (4.52) and (4.53) are given by

$$\begin{aligned} \alpha_{ii} &= \frac{44\,447}{1\,987\,200}, \quad \alpha_{i,i-1} = -\frac{24\,529}{2\,649\,600}, \\ \alpha_{i,i-2} &= \frac{3\,443}{1\,324\,800}, \quad \alpha_{i,i-3} = -\frac{569}{1\,589\,760}, \\ \alpha_{i,i-j} &= O(h^5), \quad |i-j| \geq 4. \end{aligned} \quad (4.54)$$

Now we consider the following systems

$$m_{i-2} + 56m_{i-1} + 246m_i + 56m_{i+1} + m_{i+2} = b_i, \quad (4.55)$$

$$n_{i-2} + 56n_{i-1} + 246n_i + 56n_{i+1} + n_{i+2} = d_i, \quad i = 2(1)n - 2, \quad (4.56)$$

where $m_i = S_5'(x_i)$ and $n_i = S_5'''(x_i)$ and

$$b_i = \frac{30}{h^2} (-I_{i-1} - 9I_i + 9I_{i+1} + I_{i+2}), \quad d_i = \frac{360}{h^4} (-I_{i-1} + 3I_i - 3I_{i+1} + I_{i+2}). \quad (4.57)$$

These systems arise in constructing integro quintic spline [2]. An explicit and approximate solution of system (4.55) can be found in [15] as

$$m_i = \frac{1}{180h^2} (-2I_{i-2} + 25I_{i-1} - 245I_i + 245I_{i+1} - 25I_{i+2} + 2I_{i+3}) + O(h^5), \quad i = 2(1)n - 4. \quad (4.58)$$

In this case, using analogous technique, as above, we find that

$$\begin{aligned}
 81\alpha_{i,i-3} + 16\alpha_{i,i-2} + \alpha_{i,i-1} &= \frac{1}{2\,700}, \\
 295\alpha_{i,i-3} + 30\alpha_{i,i-2} - 9\alpha_{i,i-1} - \alpha_{ii} &= \frac{25}{5\,400}, \\
 504\alpha_{i,i-3} + 34\alpha_{i,i-2} - 8\alpha_{i,i-1} + 9\alpha_{ii} &= \frac{245}{5\,400}, \\
 2\alpha_{i,i-2} + 112\alpha_{i,i-1} + 246\alpha_{ii} &= 1, \\
 \alpha_{i,i-j} &= O(h^5), \quad |i-j| \geq 4.
 \end{aligned} \tag{4.59}$$

The solutions of (4.59) are

$$\begin{aligned}
 \alpha_{ii} &= \frac{9\,979}{2\,195\,100}, \quad \alpha_{i,i-1} = -\frac{8\,273}{7\,804\,800}, \\
 \alpha_{i,i-2} &= \frac{577}{2\,926\,800}, \quad \alpha_{i,i-3} = -\frac{299}{14\,048\,640}, \\
 \alpha_{i,i-j} &= O(h^5), \quad |i-j| \geq 4.
 \end{aligned} \tag{4.60}$$

Using (3.35), (4.56), (4.57) and (4.60), we obtain (3.36).

Note that, in [13] Z-folding algorithm was proposed for solving the penta-diagonal system of linear equations, which allows us to reduce the system by solving two tri-diagonal systems sequentially. We can find the approximate inverse of penta-diagonal matrices by using the Z-folding algorithm and (3.46), but this approach is not suitable to obtain explicit formulas as (3.36) and (4.58).

5 Conclusion

For some application cases it is not necessary to find all entries of the inverse matrices of band matrices. The main advantage of our approach are simple and explicit formulas for only main diagonal and its few adjacent diagonals entries of the inverse matrices.

Acknowledgments

The authors acknowledge the many helpful suggestions of the referees during the preparation of the paper. The work was partially supported by Foundation of Science and Technology of Mongolia (No. SST_007/2015).

References

- [1] S. Barnett, *Matrices: Methods and Applications*, Oxford University Press, New York, (1990).
- [2] H. Behforooz, Interpolation by integro quintic splines, *Appl. Math. Comput*, 216 (2010) 364-367.
<https://doi.org/10.1016/j.amc.2010.01.009>
- [3] P. Bickel, M. Lindner, Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics, *Theory Probab. Appl*, 56 (2012) 1-20.
<https://doi.org/10.1137/S0040585X97985224>
- [4] S. Demko, Inverses of band matrices and local convergence of spline projections, *SIAM J. Numer. Anal*, 14 (1977) 616-619.
<https://doi.org/10.1137/0714041>
- [5] C. F. Fischer, R. A. Usmani, Properties of some tridiagonal matrices and their application to boundary value problems, *SIAM J. Numer. Anal*, 6 (1969) 127-142.
<https://doi.org/10.1137/0706014>

- [6] W. K. Grassmann, Real eigenvalues of certain tridiagonal matrix polynomials, with queueing applications, *Linear Algebra Appl*, 342 (2002) 93-106.
[https://doi.org/10.1016/S0024-3795\(01\)00462-1](https://doi.org/10.1016/S0024-3795(01)00462-1)
- [7] J. Jia, S. Li, Symbolic algorithms for the inverses of general k -tridiagonal matrices, *Comput. Math. Appl*, 70 (2015) 3032-3042.
<https://doi.org/10.1016/j.camwa.2015.10.018>
- [8] P. Sablonnière, Univariate spline quasi-interpolants and applications to numerical analysis, *Rend. Sem. Mat. Univ. Pol. Torino*, 63 (2005) 107-118.
- [9] D. C. Smolarski, Diagonally-stripped matrices and approximate inverse preconditioners, *J. Comput. Appl. Math*, 186 (2006) 416-431.
<https://doi.org/10.1016/j.cam.2005.02.012>
- [10] J. Wu, X. Zhang, Integro quadratic spline interpolation, *Appl. Math. Modell*, 39 (2015) 2973-2980.
<https://doi.org/10.1016/j.apm.2014.11.015>
- [11] T. Yamamoto, Inversion formulas for tridiagonal matrices with applications to boundary value problems, *Numer. Funct. Anal. Optim*, 22 (2001) 357-385.
<https://doi.org/10.1081/NFA-100105108>
- [12] Yu. S. Zav'yalov, B. I. Kvasov, V. L. Miroschnichenko, *Spline function methods* Nauka Moscow, (1980) in Russian.
- [13] T. Zhanlav, Z -folding and its applications, *Mong. Math. J*, 17 (2013) 68-74.
- [14] T. Zhanlav, R. Mijiddorj, On local integro quartic splines, *Appl. Math. Comput*, 269 (2015) 301-307.
<https://doi.org/10.1016/j.amc.2015.07.077>
- [15] T. Zhanlav, R. Mijiddorj, H. Behforooz, Construction of local integro quintic splines, *Communications in Numerical Analysis*, 2 (2016) 167-179.
<https://doi.org/10.5899/2016/cna-00267>



Construction of a Family of C^1 Convex Integro Cubic Splines

Zhanlav Tugal¹ and Mijiddorj Renchin-Ochir^{*1,2}

¹Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia

²Department of Informatics, Mongolian National University of Education, Ulaanbaatar, Mongolia

*Corresponding author: mijiddorj@msue.edu.mn

Abstract. We construct a family of monotone and convex C^1 integro cubic splines under a strictly convex position of the dataset. Then, we find an optimal spline by considering its approximation properties. Finally, we give some examples to illustrate the convex-preserving properties of these splines.

Keywords. Shape-preserving; Approximation; Integro spline

MSC. 41A15; 65D10

Received: April 1, 2020

Accepted: August 28, 2020

Copyright © 2020 Zhanlav Tugal and Mijiddorj Renchin-Ochir. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. Introduction

Let Δ_k be the non-uniform partition on $[a, b]$, $a = x_0 < x_1 < \dots < x_k = b$, and $h_{i+1} = x_{i+1} - x_i$, $i = 0, \dots, k-1$, are step sizes. Let $S(x)$ be a cubic spline that approximates a function $u(x)$. We assume that the function values $u_i = u(x_i)$ are not given, but the integral values $h_{i+1}I_{i+1}$ on the subintervals $[x_i, x_{i+1}]$ of $u(x)$ are known. The problem of the construction of an integro cubic spline (see [2]) is to find $S(x)$ such that

$$\int_{x_i}^{x_{i+1}} S(x)dx = \int_{x_i}^{x_{i+1}} u(x)dx = h_{i+1}I_{i+1}, \quad i = 0, \dots, k-1. \quad (1)$$

We use a notation $m_i = S'(x_i)$ and piecewise polynomial representation

$$S(x) = (1-t)^2(1+2t)S(x_i) + t^2(3-2t)S(x_{i+1}) + h_{i+1}t(1-t)\{(1-t)m_i - tm_{i+1}\},$$

$$x \in [x_i, x_{i+1}], \quad t = \frac{x-x_i}{h_{i+1}}, \quad i = 0, \dots, k-1, \quad t \in [0, 1]. \quad (2)$$

The sufficient conditions for convexity of cubic histosplines derived in [8, 9] can be written in the form:

$$2m_{i-1} + m_i \leq \frac{3}{h_i}(S(x_i) - S(x_{i-1})) \leq m_{i-1} + 2m_i, \quad i = 1, \dots, k,$$

$$\frac{1}{2}(S(x_{i-1}) + S(x_i)) + \frac{h_i}{12}(m_{i-1} - m_i) = I_i, \quad i = 1, \dots, k. \quad (3)$$

Obviously, the system (3) has an infinite number of solutions. In [9–11], the staircase algorithm with three terms of recurrence relations was used to find the solutions of (3). Moreover, shape-preserving approximations of histosplines have been studied in [3, 4, 15] and references therein. There are two traditional approaches to constructing shape-preserving histosplines: additional knots of spline and splines of a higher order with less smoothness [4]. It is well known that the interpolating cubic spline of the class C^2 does not preserve the monotonicity and convexity of the input data. Recently, the shape-preserving properties of the C^2 local integro cubic spline have been investigated only on a uniform partition in [14].

Usually, the monotonicity and convexity preserving property of the spline $S(x)$ are discussed based on the properties of the data $u_i = u(x_i)$. Now, we discuss the properties of monotonicity and convexity of the spline $S(x)$ based on data I_i . We construct a family of monotone and convex C^1 integro cubic splines under a strictly convex position of the data set. In this paper, we will give a simple constructive algorithm for C^1 integro cubic splines (or histosplines) that preserve monotonicity and convexity. The remainder of this paper is organized as follows. In Section 2, a simple method for constructing the family of C^1 integro cubic splines (depending on the parameter α) is given. We discuss sufficient conditions of monotonicity and convexity of the presented integro cubic splines. We also consider an error analysis of the integro cubic splines in Section 3. Some numerical examples are given in Section 4 to illustrate the convexity preserving property.

2. Construction of Convex Integro Cubic Splines

Using the ideas in [12, 13] instead of inequalities in (3), we consider the following relations:

$$\frac{3}{h_i}(S(x_i) - S(x_{i-1})) = \alpha(m_{i-1} + 2m_i) + (1-\alpha)(2m_{i-1} + m_i)$$

$$= (2-\alpha)m_{i-1} + (1+\alpha)m_i, \quad \alpha \in [0, 1]. \quad (4)$$

The right-hand side of (4) is a linear combination of $m_{i-1} + 2m_i$ and $2m_{i-1} + m_i$, and is a linear function with respect to α . Hence, from (4), it follows when $m_{i-1} \leq m_i$ that

$$2m_{i-1} + m_i \leq \frac{3}{h_i}(S(x_i) - S(x_{i-1})) \leq m_{i-1} + 2m_i. \quad (5)$$

That is, instead of (3), it is possible to consider

$$\begin{aligned} \frac{3}{h_i}(S(x_i) - S(x_{i-1})) &= (2 - \alpha)m_{i-1} + (1 + \alpha)m_i, \\ \frac{3}{h_i}(S(x_i) + S(x_{i-1})) &= \frac{6}{h_i}I_i - \frac{1}{2}(m_{i-1} - m_i). \end{aligned} \tag{6}$$

By adding and subtracting these two equations, we get

$$S(x_i) = I_i + \frac{h_i}{12}\{(3 - 2\alpha)m_{i-1} + (3 + 2\alpha)m_i\}, \quad i = 1, \dots, k, \tag{7}$$

$$S(x_{i-1}) = I_i + \frac{h_i}{12}\{(2\alpha - 5)m_{i-1} - (2\alpha + 1)m_i\}, \quad i = 1, \dots, k. \tag{8}$$

From (7) and (8), it follows that

$$\lambda_i(3 - 2\alpha)m_{i-1} + \{\lambda_i(3 + 2\alpha) + \mu_i(5 - 2\alpha)\}m_i + \mu_i(2\alpha + 1)m_{i+1} = 6\delta I_i, \quad i = 1, \dots, k - 1, \tag{9}$$

where

$$\lambda_i = \frac{h_i}{h_i + h_{i+1}}, \quad \mu_i = 1 - \lambda_i, \quad \delta I_i = \frac{I_{i+1} - I_i}{\bar{h}_i}, \quad \bar{h}_i = \frac{h_i + h_{i+1}}{2}. \tag{10}$$

Using the eq. (7), (8), and (9), we get a closed system of equations

$$\begin{aligned} (5 - 2\alpha)m_0 + (2\alpha + 1)m_1 &= \frac{12}{h_1}(I_1 - S(x_0)), \\ a_i m_{i-1} + c_i m_i + b_i m_{i+1} &= 6\delta I_i, \quad i = 1, \dots, k - 1, \\ (3 - 2\alpha)m_{k-1} + (3 + 2\alpha)m_k &= \frac{12}{h_k}(S(x_k) - I_k), \end{aligned} \tag{11}$$

where

$$a_i = \lambda_i(3 - 2\alpha) > 0, \quad b_i = \mu_i(1 + 2\alpha) > 0, \quad c_i = a_i + b_i + 4(\alpha\lambda_i + (1 - \alpha)\mu_i) > a_i + b_i > 0. \tag{12}$$

Since,

$$\begin{aligned} 5 - 2\alpha - 2\alpha - 1 &= 4(1 - \alpha) \geq 0, \\ \lambda_i(3 + 2\alpha) + \mu_i(5 - 2\alpha) - \lambda_i(3 - 2\alpha) - \mu_i(2\alpha + 1) &= 4\lambda_i\alpha + 4\mu_i(1 - \alpha) > 0, \quad i = 1, \dots, k - 1, \\ 3 + 2\alpha - 3 + 2\alpha &= 4\alpha \geq 0, \end{aligned}$$

the matrix of the system (11) has diagonal dominance. Hence, the system (11) has a unique solution (m_0, m_1, \dots, m_k) for each $\alpha \in [0, 1]$, and it can be easily solved by using the tridiagonal LU decomposition algorithm. Here, $S(x_0)$ and $S(x_k)$ are assumed to be given for now. The values of $S(x_i)$ are determined by means of (7) or (8), and the spline S is given by (2). Then, $S(x)$ will be C^1 cubic integro splines depending on the parameter $\alpha \in [0, 1]$. Thus, the family of $S(x, \alpha)$ depending on the parameter α is determined completely. As usual, the given data I_i is called monotonically increasing if

$$\delta I_i = \frac{I_{i+1} - I_i}{\bar{h}_i} \geq 0, \quad i = 1, \dots, k - 1, \tag{13}$$

and convex if

$$\frac{I_{i+1} - I_i}{\bar{h}_i} - \frac{I_i - I_{i-1}}{\bar{h}_{i-1}} \geq 0, \quad i = 2, \dots, k - 1, \tag{14a}$$

or

$$\delta I_i \geq \delta I_{i-1}. \quad (14b)$$

Using the Taylor expansion of $S(x)$ in (2), we obtain

$$S(x_0) = I_1 + \frac{h_1}{12} \left\{ \frac{\mu_1(1+2\alpha)(2\alpha-5)(\delta I_1 - \delta I_2)}{\lambda_1(3-2\alpha)} - 6\delta I_1 \right\},$$

$$S(x_k) = I_k + \frac{h_k}{12} \left\{ \frac{\lambda_{k-1}(9-4\alpha^2)(\delta I_{k-1} - \delta I_{k-2})}{\mu_{k-1}(1+2\alpha)} + 6\delta I_{k-1} \right\}.$$

To study the shape-preserving properties of (2), one must use the derivatives of (2), which are

$$S'(x) = 6t(1-t) \frac{S(x_{i+1}) - S(x_i)}{h_{i+1}} + (1-t)(1-3t)m_i + t(3t-2)m_{i+1}, \quad (15)$$

and

$$S''(x) = 6(1-2t) \frac{S(x_{i+1}) - S(x_i)}{h_{i+1}^2} + \frac{1}{h_{i+1}} \{(6t-4)m_i + (6t-2)m_{i+1}\}. \quad (16)$$

Using (6) in (15) and (16), we obtain

$$S'(x) = (1-t)(1+(1-2\alpha)t)m_i + t(2\alpha+t(1-2\alpha))m_{i+1}, \quad (17)$$

and

$$S''(x) = 2 \frac{\alpha+t(1-2\alpha)}{h_{i+1}} (m_{i+1} - m_i). \quad (18)$$

It is easy to show that

$$1+(1-2\alpha)t \geq 0, \quad 2\alpha+t(1-2\alpha) \geq 0, \quad \text{for } \alpha \in [0, 1].$$

Hence, from (17), it follows that

$$S'(x) \geq 0, \quad x \in [x_i, x_{i+1}] \text{ if } m_i \geq 0 \text{ and } m_{i+1} \geq 0. \quad (19)$$

Since $\alpha+t(1-2\alpha) \geq 0$ then from (18), it follows that

$$S''(x) \geq 0, \quad x \in [x_i, x_{i+1}] \text{ if } m_{i+1} - m_i \geq 0. \quad (20)$$

Thus, from (19), we conclude that $S(x, \alpha)$ will monotonically increase if the solution to (11) is nonnegative. In order to study the solution to (11), we use the following theorem given in [5].

Theorem 1. For the system $\mathbf{Ax} = \mathbf{f}$, suppose that

$$a_{ij} \geq 0, \quad a_{ii} > 0, \quad f_i > 0, \quad i, j = 1, \dots, k, \quad i \neq j.$$

If for all i , $i = 1, \dots, k$,

$$f_i > \sum_{j=1, j \neq i}^k a_{ij} \frac{f_j}{a_{jj}},$$

then \mathbf{A} is invertible, and $x_i = (\mathbf{A}^{-1}\mathbf{f})_i > 0$ for all i .

We show that the assumptions given in Theorem 1 are fulfilled for our system (11) under conditions

$$\frac{2a_1(I_1 - S(x_0))}{h_1(5-2\alpha)} + \frac{b_1\delta I_2}{c_2} < \delta I_1 < \frac{2c_1(I_1 - S(x_0))}{h_1(2\alpha+1)}, \quad I_1 - S(x_0) > 0, \quad (21a)$$

$$\frac{\alpha_j \delta I_{j-1}}{c_{j-1}} + \frac{b_j \delta I_{j+1}}{c_{j+1}} < \delta I_j, \quad j = 2, \dots, k-2, \tag{21b}$$

$$\frac{2b_{k-1}(S(x_k) - I_k)}{h_k(3 + 2\alpha)} + \frac{a_{k-1} \delta I_{k-2}}{c_{k-2}} < \delta I_{k-1} < \frac{2c_{k-1}(S(x_k) - I_k)}{h_k(3 - 2\alpha)}, \quad S(x_k) - I_k > 0. \tag{21c}$$

Let us summarize the obtained above results as:

Theorem 2. *Let the integro cubic splines $S(x, \alpha) \in C^1[a, b]$ be defined by (2), (7), and (11), and the data I_i monotonically increase. If the inequalities (21) are valid then $m_i > 0$ for all $i = 0, \dots, k$ and thereby $S'(x) > 0$ on $[x_0, x_k]$, that is, S is monotonically increasing on $[a, b]$.*

Now, we proceed to study the convexity property of $S(x, \alpha)$. To this end, we pass from (11) to the following system

$$(2\alpha + 1)(m_1 - m_0) = \frac{12}{h_1}(I_1 - S(x_0)) - 6m_0, \tag{22a}$$

$$\begin{aligned} \alpha_i(m_{i-1} - m_{i-2}) + c_i(m_i - m_{i-1}) + b_i(m_{i+1} - m_i) \\ = 6(\delta I_i - \delta I_{i-1}) + (\alpha_{i-1} - \alpha_i)m_{i-2} + (c_{i-1} - c_i)m_{i-1} + (b_{i-1} - b_i)m_i, \quad i = 2, \dots, k-1, \end{aligned} \tag{22b}$$

$$(3 + 2\alpha)(m_k - m_{k-1}) = \frac{12}{h_k}(S(x_k) - I_k) - 6m_{k-1}. \tag{22c}$$

If the following equalities hold:

$$\alpha_{i-1} - \alpha_i = 0, \quad c_{i-1} - c_i = 0, \quad b_{i-1} - b_i = 0, \tag{23}$$

then the equation (22b) for $i = 2, \dots, k-1$ leads to

$$\alpha_i(m_{i-1} - m_{i-2}) + c_i(m_i - m_{i-1}) + b_i(m_{i+1} - m_i) = 6(\delta I_i - \delta I_{i-1}), \quad i = 2, \dots, k-1. \tag{24}$$

As above, it is easy to verify that the assumptions of Theorem 1 are fulfilled for the system (22a), (22c), and (24) under conditions

$$\frac{\alpha_2(\frac{2(I_1 - S(x_0))}{h_1} - m_0)}{2\alpha + 1} + \frac{b_2(\delta I_3 - \delta I_2)}{c_3} < \delta I_2 - \delta I_1, \tag{25a}$$

$$\frac{\alpha_j(\delta I_{j-1} - \delta I_{j-2})}{c_{j-1}} + \frac{b_j(\delta I_{j+1} - \delta I_j)}{c_{j+1}} < \delta I_j - \delta I_{j-1}, \quad j = 3, \dots, k-2, \tag{25b}$$

$$\frac{b_{k-1}(\frac{2}{h_k}(S(x_k) - I_k) - m_{k-1})}{3 + 2\alpha} + \frac{a_{k-1}(\delta I_{k-2} - \delta I_{k-3})}{c_{k-2}} < \delta I_{k-1} - \delta I_{k-2}, \tag{25c}$$

where $\frac{2}{h_1}(I_1 - S(x_0)) - m_0 > 0$ and $\frac{2}{h_k}(S(x_k) - I_k) - m_{k-1} > 0$. Thus, we have:

Theorem 3. *Let the integro cubic splines $S(x, \alpha) \in C^1[a, b]$ be defined by (2), (7), and (11), and the data I_i are convex, and m_0 and m_{k-1} are given. If (23) and (25) are valid then $m_i - m_{i-1} > 0$ for all $i = 1, \dots, k$ and thereby $S''(x) > 0$ on $[x_0, x_k]$, that is, $S(x)$ is convex on $[a, b]$.*

Note that the equalities (23) hold true if the step sizes of grid satisfy

$$h_i = \sqrt{h_{i-1}h_{i+1}}, \quad i = 1, \dots, k-1. \tag{26}$$

Of course, the conditions (26) are fulfilled on a uniform partition. Now, we are interested in the dependence of m_i on parameter α . To this end, differentiating the system (11) with respect to

α , we obtain

$$\begin{aligned} (5 - 2\alpha)m'_0(\alpha) + (2\alpha + 1)m'_1(\alpha) &= 2(m_0 - m_1), \\ \alpha m'_{i-1}(\alpha) + c_i m'_i(\alpha) + b_i m'_{i+1}(\alpha) &= 2\lambda_i(m_{i-1} - m_i) + 2\mu_i(m_i - m_{i+1}), \quad i = 1, \dots, k-1, \\ (3 - 2\alpha)m'_{k-1}(\alpha) + (3 + 2\alpha)m'_k(\alpha) &= 2(m_{k-1} - m_k). \end{aligned} \quad (27)$$

From (27), it is clear that the right-hand side of the system (27) is negative if (23) and (25) are fulfilled.

Theorem 4. Assume that (23) and (25) are fulfilled. Then, $m_i(\alpha)$ is a decreasing function with respect to α for all $i = 1, \dots, k-1$, i.e.,

$$m_i(0) \geq m_i(\alpha) \geq m_i(1), \quad i = 1, \dots, k-1. \quad (28)$$

Proof. By Theorems 1 and 3, the solution to system (27) is negative, that is, $m'_i(\alpha) < 0$ for all $i = 1, \dots, k-1$. Hence, (28) is valid. \square

Thus, we obtain feasible intervals $[m_i(1), m_i(0)]$ of $m_i(\alpha)$ that ensure the monotonicity and convexity of splines $S(x, \alpha)$. From (7) and (28), we also derive the interval of $S(x_i, \alpha)$:

$$S(x_i, \alpha) \in \left[I_i + \frac{h_i}{2} m_{i-1}(1), I_i + \frac{h_i}{2} m_i(0) \right], \quad i = 1, \dots, k-1. \quad (29)$$

Theorem 5. Let the assumptions of Theorem 3 be fulfilled and $m_0 > 0$. Then, $S(x_i)$ are in strictly convex positions as the data I_i , that is,

$$\delta S(x_i) > \delta S(x_{i-1}) \geq 0, \quad i = 1, \dots, k-1. \quad (30)$$

Proof. By (4) and Theorem 3, we have

$$\delta S(x_i) = \frac{S(x_{i+1}) - S(x_i)}{h_{i+1}} = \frac{1}{3} \{ (2 - \alpha)m_i + (1 + \alpha)m_{i+1} \} \geq m_i \geq 0.$$

By Theorem 3, we have

$$m_{i+1} - m_i > 0, \quad i = 0, \dots, k-1,$$

$$m_i - m_{i-1} > 0, \quad i = 1, \dots, k.$$

From this, we obtain

$$(1 + \alpha)(m_{i+1} - m_i) + (2 - \alpha)(m_i - m_{i-1}) > 0,$$

which leads to

$$(1 + \alpha)m_{i+1} + (2 - \alpha)m_i > (1 + \alpha)m_i + (2 - \alpha)m_{i-1}, \quad i = 1, \dots, k-1. \quad (31)$$

Using (6) in (31), we get (30). \square

The well-known convex interval interpolation problem was solved by three-term staircase algorithm in [7]. From (14) and (29), one can easily see that we solved the convex interval interpolation problem $S(x_i) \in [l_i, v_i]$, $i = 0, \dots, k$, for a particular case with $l_i = I_i + \frac{h_i}{2} m_{i-1}(1)$, $v_i = I_i + \frac{h_i}{2} m_i(0)$.

3. Error Analysis

Now, we consider the approximation properties of convex integro cubic splines $S(x, \alpha)$. Using the Taylor expansion of $u \in C^3[a, b]$, one can easily obtain

$$\delta I_i = u'_i + \frac{h_{i+1} - h_i}{3} u''_i + O(\bar{h}^2), \tag{32}$$

$$I_i = u_i - \frac{h_i}{2} u'_i + \frac{h_i^2}{6} u''_i + O(\bar{h}^3), \tag{33}$$

where $\bar{h} = \max_{1 \leq i \leq k} h_i$.

From (32), it is clear that

$$\delta I_i = u'_i + O(\bar{h}^2), \tag{34}$$

under condition

$$h_{i+1} - h_i = O(\bar{h}^2). \tag{35}$$

Theorem 6. Let $S(x, \alpha)$ be C^1 integro cubic splines defined by (2), (7), (11), and $S(x_i) = u_i + O(\bar{h}^3)$, $i = 0, k$. Then, for $u \in C^3$, we have estimations

$$S^{(r)}(x_i) - u_i^{(r)} = O(\bar{h}^{\sigma+1-r}), \quad r = 0, 1, \quad i = 0, \dots, k. \tag{36}$$

under (35). Here $\sigma = 1$ when $\alpha \neq \frac{1}{2}$ and $\sigma = 2$ when $\alpha = \frac{1}{2}$.

Proof. First, let us estimate $q_i = m_i - u'_i$, $i = 0, \dots, k$. To this end, we pass from (11) to the system

$$\begin{aligned} (5 - 2\alpha)q_0 + (2\alpha + 1)q_1 &= d_0, \\ a_i q_{i-1} + c_i q_i + b_i q_{i+1} &= d_i, \quad i = 1, \dots, k-1, \\ (3 - 2\alpha)q_{k-1} + (3 + 2\alpha)q_k &= d_k, \end{aligned} \tag{37}$$

where

$$\begin{aligned} d_0 &= \frac{12}{h_1}(I_1 - S(x_0)) - \{(5 - 2\alpha)u'_0 + (2\alpha + 1)u'_1\}, \\ d_i &= 6\delta I_i - \{a_i u'_{i-1} + c_i u'_i + b_i u'_{i+1}\}, \\ d_k &= \frac{12}{h_k}(S(x_k) - I_k) - \{(3 - 2\alpha)u'_{k-1} + (3 + 2\alpha)u'_k\}. \end{aligned} \tag{38}$$

Using (33), (34), (35), and the Taylor expansion of function $u \in C^3[a, b]$ in (38), one can easily obtain

$$d_i = O(\bar{h}^\sigma). \tag{39}$$

Then, from (37) and (39), it follows (36) for $r = 1$. From (7), we get

$$\begin{aligned} S(x_i) - u_i &= I_i - u_i + \frac{h_i}{12} \{(3 - 2\alpha)(m_{i-1} - u'_{i-1}) + (3 + 2\alpha)(m_i - u'_i)\} \\ &\quad + \frac{h_i}{12} \{(3 - 2\alpha)u'_{i-1} + (3 + 2\alpha)u'_i\}, \quad i = 1, \dots, k-1. \end{aligned} \tag{40}$$

As above, using (33), (36) for $r = 1$ and the Taylor expansion of function $u \in C^3[a, b]$ in (40), we obtain

$$S(x_i) - u_i = O(\bar{h}^{\sigma+1}), \quad i = 1, \dots, k-1,$$

i.e., the estimate (36) is proved for $r = 0$. This completes the proof of Theorem 6. \square

Using (34) and (36) for $r = 1$, it is easy to show that

$$S''_{i+0} - S''_{i-0} = O(\bar{h}^{\sigma-1}), \quad i = 1, \dots, k-1. \quad (41)$$

From the estimations (36) and (41), it is clear that the best or optimal C^1 integro cubic spline (abbr. OCICS) is derived when $\alpha = \frac{1}{2}$ in the sense of approximation properties. This selection shows that using an optimal choice of parameter one can rise the order of approximation.

4. Numerical Experiments

In this section, we apply the proposed method to some numerical examples.

Example 1. We consider the histogram $I = \{1, 2, 4\}$ on $\Delta_3 = \{0 < 4 < 6 < 7\}$ in [9]. A convex integro cubic spline (abbr. CICS) curve with $\alpha = 0.5$ is shown in Figure 1, and with $\alpha = 1$ is shown in Figure 2.

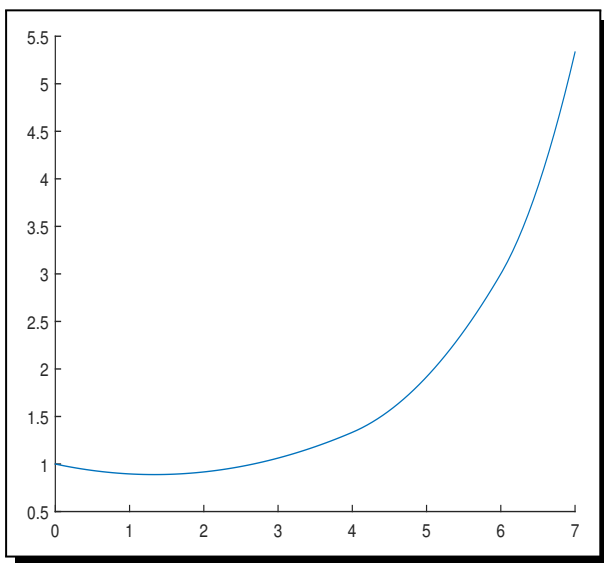


Figure 1. Approximation by OCICS with $\alpha = 0.5$ for Example 1

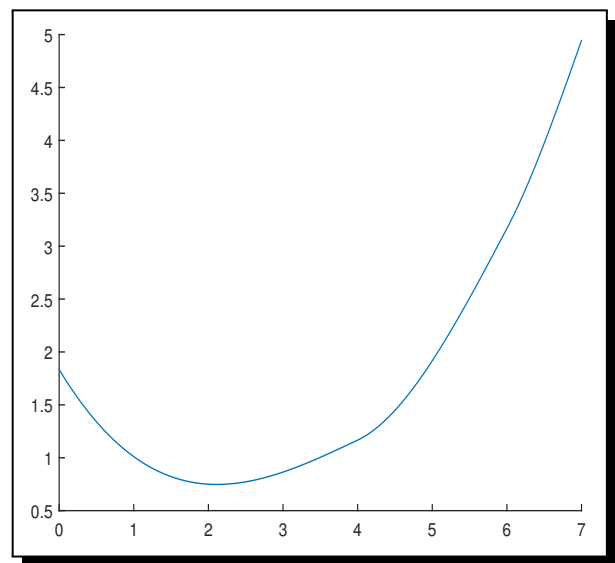


Figure 2. Approximation by CICS with $\alpha = 1$ for Example 1

Example 2. Next, we take $u(x) = 2 - \sqrt{x(2-x)}$, $0 \leq x \leq 2$ [6]. This function is approximated by the CICS on a uniform mesh in x , for $k = 10$, in Figure 3. In Figure 4, we consider the CICS for this function on a non-uniform grid $\Delta_{10} = \{0 < 0.05 < 0.1 < 0.4 < 0.7 < 1 < 1.3 < 1.6 < 1.9 < 1.95 < 2\}$. Near the end knots, the fitting result of the spline curve in Figure 4 is better than that of the spline curve in Figure 3. From this example, we can see that the constructed CICS possesses

convexity-preserving property and convergence. The purpose of this example is to observe the effects of the changes in the step size.

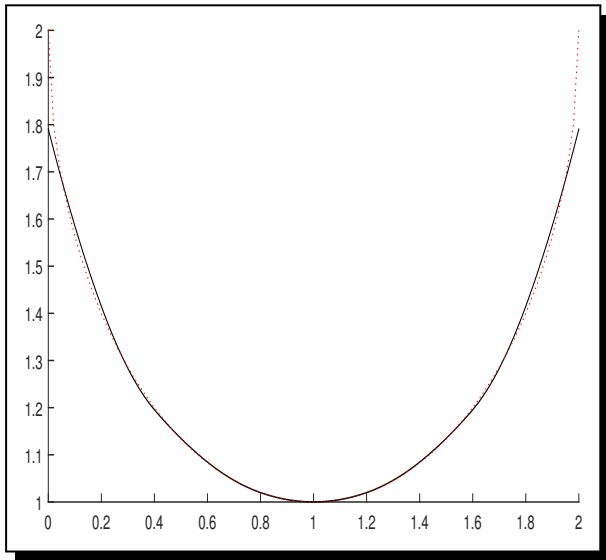


Figure 3. Approximation by OCICS for $u(x)$ on Δ_{10}

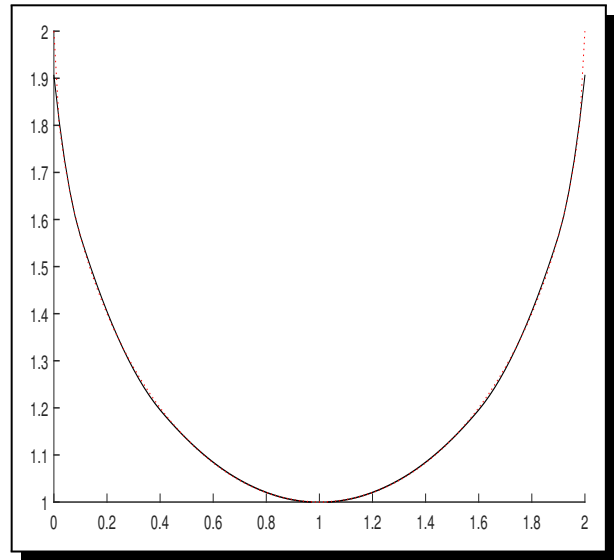


Figure 4. Approximation by OCICS for Akima's data

Example 3. Then, we consider the histogram $I = \{2.86, 1, 0.5, 1, 2, 2.86\}$ on $\Delta_6 = \{0 < 1 < 2 < 4 < 6 < 7 < 8\}$ which is in convex positions. Figure 5 shows that the fitting result is the same as that presented in [9].

Fortunately, for the data of the examples above, the conditions (25) are fulfilled.

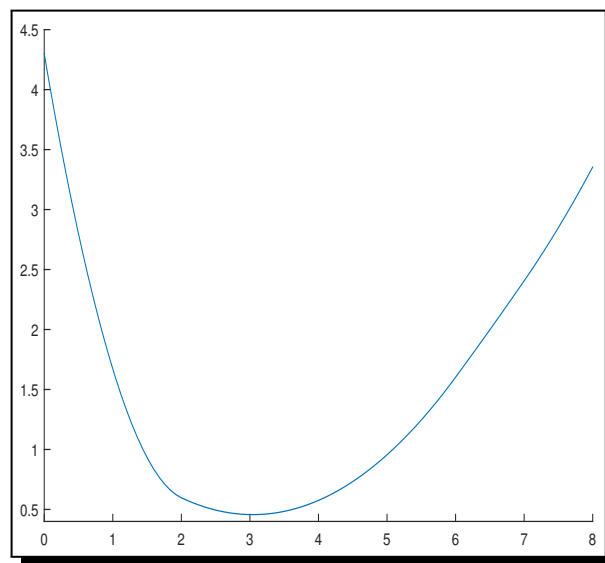


Figure 5. Approximation by OCICS for Example 3

Example 4. The data for the last example were taken from [1] (see Table 1), where the conditions (25) are not fulfilled. As for Akima's data, the solution of the system (11) does not satisfy the condition $m_i \leq m_{i+1}$, $i = 0, \dots, k-1$, so we can not construct the shape-preserving integro spline with the proposed method. Now, we can simply choose m_i by

$$m_i = \delta I_i, \quad i = 1, \dots, k-1,$$

because of Theorem 6. The remainder m_0 and m_k are obtained from (11) setting $i = 1, k-1$, respectively. The values of $S(x_i)$ are completely determined by (7) and (8). Figure 6 shows that the last integro cubic spline with $\alpha = \frac{1}{2}$ has a better convexity and monotonicity property.

Table 1. Akima's data [1]

x_i	0	2	3	5	6	8	9	11	12	14
I_i	10	10	10	10	10	10	10.5	15	50	

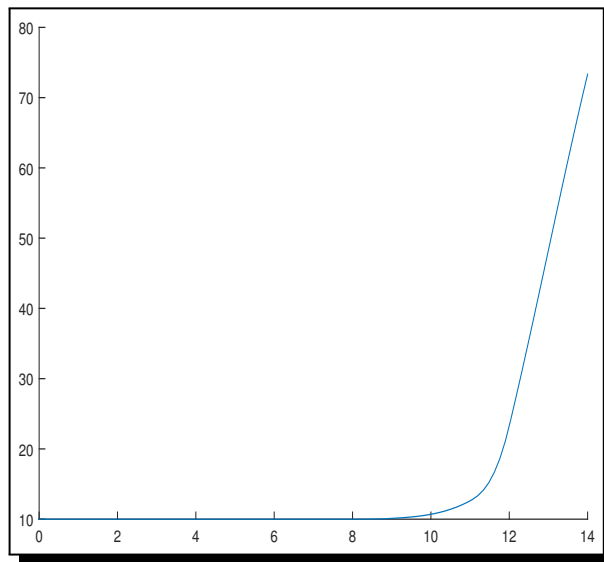


Figure 6. Approximation by OCICS for Example 3

5. Conclusion

In this paper, we derive a family of C^1 convex integro cubic splines based on sufficient conditions for convexity. We give some sufficient convexity and monotonicity conditions for constructed integro splines. The proposed family of splines has good approximation properties. The best convex integro spline is obtained when the α parameter is equal to $\frac{1}{2}$. The shape-preserving properties of splines are demonstrated by numerical examples.

Acknowledgement

The work was partially supported by the Foundation of Science and Technology of Mongolia (No. SST_18/2018).

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

- [1] H. Akima, A new method of interpolation and smooth curve fitting based on local procedures, *Journal of Association for Computing Machinery* **17** (1970), 589 – 602, DOI: 10.1145/321607.321609.
- [2] H. Behforooz, Approximation by integro cubic splines, *Applied Mathematics and Computation* **175** (2006), 8 – 15, DOI: 10.1016/j.amc.2005.07.066.
- [3] M. Fischer and P. Oja, Monotonicity preserving rational spline histopolation, *Journal of Computational and Applied Mathematics* **175** (2005), 195 – 208, DOI: 10.1016/j.cam.2004.05.009.
- [4] M. Fischer, P. Oja and H. Trossmann, Comonotone shape-preserving spline histopolation, *Journal of Computational and Applied Mathematics* **200** (2007), 127 – 139, DOI: 10.1016/j.cam.2005.12.010.
- [5] M. Kaykobad, Positive solutions of positive linear systems, *Linear Algebra and its Applications* **64** (1985), 133 – 140, DOI: 10.1016/0024-3795(85)90271-X.
- [6] T.-W. Kim and B. Kvasov, A shape-preserving approximation by weighted cubic splines, *Journal of Computational and Applied Mathematics* **236** (2012), 4383 – 4397, DOI: 10.1016/j.cam.2012.04.001.
- [7] B. Mulansky and J. W. Schmidt, Convex interval interpolation using a three-term staircase algorithm, *Numerische Mathematik* **82** (1999), 313 – 337, DOI: 10.1007/s002110050421.
- [8] E. Neuman, Uniform approximation by some Hermite interpolating splines, *Journal of Computational and Applied Mathematics* **4** (1978), 7 – 9, <https://core.ac.uk/download/pdf/81978018.pdf>.
- [9] J. W. Schmidt and W. Heß, Shape preserving C^2 -spline histopolation, *Journal of Approximation Theory* **75** (1993), 325 – 345, DOI: 10.1006/jath.1993.1106.
- [10] J. W. Schmidt and W. Heß, An allways successful method in univariate convex C^2 -interpolation, *Numerische Mathematik* **71** (1995), 237 – 252, DOI: 10.1007/s002110050143.
- [11] J. W. Schmidt, Staircase algorithm and construction of convex spline interpolats up to the continuity C^3 , *Computers & Mathematics with Applications* **31** (1996), 67 – 79, DOI: 10.1016/0898-1221(95)00218-9.
- [12] T. Zhanlav, Shape preserving properties of C^1 cubic spline approximations, *Scientific Transaction NUM* **7** (2000), 14 – 20, URL https://www.researchgate.net/profile/T_Zhanlav/publication/307632050_Shape_preserving_properties_of_C1_cubic_spline_approximations/links/5f555a5a458515e96d35c24f/Shape-preserving-properties-of-C1-cubic-spline-approximations.pdf.
- [13] T. Zhanlav, Shape preserving properties of some C^2 cubic spline approximations, *Scientific Transaction NUM* **7** (2000), 21 – 35, URL https://www.researchgate.net/profile/T_Zhanlav/publication/307631606_Shape_preserving_properties_of_some_C2_cubic_spline_approximations/links/5f555c4c92851c250b995dc1/Shape-preserving-properties-of-some-C2-cubic-spline-approximations.pdf.

- [14] T. Zhanlav and R. Mijiddorj, Convexity and monotonicity properties of the local integro cubic spline, *Applied Mathematics and Computation* **293** (2017), 131–137, DOI: 10.1016/j.amc.2016.08.017.
- [15] P. Ženčák, The convex interpolation of histogram by polynomial splines: The existence theorem, *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **41** (2002), 175–182, URL: https://dml.cz/bitstream/handle/10338.dmlcz/120449/Acta01om_41-2002-1_16.pdf.

Integro Cubic Splines on Non-Uniform Grids and Their Properties

T. Zhanlav¹ and R. Mijiddorj^{2,*}

¹*Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia.*

²*Department of Informatics, Mongolian National University of Education, Ulaanbaatar, Mongolia.*

Received 3 September 2020; Accepted (in revised version) 25 December 2020.

Abstract. Integro cubic splines on a non-uniform grid using the integral values of an unknown function are constructed. We establish a consistency relation for integro cubic spline and derive a local integro cubic spline on non-uniform partitions. Approximation and convexity properties of the local integro cubic splines are also studied.

AMS subject classifications: 65D05, 65D07

Key words: Integro cubic spline, local construction, non-uniform grid, error analysis.

1. Introduction

Researchers from our university investigating the location of a robot equipped with a rotary encoder device, wanted us to determine the velocity $v(t)$ of the wheel of the rotary encoder, which registers the time series when it runs a constant distance or Area= \square , cf. Fig. 1.

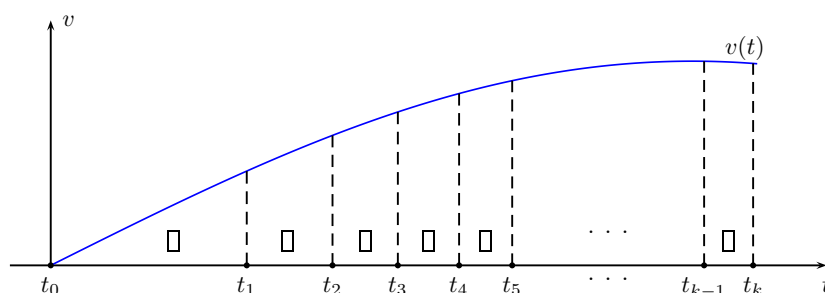


Figure 1: Time series registered with a rotary encoder.

*Corresponding author. Email address: mijiddorj@msue.edu.mn (R. Mijiddorj)

Of course, the problem of determining $v(t)$ concerns the histo-spline and the integro spline. There are many papers constructing integro splines [1–4, 10, 16, 17], but they are mainly focus on uniform partitions. Wu and Zhang [9] suggested an integro quadratic spline and Kirsiaed *et al.* [7] constructed a cubic spline histopolation on a non-uniform partition and studied its approximating properties. However, such splines do not solve the above problem, since the corresponding construction requires full information for a specified interval $[t_0, t_k]$. We have to provide real-time velocity $v(t)$ so that in our situation the local construction of the integro spline is more suitable. It is well-known [14] that the local construction of an integro spline has a lower computational cost than the constructions based on solving systems of linear equations.

This work is organised as follows. In Section 2, we consider an integro cubic spline on a non-uniform grid. Section 3 discusses the local construction of the integro cubic spline. In Section 4, we study the errors and convexity property of the spline proposed. Numerical examples presented in the last section illustrate the accuracy of the methods used.

2. Construction of Integro Cubic Splines

Let $\mathcal{T}_k := \{t_0 < t_1 < \dots < t_k\}$ be a non-uniform partition of $[t_0, t_k]$ and $h_{i+1} = t_{i+1} - t_i$ are the step sizes. We have no information about the values of the function $v(t)$. However, it is known that for any subinterval $[t_i, t_{i+1}]$ the area \square under the graph of $v(t)$ is the same.

The problem of construction of integro cubic spline consists in finding an $S(t)$ such that the following conditions hold:

- (i) On each subinterval $[t_i, t_{i+1}]$, $S(t)$ coincides with a polynomial of degree three.

$$(ii) \frac{1}{h_i} \int_{t_{i-1}}^{t_i} S(t)dt = \frac{1}{h_i} \int_{t_{i-1}}^{t_i} v(t)dt = I_i, \quad i = 1, 2, \dots, k.$$

It follows from Fig. 1 and condition (ii) that $\square = h_i I_i$. We denote by $S_3(\mathcal{T}_k)$ the space of cubic splines over the partition \mathcal{T}_k , i.e.

$$S_3(\mathcal{T}_k) = \{p(x) | p(x) \in C^2[t_0, t_k]\},$$

where $p(x)$ is a polynomial of degree at most three on \mathcal{T}_k . According to [11], the elements $S \in S_3(\mathcal{T}_k)$ can be represented in one of the forms

$$S(t) = (1 - \xi)^2(1 + 2\xi)S_{i-1} + \xi^2(3 - 2\xi)S_i + h_i \xi(1 - \xi) \{(1 - \xi)S'_{i-1} - \xi S'_i\}, \quad (2.1)$$

or

$$S(t) = (1 - \xi)S_{i-1} + \xi S_i - \frac{h_i^2}{6} \xi(1 - \xi) [(2 - \xi)S''_{i-1} + (1 + \xi)S''_i], \quad (2.2)$$

$$t \in [t_{i-1}, t_i], \quad \xi = \frac{t - t_{i-1}}{h_i}, \quad \xi \in [0, 1],$$

where $S_i = S(t_i)$, $S'_i = S'(t_i)$, and $S''_i = S''(t_i)$. Using (2.1) and (2.2) in (ii) yields

$$S_{i-1} + S_i = 2I_i - \frac{h_i}{6} (S'_{i-1} - S'_i), \quad i = 1, 2, \dots, k, \quad (2.3a)$$

$$S_{i-1} + S_i = 2I_i + \frac{h_i^2}{12} (S''_{i-1} + S''_i), \quad i = 1, 2, \dots, k. \quad (2.3b)$$

Subtracting (2.3a) from (2.3b), we obtain

$$S''_{i-1} + S''_i = \frac{2}{h_i} (S'_i - S'_{i-1}), \quad i = 1, 2, \dots, k. \quad (2.4)$$

The Eq. (2.4) implies

$$\begin{aligned} & h_i^2 S''_{i-1} + (h_i^2 - h_{i+1}^2) S''_i - h_{i+1}^2 S''_{i+1} \\ &= 2(-h_i S'_{i-1} + (h_i + h_{i+1}) S'_i - h_{i+1} S'_{i+1}), \end{aligned} \quad (2.5)$$

and consequently,

$$\mu_i S''_{i-1} + S''_i + \lambda_i S''_{i+1} = \frac{2}{h_i + h_{i+1}} (S'_{i+1} - S'_{i-1}), \quad i = 1, 2, \dots, k-1, \quad (2.6)$$

where

$$\mu_i = \frac{h_i}{h_i + h_{i+1}}, \quad \lambda_i = 1 - \mu_i.$$

Replacing the index i by $i + 1$ in (2.3b) and adding/subtracting the resulting equation to/from (2.3a) gives

$$I_i + I_{i+1} = \frac{1}{2} (S_{i-1} + 2S_i + S_{i+1}) - \frac{1}{24} (h_i^2 S''_{i-1} + (h_i^2 + h_{i+1}^2) S''_i + h_{i+1}^2 S''_{i+1}), \quad (2.7)$$

$$I_{i+1} - I_i = \frac{1}{2} (S_{i+1} - S_{i-1}) + \frac{1}{24} (h_i^2 S''_{i-1} + (h_i^2 - h_{i+1}^2) S''_i - h_{i+1}^2 S''_{i+1}). \quad (2.8)$$

It follows from (2.2) that

$$S'''(t_i - 0) = \frac{S''_i - S''_{i-1}}{h_i}, \quad S'''(t_i + 0) = \frac{S''_{i+1} - S''_i}{h_{i+1}}. \quad (2.9)$$

Using (2.9) and the Taylor expansions of S_{i-1} and S_{i+1} in (2.7) and (2.8), we get

$$\begin{aligned} I_i + I_{i+1} &= 2S_i + \frac{h_{i+1} - h_i}{2} S'_i + \frac{h_i^2 + h_{i+1}^2}{6} S''_i + \frac{1}{24} (h_{i+1}^3 S'''(t_i + 0) - h_i^3 S'''(t_i - 0)), \\ I_{i+1} - I_i &= \frac{h_i + h_{i+1}}{2} S'_i + \frac{h_{i+1}^2 - h_i^2}{6} S''_i + \frac{1}{24} (h_i^3 S'''(t_i - 0) + h_{i+1}^3 S'''(t_i + 0)), \\ & i = 1, 2, \dots, k-1. \end{aligned} \quad (2.10)$$

Substituting (2.9) into (2.10) yields

$$S'_i = \frac{2}{h_i + h_{i+1}}(I_{i+1} - I_i) + \frac{1}{12}(h_i\mu_i S''_{i-1} + 3(h_i - h_{i+1})S''_i - h_{i+1}\lambda_i S''_{i+1}), \quad (2.11)$$

and using (2.5) in (2.11), we get

$$\begin{aligned} & \mu_i S'_{i-1} + 5S'_i + \lambda_i S'_{i+1} \\ &= \frac{12}{h_i + h_{i+1}}(I_{i+1} - I_i) + (h_i - h_{i+1})S''_i, \quad i = 1, 2, \dots, k-1. \end{aligned} \quad (2.12)$$

Besides, substituting (2.9) into (2.6), we obtain

$$S''_i \approx \frac{S'_{i+1} - S'_{i-1}}{h_i + h_{i+1}} \quad (2.13)$$

with accuracy $\mathcal{O}(h_i - h_{i+1})$. We now can replace the term S'' in (2.12) by (2.13), thus obtaining

$$\begin{aligned} & (2 - 3\lambda_i)S'_{i-1} + 5S'_i + (3\lambda_i - 1)S'_{i+1} \\ &= \frac{12}{h_i + h_{i+1}}(I_{i+1} - I_i), \quad i = 1, 2, \dots, k-1. \end{aligned} \quad (2.14)$$

Note that, using the continuity property of (2.1) and (2.3a), (2.13) we also arrive at (2.14). Thus for uniform partitions, the Eqs. (2.14) can be exact consistency relations [2], but for integro cubic splines on non-uniform partition they are only approximate ones. The Eqs. (2.14) and the end conditions S'_0, S'_k forms a closed system. In order to study the solvability of this system, we calculate the term

$$r_i = a_{ii} - \sum_{j \neq i} |a_{ij}|,$$

where a_{ij} are entries of the matrix of (2.14). There are three cases — viz.

1. If $0 \leq \lambda_i \leq 1/3$, then $3\lambda_i - 1 \leq 0$, $2 - 3\lambda_i \geq 0$ and $r_i = 2(1 + 3\lambda_i) \geq 2$.
2. If $1/3 \leq \lambda_i \leq 2/3$, then $3\lambda_i - 1 \geq 0$, $2 - 3\lambda_i \geq 0$, and $r_i = 4$.
3. If $2/3 \leq \lambda_i \leq 1$, then $3\lambda_i - 1 \geq 0$, $2 - 3\lambda_i \leq 0$, and $r_i = 8 - 6\lambda_i \geq 2$.

The matrix of the system (2.14) is diagonally dominant and hence, the system (2.14) along with the end conditions S'_0 and S'_k has a unique solution. Unlike the interpolatory cubic splines, besides of S'_0 and S'_k , an additional end condition S_0 or S_k is needed. Thus, values S_i and S'_i for $i = 0, 1, \dots, k$ are found from (2.3a) and (2.14) provided that S'_0, S'_k , and S_0 (or S_k) are known. Then the integro cubic spline $S(t)$ is constructed using its piecewise polynomial presentation (2.1).

For the integro spline, more suitable end conditions are the third derivative continuity conditions (or not-a-knot end conditions), viz.

$$S'''(t_i - 0) = S'''(t_i + 0), \quad i = 1, 2, k - 2, k - 1.$$

They can be also rewritten in terms S_i'' as

$$\lambda_i S_{i-1}'' - S_i'' + \mu_i S_{i+1}'' = 0, \quad i = 1, 2, k - 2, k - 1. \tag{2.15}$$

Well-known continuity property of (2.2) has the form

$$\frac{6}{h_i h_{i+1}} (\lambda_i S_{i-1} - S_i + \mu_i S_{i+1}) = \mu_i S_{i-1}'' + 2S_i'' + \lambda_i S_{i+1}'', \quad i = 1, 2, \dots, k - 1, \tag{2.16}$$

and (2.3b) as well as with three of (2.15) together consist of $2k + 2$ equations with unknowns S_0, S_1, \dots, S_k , and $S_0'', S_1'', \dots, S_k''$. Solving this system of linear equations, we can also construct the integro cubic spline by (2.2).

3. Construction of a Local Integro Cubic Spline

The continuity property (2.16) and (2.3b) give

$$S_i = \lambda_i I_i + \mu_i I_{i+1} - \frac{h_i h_{i+1}}{24} (\mu_i S_{i-1}'' + 3S_i'' + \lambda_i S_{i+1}''), \quad i = 1, 2, \dots, k - 1. \tag{3.1}$$

The Eqs. (2.12) and (3.1) allow us to easily construct an integro spline if $S_0'', S_1'', \dots, S_k''$ are given. We call the grid \mathcal{T}_k almost uniform if $h_{i+1} - h_i = \mathcal{O}(\bar{h}^2), i = 1, 2, \dots, k - 1$, where $\bar{h} = \max_{1 \leq i \leq k} \{h_i\}$. Therefore, on an almost uniform grid the relation (2.13) is valid with accuracy $\mathcal{O}(\bar{h}^2)$. Note that the second term on the right hand side of (2.11) equals to $\mathcal{O}(\bar{h}^2)$. Thus ignoring it, we obtain the approximate formula

$$S_i' \approx \frac{2}{h_i + h_{i+1}} (I_{i+1} - I_i). \tag{3.2}$$

However, we need a more accurate formula than (3.2). Therefore, our first task is to find S_i'' with accuracy $\mathcal{O}(\bar{h}^2)$. To this end, we use (2.11) in (2.13). As the result, we have

$$S_i'' = \frac{2}{h_i + h_{i+1}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) + \frac{1}{12(h_i + h_{i+1})} (A_{i+1} - A_{i-1}), \tag{3.3}$$

where

$$A_i = h_i \mu_i S_{i-1}'' + 3(h_i - h_{i+1}) S_i'' - h_{i+1} \lambda_i S_{i+1}''.$$

According to (2.9), we have

$$\begin{aligned} A_{i+1} - A_{i-1} = & 4(h_{i+1} + h_i - h_{i-1} - h_{i+2}) S_i'' + [4h_{i+1}(h_{i+1} - h_{i+2}) S'''(t_i + 0) \\ & - 4h_i(h_{i-1} - h_i) S'''(t_i - 0) + D_{i-1} - D_{i+1}], \end{aligned} \tag{3.4}$$

where

$$D_i = \frac{h_i^3 S'''(t_i - 0) + h_{i+1}^3 S'''(t_i + 0)}{h_i + h_{i+1}}.$$

Assume that

$$D_{i+1} - D_{i-1} = \mathcal{O}(\bar{h}^3), \quad h_{i+1} - h_i = \mathcal{O}(\bar{h}^2).$$

Then the expression in the square bracket in (3.4) is $\mathcal{O}(\bar{h}^3)$. Substituting (3.4) into (3.3) and ignoring the term of $\mathcal{O}(\bar{h}^2)$ gives

$$\tilde{S}_i'' = \frac{6}{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right), \quad i = 2, 3, \dots, k-2. \quad (3.5)$$

The remainders \tilde{S}_i'' , $i = 1, 2, k-2, k-1$ are determined by the Eq. (2.15). Substituting (3.5) into (2.11) and (3.1), we obtain approximations \tilde{S}'_i and \tilde{S}_i respectively at the interior knots. The remainders \tilde{S}_i are uniquely determined from (2.3b) by setting $i = 1, k$, and \tilde{S}'_i are determined from (2.12) by setting $i = 1, k-1$. For uniform grids, the explicit formulae (2.11), (3.1), and (3.5) coincide with the ones in [14]. In order to construct a spline of $C^2[t_0, t_k]$ based on these approximate values $\tilde{S}_i, \tilde{S}'_i$, and \tilde{S}_i'' , we can use the following B -spline representation:

$$\tilde{S}(t) = \sum_{i=-1}^{k+1} \tilde{\alpha}_i B_i(t) \quad (3.6)$$

of the cubic splines — cf. [11]. The coefficients in (3.6) have the form

$$\begin{aligned} \tilde{\alpha}_{-1} &= \tilde{S}_0, \\ \tilde{\alpha}_i &= \tilde{S}_i + \frac{h_{i+1} - h_i}{3} \tilde{S}'_i - \frac{h_i h_{i+1}}{6} \tilde{S}_i'', \quad i = 0, 1, \dots, k, \\ \tilde{\alpha}_{k+1} &= \tilde{S}_k, \end{aligned} \quad (3.7)$$

where $t_{-3} = t_{-2} = t_{-1} = t_0$ and $t_k = t_{k+1} = t_{k+2} = t_{k+3}$ [12]. Substituting the approximate values of $\tilde{S}_i, \tilde{S}'_i$, and \tilde{S}_i'' into (3.7), we obtain the coefficients $\tilde{\alpha}_i$ in (3.6). Thus we found a local integro cubic spline in term of B -spline representation and $\tilde{S}(t) \in C^2[t_0, t_k]$, cf. Fig. 2. For uniform grids, the coefficients $\tilde{\alpha}_i$ coincide with the ones in [13].

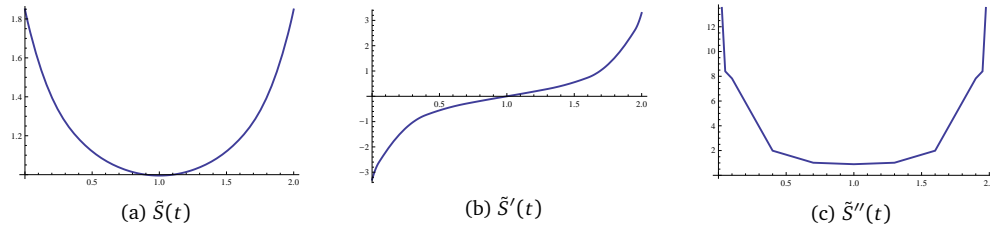


Figure 2: The local integro cubic spline $\tilde{S}(t) \in C^2[t_0, t_k]$ for $v(t) = 2 - \sqrt{t(2-t)}$.

4. Error Analysis and Convexity of Local Integro Cubic Splines

Let us consider the case of the data set I_i obtained from a smooth function $v(t) \in C^5[t_0, t_k]$. Similar to the previous considerations, we use the Taylor expansion of function $v(t)$ in (ii) to obtain

$$I_{i+1} = v_i + \frac{h_{i+1}}{2}v'_i + \frac{h_{i+1}^2}{6}v''_i + \frac{h_{i+1}^3}{24}v'''_i + \mathcal{O}(h_{i+1}^4), \quad (4.1)$$

$$I_{i+1} = v_{i+1} - \frac{h_{i+1}}{2}v'_{i+1} + \frac{h_{i+1}^2}{6}v''_{i+1} - \frac{h_{i+1}^3}{24}v'''_{i+1} + \mathcal{O}(h_{i+1}^4),$$

$$I_i = v_i - \frac{h_i}{2}v'_i + \frac{h_i^2}{6}v''_i - \frac{h_i^3}{24}v'''_i + \mathcal{O}(h_i^4), \quad (4.2)$$

where $v_i = v(t_i)$, $v'_i = v'(t_i)$, $v''_i = v''(t_i)$, and $v'''_i = v'''(t_i)$. Using the Taylor expansion and replacing i by $i + 1$ in (4.1) gives

$$\begin{aligned} I_{i+2} = v_i + \frac{2h_{i+1} + h_{i+2}}{2}v'_i + \frac{3h_{i+1}^2 + 3h_{i+1}h_{i+2} + h_{i+2}^2}{6}v''_i \\ + \frac{4h_{i+1}^3 + 6h_{i+1}^2h_{i+2} + 4h_{i+1}h_{i+2}^2 + h_{i+2}^3}{24}v'''_i + \mathcal{O}(\bar{h}^4). \end{aligned} \quad (4.3)$$

Analogously, replacing i by $i - 1$ in (4.2) gives

$$\begin{aligned} I_{i-1} = v_i - \frac{2h_i + h_{i-1}}{2}v'_i + \frac{3h_i^2 + 3h_{i-1}h_i + h_{i-1}^2}{6}v''_i \\ - \frac{4h_i^3 + 6h_i^2h_{i-1} + 4h_{i-1}^2h_i + h_{i-1}^3}{24}v'''_i + \mathcal{O}(\bar{h}^4). \end{aligned}$$

Subtracting (4.1) from (4.3), we have

$$\begin{aligned} \frac{2}{h_{i+1} + h_{i+2}}(I_{i+2} - I_{i+1}) = v'_i + \frac{2h_{i+1} + h_{i+2}}{3}v''_i \\ + \frac{3h_{i+1}^2 + 3h_{i+1}h_{i+2} + h_{i+2}^2}{12}v'''_i + \mathcal{O}(\bar{h}^3). \end{aligned} \quad (4.4)$$

Similar considerations show that

$$\begin{aligned} \frac{2}{h_i + h_{i+1}}(I_{i+1} - I_i) = v'_i + \frac{h_{i+1} - h_i}{3}v''_i + \frac{h_i^3 + h_{i+1}^3}{12(h_i + h_{i+1})}v'''_i \\ + \frac{h_{i+1}^4 - h_i^4}{60(h_i + h_{i+1})}v_i^{(4)} + \mathcal{O}(\bar{h}^4), \end{aligned} \quad (4.5)$$

$$\begin{aligned} \frac{2}{h_{i-1} + h_i}(I_i - I_{i-1}) = v'_i - \frac{2h_i + h_{i-1}}{3}v''_i \\ + \frac{3h_i^2 + 3h_ih_{i-1} + h_{i-1}^2}{12}v'''_i + \mathcal{O}(\bar{h}^3). \end{aligned} \quad (4.6)$$

We are ready to prove the following theorem.

Theorem 4.1. Let $v(t) \in C^5[t_0, t_k]$, \tilde{S}_i be determined by (2.3b), (3.1), \tilde{S}'_i by (2.11), (2.12), \tilde{S}''_i by (2.15), (3.5) and the grid is almost uniform. Then the following estimates

$$\tilde{S}''_i - v''_i = \mathcal{O}(\bar{h}^2), \quad i = 0, 1, \dots, k, \tag{4.7}$$

$$\tilde{S}'_i - v'_i = \mathcal{O}(\bar{h}^4), \quad i = 0, 1, \dots, k, \tag{4.8}$$

$$\tilde{S}_i - v_i = \mathcal{O}(\bar{h}^4), \quad i = 0, 1, \dots, k \tag{4.9}$$

hold.

Proof. First, we establish the estimate (4.7). It follows from (4.4) and (4.6) that

$$\frac{6}{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) = v''_i + \frac{b_i}{12} v'''_i + \mathcal{O}(\bar{h}^2), \tag{4.10}$$

where

$$b_i = \frac{3(h_{i+1}^2 - h_i^2) + 3(h_{i+1}h_{i+2} - h_i h_{i-1}) + h_{i+2}^2 - h_{i-1}^2}{\hat{h}_i}. \tag{4.11}$$

On an almost uniform grid, the coefficients in v'''_i are of $\mathcal{O}(h_{i+1} - h_i) = \mathcal{O}(\bar{h}^2)$. The relations (3.5) and (4.10) show that

$$\tilde{S}''_i - v''_i = \mathcal{O}(\bar{h}^2), \quad i = 2, 3, \dots, k - 2.$$

For remaining i , the estimates (4.7) follow from the relation

$$\lambda_i (\tilde{S}''_{i-1} - v''_{i-1}) - (\tilde{S}''_i - v''_i) + \mu_i (\tilde{S}''_{i+1} - v''_{i+1}) = -(\lambda_i v''_{i-1} - v''_i + \mu_i v''_{i+1}) = \mathcal{O}(\bar{h}^2),$$

which is a consequence of the not-a-knot relation. If $v(t) \in C^5$, then using (4.5) in (2.11), we get

$$\begin{aligned} \tilde{S}'_i - v'_i &= \frac{1}{12} (h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(h_i - h_{i+1}) (\tilde{S}''_i - v''_i) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) \\ &\quad + \frac{2}{h_i + h_{i+1}} (I_{i+1} - I_i) - v'_i + \frac{1}{12} (h_i \mu_i v'''_{i-1} + 3(h_i - h_{i+1}) v'''_i - h_{i+1} \lambda_i v'''_{i+1}) \\ &= \frac{1}{12} (h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) + 3(h_i - h_{i+1}) (\tilde{S}''_i - v''_i) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1})) + \mathcal{O}(\bar{h}^4). \end{aligned}$$

Using (4.10) and (4.11) for $i - 1$, and $i + 1$ in the last equality, we get

$$h_i \mu_i (\tilde{S}''_{i-1} - v''_{i-1}) - h_{i+1} \lambda_i (\tilde{S}''_{i+1} - v''_{i+1}) = \frac{h_i \mu_i b_{i-1} - h_{i+1} \lambda_i b_{i+1}}{12} v_i^{(3)} + \mathcal{O}(\bar{h}^4).$$

Therefore, we have

$$\tilde{S}'_i - v'_i = \mathcal{O}(\bar{h}^4), \quad i = 3, 4, \dots, k - 2.$$

For the remaining i , the proof of the estimates (4.8) is based on the relations

$$\mu_i (\tilde{S}'_{i-1} - v'_{i-1}) + 5(\tilde{S}'_i - v'_i) + \lambda_i (\tilde{S}'_{i+1} - v'_{i+1})$$

$$\begin{aligned}
&= \frac{12}{h_i + h_{i+1}}(I_{i+1} - I_i) + (h_i - h_{i+1})(\tilde{S}_i'' - v_i'') + (h_i - h_{i+1})v_i'' - \mu v_{i-1}' - 5v_i' - \lambda_i v_{i+1}' \\
&= \mathcal{O}(\bar{h}^4),
\end{aligned}$$

which are the consequence of (2.12). Further, it follows from (3.1), (4.1), and (4.2) that

$$\begin{aligned}
\tilde{S}_i - v_i &= \lambda_i I_i + \mu_i I_{i+1} - \frac{h_i h_{i+1}}{24} \left(\mu_i (\tilde{S}_{i-1}'' - v_{i-1}'') + 3(\tilde{S}_i'' - v_i'') + \lambda_i (\tilde{S}_{i+1}'' - v_{i+1}'') \right) \\
&\quad - v_i - \frac{h_i h_{i+1}}{24} (\mu_i v_{i-1}'' + 3v_i'' + \lambda_i v_{i+1}'') \\
&= \frac{h_i h_{i+1}}{24} \left(\mu_i (\tilde{S}_{i-1}'' - v_{i-1}'') + 3(\tilde{S}_i'' - v_i'') + \lambda_i (\tilde{S}_{i+1}'' - v_{i+1}'') \right) + \mathcal{O}(\bar{h}^4). \quad (4.12)
\end{aligned}$$

Consequently, (4.12) implies (4.9) for $i = 1, 2, \dots, k-1$. If $i = 0$ or $i = k$, the estimations (4.9) follow directly from (2.3). \square

Remark 4.1. The proof of Theorem 4.1 show that on any non-uniform grid we have

$$\tilde{S}_i'' - v_i'' = \mathcal{O}(\bar{h}), \quad \tilde{S}_i' - v_i' = \mathcal{O}(\bar{h}^2), \quad \tilde{S}_i - v_i = \mathcal{O}(\bar{h}^3).$$

Remark 4.2. The estimates (4.7)-(4.9) are pointwise ones. However, [12, Theorem 2] and Theorem 4.1 allow to establish global approximation errors. More exactly, if the grid is almost uniform, then under the assumptions of Theorem 4.1, the local integro cubic spline $\tilde{S}(t)$ in (3.6) satisfies the estimates

$$\|\tilde{S}^{(r)}(t) - v^{(r)}(t)\|_\infty = \mathcal{O}(\bar{h}^{4-r}), \quad r = 0, 1, 2.$$

Theorem 4.2. *If the grid is almost uniform, then under the assumptions of Theorem 4.1, the spline $S(t)$ in (2.2) with the coefficients \tilde{S}_i and \tilde{S}_i'' satisfies the estimates*

$$\|S^{(r)}(t) - v^{(r)}(t)\|_\infty = \mathcal{O}(\bar{h}^{4-r}), \quad r = 0, 1, 2. \quad (4.13)$$

Proof. First of all, we show (4.13) for $r = 2$. Exploiting the second derivative of (2.2), we consider the equation

$$\begin{aligned}
S''(t) - v''(t) &= (1-t)(S_i'' - v_i'') + t(S_{i+1}'' - v_{i+1}'') \\
&\quad + [(1-t)v_i'' + tv_{i+1}'' - v''(t)]
\end{aligned} \quad (4.14)$$

on the interval $[t_i, t_{i+1}]$. The Taylor expansions

$$\begin{aligned}
v_i'' &= v''(t) - v'''(t)th_{i+1} + \frac{v^{(4)}(t)}{2}t^2h_{i+1}^2 + \mathcal{O}(h_{i+1}^3), \\
v_{i+1}'' &= v''(t) + v'''(t)(1-t)h_{i+1} + \frac{v^{(4)}(t)}{2}(1-t)^2h_{i+1}^2 + \mathcal{O}(h_{i+1}^3),
\end{aligned}$$

lead to the relation

$$(1-t)v_i'' + tv_{i+1}'' - v''(t) = \mathcal{O}(h_{i+1}^2).$$

Combining it with (4.7), (4.8) in (4.14) shows that

$$S''(t) - v''(t) = \mathcal{O}(\bar{h}^2), \quad t \in [t_i, t_{i+1}], \quad i = 0, 1, \dots, k-1.$$

Thus (4.13) is proven for $r = 2$. Now we can use (2.2) and obtain

$$\begin{aligned} S(t) - v(t) &= (1-t)(S_i - v_i) + t(S_{i+1} - v_{i+1}) - \frac{h_{i+1}^2}{6}t(1-t) \\ &\quad \times \left[(2-t)(S_i'' - v_i'') + (1+t)(S_{i+1}'' - v_{i+1}'') \right] + E_i, \end{aligned} \quad (4.15)$$

where

$$E_i = (1-t)v_i + tv_{i+1} - \frac{h_{i+1}^2}{6}t(1-t) \left[(2-t)v_i'' + (1-t)v_{i+1}'' \right] - v(t).$$

As before, one can use the Taylor expansion of $v \in C^5$ to check that

$$E_i = \mathcal{O}(\bar{h}^4). \quad (4.16)$$

The relations (4.16), (4.7), (4.8) and (4.15) show that $S(t) - v(t) = \mathcal{O}(\bar{h}^4)$.

From the first derivative of (2.2), we get

$$S_i - S_{i-1} = \frac{h_i}{2}(S_i' + S_{i-1}') - \frac{h_i^2}{12}(S_i'' - S_{i-1}''),$$

which yields

$$S'(t) = \frac{1}{2}(S_i' + S_{i+1}') - \frac{h_{i+1}}{4} \left[(1-4t+2t^2)S_i'' + (1-2t^2)S_{i+1}'' \right].$$

Considering the residue

$$\begin{aligned} S'(t) - v'(t) &= \frac{(S_i' - v_i') + (S_{i+1}' - v_{i+1}')}{2} - \frac{h_{i+1}}{4} \\ &\quad \times \left[(1-4t+2t^2)(S_i'' - v_i'') + (1-2t^2)(S_{i+1}'' - v_{i+1}'') \right] + F_i, \end{aligned} \quad (4.17)$$

where

$$F_i = \frac{1}{2}(v_i' + v_{i+1}') - \frac{h_{i+1}}{4} \left[(1-4t+2t^2)v_i'' + (1-2t^2)v_{i+1}'' \right] - v'(t)$$

and using the Taylor expansion of $v \in C^4$ shows that

$$F_i = \mathcal{O}(\bar{h}^3). \quad (4.18)$$

Combining (4.7), (4.8), (4.18) and (4.17) produces (4.13) for $r = 1$. □

Remark 4.3. If the grid is almost uniform, then under the assumptions of Theorems 4.1 and 4.2 the following estimates

$$\|\tilde{S}^{(r)}(t) - S^{(r)}(t)\|_\infty = \mathcal{O}(\bar{h}^{4-r}), \quad r = 0, 1, 2$$

hold.

Now, we study the convexity of the local integro cubic splines. As usual, given data I_i are called convex if

$$a_i - a_{i-1} \geq 0, \quad i = 2, 3, \dots, k-1, \tag{4.19}$$

where $a_i = 2(I_{i+1} - I_i)/(h_i + h_{i+1})$. This definition agrees with the definition of the convex data — cf. [5, 6, 15].

Theorem 4.3. *Let the data I_i be convex and*

$$\hat{h}_i := \frac{h_{i-1} + 2h_i + 2h_{i+1} + h_{i+2}}{3}.$$

If

$$\hat{h}_2 a_2 + \hat{h}_3 a_3 \geq \hat{h}_3 a_1 + \hat{h}_2 a_4, \tag{4.20a}$$

$$\hat{h}_{k-2} a_{k-4} + \hat{h}_{k-3} a_{k-1} \geq \hat{h}_{k-3} a_{k-3} + \hat{h}_{k-2} a_{k-2}, \tag{4.20b}$$

then $\tilde{S}''(t) > 0$ for all $t \in [t_0, t_k]$, i.e. $\tilde{S}(t)$ is convex on $[t_0, t_k]$.

Proof. According to (3.5) and (4.19), we have

$$\tilde{S}_i'' \geq 0, \quad i = 2, 3, \dots, k-2.$$

The not-a-knot end point condition (2.15) implies

$$\tilde{S}_0'' = \frac{\tilde{S}_1'' - \mu_1 \tilde{S}_2''}{\lambda_1} = \frac{(1 - \mu_1 \lambda_2) \tilde{S}_2'' - \mu_2 \tilde{S}_3''}{\lambda_1 \lambda_2}. \tag{4.21}$$

Since $1 - \mu_1 \lambda_2 = \mu_2 + \lambda_1 \lambda_2 > \mu_2$ and $\tilde{S}_2'' \geq \tilde{S}_3''$ by (4.20a), the Eq. (4.21) yields $S''(t_0) \geq 0$. Consequently, taking into account the Eqs. (2.15), we obtain

$$\tilde{S}_1'' = \lambda_1 \tilde{S}_0'' + \mu_1 \tilde{S}_2'' \geq 0.$$

Analogously, conditions (2.15) and (4.20b) yield

$$\tilde{S}_k'' \geq 0, \quad \tilde{S}_{k-1}'' \geq 0,$$

and using the representation (2.2) for $\tilde{S}(t)$, one obtains

$$\tilde{S}''(\xi) = (1 - \xi) \tilde{S}_{i-1}'' + \xi \tilde{S}_i'' \geq 0, \quad \xi \in [0, 1]. \quad \square$$

If v_i are the data in the interpolation problem, the convexity is defined as

$$\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \geq 0, \quad i = 1, 2, \dots, k - 1, \tag{4.22}$$

cf. [6].

Note that for a smooth function, we have

$$\frac{2}{h_i + h_{i+1}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right) = v_i'' + \mathcal{O}(\bar{h}^2). \tag{4.23}$$

From (3.5), (4.7), and (4.23), we see that

$$\frac{2}{\hat{h}_i} \left(\frac{I_{i+2} - I_{i+1}}{h_{i+1} + h_{i+2}} - \frac{I_i - I_{i-1}}{h_{i-1} + h_i} \right) \approx \frac{2}{h_i + h_{i+1}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right)$$

with the accuracy $\mathcal{O}(\bar{h}^2)$. Therefore, (4.19) implies (4.22) with the accuracy $\mathcal{O}(\bar{h}^2)$. Theorems 4.2 and 4.3 show that the local integro cubic spline have good approximation and convexity properties. Note that for the interpolation cubic spline $S(t) \in C^2$ the sufficient conditions for convexity are

$$\begin{aligned} 2\Delta_i^2 - \mu_i \Delta_{i-1}^2 - \lambda_i \Delta_{i+1}^2 &\geq 0, \quad i = 1, 2, \dots, k - 1, \\ 2\Delta_0^2 - \Delta_1^2 &\geq 0, \quad 2\Delta_k^2 - \Delta_{k-1}^2 \geq 0, \end{aligned}$$

where $\Delta_i^2 = f[t_{i-1}, t_i, t_{i+1}]$, cf. [8]. On the other hand, the local integro cubic spline is convex under conditions (4.19) and (4.20). This is one of the advantages of our local integro spline as compared to the interpolation C^2 cubic spline.

5. Numerical Examples

We tested the formulae (4.7)-(4.9) for a uniform grid and some test functions, and the results are consistent with the theoretical ones. Suppose that the wheel of the robot starts at $t_0 = 0$ and runs by $v(t) = \sin(t)$ with $\square = 0.2$, then stops at $t_k = \pi$. We construct the velocity by (2.2), using the coefficients obtained from a system of $(2k+2)$ linear equations and the local integro spline (3.6). Fig. 3 shows the proximity of the curves appear but the local one (green online) rises up a bit at the ends. Next, we want to know the behavior of the real-time situation. The wheel with an encoder starts at $t_0 = 0$ and runs by $v(t) = 0.9 \sin(\pi t) + 0.76t$ with $\square = 0.1$. Local construction (dashed line in Fig. 4) fits the original velocity. As the wheel goes slowly, errors may arise near that time interval. From this example, the local integro spline successfully discovers the velocity in real-time. At the beginning of the movement, the velocity can be computed with seven values $[t_0, \dots, t_6]$, and in the middle of the movement — with six values $[h_{i-2}, h_{i-1}, \dots, h_{i+3}]$. In order to test the convergence order of the method for $v(t) \in C^5[t_0, t_k]$, we consider the uniform grid on the interval $t \in [0, 1]$ with $k = 20$ and create an almost uniform grid \mathcal{T}_{20} by adding or subtracting random numbers $\text{rand}_i \in [10^{-6}, 10^{-4}]$ to each node t_i of the uniform grid.

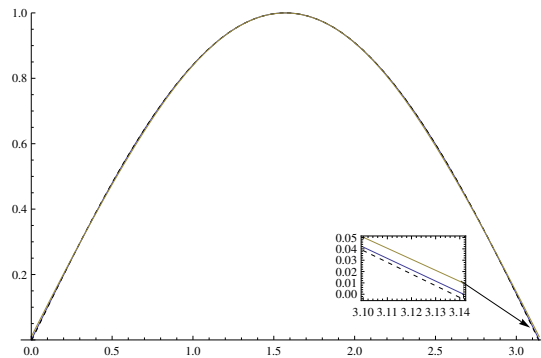


Figure 3: The trajectory of $v(t) = \sin(t)$ with Area=0.2 on $[0, \pi]$.

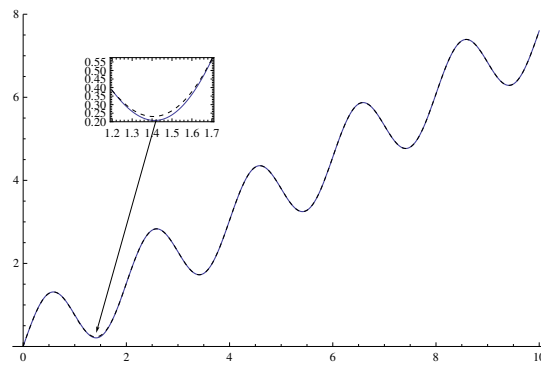


Figure 4: An unfinished movement.

Each of these almost uniform subintervals is then divided in half to create an almost uniform grid \mathcal{T}_k , $k = 40, 80, \dots$. Table 1 shows the corresponding numerical results compared to the method in [9]. The numerical convergence order (NCO) is computed by a formula from [9,13]. The boundary values of the integro quadratic spline $Q(t)$ [9] are given by exact values v_0 and v_k . Our method approximates the first and second derivatives of the function better than the method in [9]. Finally, we consider the convexity of the proposed integro splines. In Fig. 5, the velocity function $v(t) = 2 - \sqrt{t(2-t)}$ is approximated on $\mathcal{T}_{10} = \{0, 0.05, 0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 1.95, 2\}$. The dotted line (a) represents the spline (2.2) with the coefficients computed by solving the system equation, whereas the solid and the

Table 1: The numerical results for the function $v(t) = 0.9\sin(\pi t) + 0.76t$, $t \in [0, 1]$.

k	$\ \hat{S}_i - v_i\ _{\infty,k}$	NCO	$\ Q_i - v_i\ _{\infty,k}$	NCO	$\ \hat{S}'_i - v'_i\ _{\infty,k}$	NCO	$\ Q'_i - v'_i\ _{\infty,k}$	NCO	$\ \hat{S}''_i - v''_i\ _{\infty,k}$	NCO
20	7.18×10^{-5}	-	3.05×10^{-06}	-	4.16×10^{-3}	-	5.82×10^{-3}	-	1.69×10^{-1}	-
40	2.28×10^{-6}	4.97	1.90×10^{-07}	4.00	2.64×10^{-4}	3.97	1.45×10^{-3}	2.00	2.14×10^{-2}	2.97
80	1.16×10^{-7}	4.29	1.19×10^{-08}	4.00	1.66×10^{-5}	3.99	3.63×10^{-4}	2.00	3.42×10^{-3}	2.64
160	7.24×10^{-9}	3.99	7.43×10^{-10}	4.00	1.04×10^{-6}	4.00	9.08×10^{-5}	2.00	8.56×10^{-4}	1.99

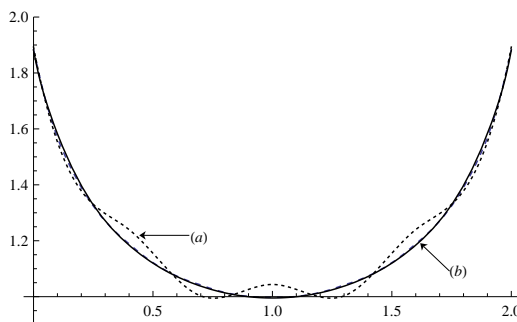


Figure 5: Convexity properties of proposed integro splines.

dashed lines (b) are the splines (2.2) and (3.6), respectively, with the coefficients given by $\tilde{S}_i, \tilde{S}'_i, \tilde{S}''_i$ and $\tilde{\alpha}_i$. We note that the solid and dashed (blue online) lines are close to each other, (b) is convex, but (a) is not.

6. Conclusion

We constructed integro cubic splines on non-uniform partitions such that the corresponding local splines have good convexity and approximation properties. Numerical experiments are consistent with the theoretical findings.

Acknowledgments

The authors thank the reviewers who helped to clarify the manuscript. This work was partially supported by the Foundation of Science and Technology of Mongolia (Grant No. SST_18/2018).

References

- [1] D. Barrera, S. Eddargani and A. Lamnii, *Uniform algebraic hyperbolic spline quasi-interpolant based on mean integral values*, *Comput. Math. Methods*, doi: 10.1002 /cmm4.1123.
- [2] H. Behforooz, *Approximation by integro cubic splines*, *Appl. Math. Comput.* **175**, 8–15 (2006).
- [3] S. Eddargani, A. Lamnii and M. Lamnii, *On algebraic trigonometric integro splines*, *Z. Angew. Math. Mech.* e201900262, doi: 10.1002/zamm.201900262 (2019).
- [4] X. Guo, X. Han and Y. Zhang, *The local integro splines with optimized knots*, *Comp. Appl. Math.* **38**:156, doi: 10.1007/s40314-019-0960-z (2019).
- [5] X. Han, *Convexity-preserving approximation by univariate cubic splines*, *J. Comput. Appl. Math.* **287**, 196–206 (2015).
- [6] T. Kim and B.I. Kvasov, *A shape-preserving approximation by weighted cubic splines*, *J. Comput. Appl. Math.* **236**, 4383–4397 (2012).
- [7] E. Kirsiaed, P. Oja and G.W. Shah, *Cubic spline histopolation*, *Mathematical modelling and analysis* **22**, 514–527 (2017).

- [8] Yu.S. Volkov, V.V. Bogdanov, V.L. Miroshnichenko and V.T. Shevaldin, *Shape-preserving interpolation by cubic splines*, Math. Notes **88**, 798–805 (2010).
- [9] J. Wu and X. Zhang, *Integro quadratic spline interpolation*, Appl. Math. Modell. **39**, 2973–2980 (2015).
- [10] J. Wu and X. Zhang, *Integro sextic spline interpolation and its super convergence*, Appl. Math. Comput. **219**, 6431–6436 (2013).
- [11] Yu.S. Zavyalov, B.I. Kvasov and V.L. Miroshnichenko, *Methods of Spline Functions (in Russian)*, Nauka, Moscow (1980).
- [12] T. Zhanlav, *B-representation of interpolatory cubic splines (in Russian)*, Vychislitel'nye Systemy, Novosibirsk, **87**, 3–10 (1981).
- [13] T. Zhanlav and R. Mijiddorj, *The local integro cubic splines and their approximation properties*, Appl. Math. Comput. **216**, 2215–2219 (2010).
- [14] T. Zhanlav and R. Mijiddorj, *A comparative analysis of local cubic splines*, Comp. Appl. Math. **37**, 5576–5586 (2018).
- [15] T. Zhanlav and R. Mijiddorj, *Convexity and monotonicity properties of the local integro cubic spline*, Appl. Math. Comput. **293**, 131–137 (2017).
- [16] T. Zhanlav and R. Mijiddorj, *Integro quintic splines and their approximation properties*, Appl. Math. Comput. **231**, 536–543 (2014).
- [17] T. Zhanlav and R. Mijiddorj, *On local integro quartic splines*, Appl. Math. Comput. **269**, 301–307 (2015).



Symbolic-Numerical Algorithms for Solving Elliptic Boundary-Value Problems Using Multivariate Simplex Lagrange Elements

A. A. Gusev¹, V. P. Gerdt^{1,2}, O. Chuluunbaatar^{1,3}, G. Chuluunbaatar^{1,2},
S. I. Vinitzky^{1,2}, V. L. Derbov⁴, A. Gózdź⁵, and P. M. Krassovitskiy^{1,6}

¹ Joint Institute for Nuclear Research, Dubna, Russia

gooseff@jinr.ru

² RUDN University, 6 Miklukho-Maklaya St., Moscow 117198, Russia

³ Institute of Mathematics, National University of Mongolia, Ulaanbaatar, Mongolia

⁴ N.G. Chernyshevsky Saratov National Research State University, Saratov, Russia

⁵ Institute of Physics, University of M. Curie-Skłodowska, Lublin, Poland

⁶ Institute of Nuclear Physics, Almaty, Kazakhstan

Abstract. We propose new symbolic-numerical algorithms implemented in Maple-Fortran environment for solving the self-adjoint elliptic boundary-value problem in a d -dimensional polyhedral finite domain, using the high-accuracy finite element method with multivariate Lagrange elements in the simplexes. The high-order fully symmetric PI-type Gaussian quadratures with positive weights and no points outside the simplex are calculated by means of the new symbolic-numerical algorithms implemented in Maple. Quadrature rules up to order 8 on the simplexes with dimension $d = 3 - 6$ are presented. We demonstrate the efficiency of algorithms and programs by benchmark calculations of a low part of spectra of exactly solvable Helmholtz problems for a cube and a hypercube.

Keywords: Elliptic boundary-value problem · Finite element method
Multivariate simplex lagrange elements
High-order fully symmetric Gaussian quadratures
Helmholtz equation for cube and hypercube

1 Introduction

The progress of modern computing power offers more possibilities for setting and numerical solution of multidimensional elliptic boundary-value problems (BVPs) with high accuracy. 3D BVPs have wide applications in such areas as vibrating membrane, electromagnetic radiation, motion of thermal neutrons in the reactor, seismology, and acoustics, see, e.g., [4], while multidimensional BVPs have applications in nuclear physics, see, e.g., [7]. For this purpose, novel numerical

methods of high accuracy order are being developed. When reducing the boundary value problem to an algebraic one in the finite element method (FEM) of the order p , one of the problems is the calculation of integrals on a finite element (we consider only simplicial finite elements) containing the products of two basis functions of Lagrange or Hermite interpolation polynomials of the order p by the coefficients for the unknown functions [5,9]. There are three possible ways to calculate the integrals:

- (i) using analytical calculation, which is possible for a limited number of cases;
- (ii) using quadrature formulas with products of two basic functions used as a weight function;
- (iii) using quadrature formulas with a single weight function.

It is well known [20] that as a result of applying the p th order FEM to the solution of the discrete spectrum problem for the elliptic (Schrödinger) equation, the eigenfunction and the eigenvalue are determined with an accuracy of the order $p + 1$ and $2p$ provided that all intermediate quantities are calculated with sufficient accuracy. It follows that for the realization of the FEM of the order p in the third case, the integrals must be computed at least with an accuracy of the order $2p$, depending on the problem considered. The most economical calculation of such integrals is achieved using the quadratures of Gaussian type. In the one-dimensional case, the nodes and the quadrature Gaussian weights are expressed analytically; in the two-, three- and four-dimensional case, the high-order quadrature formulas are determined numerically [2,6,8,10,17–19,21]. Note that for multidimensional integrals, numerous quadrature formulas of the Newton–Cotes and third-order Gaussian type are known, too (see Ref. [1]).

The paper presents a new method for constructing fully symmetric multidimensional Gaussian-type quadratures on a standard simplex. The main idea of the method is replacing the coordinates of nodes with their symmetric combinations obtained using the Vieta theorem, which simplifies the system of nonlinear algebraic equations. The construction of the desired systems of equations is performed analytically using an original algorithm implemented in Maple [13]. The derived systems up to the sixth order are solved using the built-in procedure `PolynomialSystem`, implementing the technique of Gröbner bases, and the systems of higher order are solved using the developed symbolic-numerical algorithm based on numerical methods, implemented in Maple-Fortran environment. We demonstrate the efficiency of algorithms and programs by benchmark calculations of the lower part of spectra in exactly solvable Helmholtz problems for a cube and a hypercube.

The paper is structured as follows. In Sects. 2 and 3, the FEM schemes and algorithms for solving the d -dimensional BVP are presented. In Sect. 4, the algorithms for constructing the d -dimensional fully symmetric Gaussian quadratures are presented. In Sect. 5, the benchmark calculations of the exactly solvable Helmholtz problems for the cube and hypercube are presented. In Conclusion, we discuss the results and perspectives.

2 Setting of the Problem

Consider a self-adjoint boundary-value problem for the elliptic differential equation of the second order:

$$(D - E)\Phi(z) \equiv \left(-\frac{1}{g_0(z)} \sum_{ij=1}^d \frac{\partial}{\partial z_i} g_{ij}(z) \frac{\partial}{\partial z_j} + V(z) - E \right) \Phi(z) = 0. \quad (1)$$

For the principal part coefficients of Eq. (1), the condition of uniform ellipticity holds in the bounded domain $z = (z_1, \dots, z_d) \in \Omega$ of the Euclidean space \mathcal{R}^d , i.e., the constants $\mu > 0$, $\nu > 0$ exist such that $\mu\xi^2 \leq \sum_{ij=1}^d g_{ij}(z)\xi_i\xi_j \leq \nu\xi^2$, $\xi^2 = \sum_{i=1}^d \xi_i^2$, $\forall \xi_i \in \mathcal{R}$. The left-hand side of this inequality expresses the requirement of ellipticity, while the right-hand side expresses the boundedness of the coefficients $g_{ij}(z)$. It is also assumed that $g_0(z) > 0$, $g_{ji}(z) = g_{ij}(z)$ and $V(z)$ are real-valued functions, continuous together with their generalized derivatives to a given order in the domain $z \in \bar{\Omega} = \Omega \cup \partial\Omega$ with the piecewise continuous boundary $S = \partial\Omega$, which provides the existence of nontrivial solutions obeying the boundary conditions [5] of the first kind

$$\Phi(z)|_S = 0, \quad (2)$$

or the second kind

$$\frac{\partial\Phi(z)}{\partial n_D} \Big|_S = 0, \quad \frac{\partial\Phi(z)}{\partial n_D} = \sum_{ij=1}^d (\hat{n}, \hat{e}_i) g_{ij}(z) \frac{\partial\Phi(z)}{\partial z_j}, \quad (3)$$

where $\frac{\partial\Phi_m(z)}{\partial n_D}$ is the derivative along the conformal direction, \hat{n} is the outer normal to the boundary of the domain $S = \partial\Omega$, \hat{e}_i is the unit vector of $z = \sum_{i=1}^d \hat{e}_i z_i$, and (\hat{n}, \hat{e}_i) is the scalar product in \mathcal{R}^d .

For a discrete spectrum problem, the functions $\Phi_m(z)$ from the Sobolev space $H_2^{s \geq 1}(\Omega)$, $\Phi_m(z) \in H_2^{s \geq 1}(\Omega)$, corresponding to the real eigenvalues E : $E_1 \leq E_2 \leq \dots \leq E_m \leq \dots$ satisfy the conditions of normalization and orthogonality

$$\langle \Phi_m(z) | \Phi_{m'}(z) \rangle = \int_{\Omega} dz g_0(z) \Phi_m(z) \Phi_{m'}(z) = \delta_{mm'}, \quad dz = dz_1 \dots dz_d. \quad (4)$$

The FEM solution of the boundary-value problems (1)–(4) is reduced to the determination of stationary points of the variational functional [3, 5]

$$\Xi(\Phi_m, E_m) \equiv \int_{\Omega} dz g_0(z) \Phi_m(z) (D - E_m) \Phi(z) = \Pi(\Phi_m, E_m), \quad (5)$$

where $\Pi(\Phi, E)$ is the symmetric quadratic functional

$$\Pi(\Phi, E) = \int_{\Omega} dz \left[\sum_{ij=1}^d g_{ij}(z) \frac{\partial\Phi(z)}{\partial z_i} \frac{\partial\Phi(z)}{\partial z_j} + g_0(z) \Phi(z) (V(z) - E) \Phi(z) \right].$$

3 FEM Calculation Scheme

In FEM, the domain $\Omega = \Omega_h(z) = \bigcup_{q=1}^Q \Delta_q$, specified as a polyhedral domain, is covered with finite elements, in the present case, the simplexes Δ_q with $d + 1$ vertices $\hat{z}_i = (\hat{z}_{i1}, \hat{z}_{i2}, \dots, \hat{z}_{id})$, $i = 0, \dots, d$. Each edge of the simplex Δ_q is divided into p equal parts, and the families of parallel hyperplanes $H(i, k)$ are drawn, numbered with the integers $k = 0, \dots, p$, starting from the corresponding face (see also [5]). The equation of the hyperplane is $H(i, k): H(i; z) - k/p = 0$, where $H(i; z)$ is a linear function of z .

The node points of hyperplanes crossing A_r are enumerated with the sets of integers $[n_0, \dots, n_d]$, $n_i \geq 0$, $n_0 + \dots + n_d = p$, where n_i , $i = 0, 1, \dots, d$ are the numbers of hyperplanes, parallel to the simplex face, not containing the i th vertex $\hat{z}_i = (\hat{z}_{i1}, \dots, \hat{z}_{id})$. The coordinates $\xi_r = (\xi_{r1}, \dots, \xi_{rd})$ of the node point $A_r \in \Delta_q$ are calculated using the formula

$$(\xi_{r1}, \dots, \xi_{rd}) = (\hat{z}_{01}, \dots, \hat{z}_{0d}) \frac{n_0}{p} + (\hat{z}_{11}, \dots, \hat{z}_{1d}) \frac{n_1}{p} + \dots + (\hat{z}_{d1}, \dots, \hat{z}_{dd}) \frac{n_d}{p} \quad (6)$$

from the coordinates of the vertices $\hat{z}_j = (\hat{z}_{j1}, \dots, \hat{z}_{jd})$. Then the Lagrange interpolation polynomials (LIP) $\varphi_r^p(z)$ are equal to one at the point A_r with the coordinates $\xi_r = (\xi_{r1}, \dots, \xi_{rd})$, characterized by the numbers $[n_0, n_1, \dots, n_d]$, and equal to zero at the remaining points $\xi_{r'}$, i.e., $\varphi_r^p(\xi_{r'}) = \delta_{rr'}$, have the form

$$\varphi_r^p(z) = \prod_{i=0}^d \prod_{n'_i=0}^{n_i-1} \frac{H(i; z) - n'_i/p}{H(i; \xi_r) - n'_i/p}. \quad (7)$$

As shape functions in the simplex Δ_q we use the multivariate Lagrange interpolation polynomials $\varphi_l^p(z)$ of the order p that satisfy the condition $\varphi_l^p(x_{1l'}, x_{2l'}) = \delta_{ll'}$, i.e., equal 1 at one of the points A_l and zero at the other points. In this method, the piecewise polynomial functions $N_l^p(z)$ in the domain Ω are constructed by joining the shape functions $\varphi_l^p(z)$ in the simplex Δ_q :

$$N_l^p(z) = \{\varphi_l^p(z), A_l \in \Delta_q; 0, A_l \notin \Delta_q\}$$

and possess the following properties: the functions $N_l^p(z)$ are continuous in the domain Ω ; the functions $N_l^p(z)$ equal 1 at one of the points A_l and zero at the rest of the points; $N_l^p(z_{l'}) = \delta_{ll'}$ in the entire domain Ω . Here l takes the values $l = 1, \dots, N$.

The functions $N_l^p(z)$ form a basis in the space of polynomials of the p th order. Now, the function $\Phi(z) \in \mathcal{H}^1(\Omega)$ is approximated by a finite sum of piecewise basis functions $N_l^p(z)$:

$$\Phi^h(z) = \sum_{l=1}^N \Phi_l^h N_l^p(z). \quad (8)$$

Table 1. The orbits and their number of permutations for $d = 3, 4, 5, 6$.

$d = 3$		$d = 4$		$d = 5$				$d = 6$			
Orbits	Perm.	Orbits	Perm.	Orbits	Perm.	Orbits	Perm.	Orbits	Perm.	Orbits	Perm.
S_4	1	S_5	1	S_6	1	S_{3111}	120	S_7	1	S_{4111}	210
S_{31}	4	S_{41}	5	S_{51}	6	S_{2211}	180	S_{61}	7	S_{3211}	420
S_{22}	6	S_{32}	10	S_{42}	15	S_{21111}	360	S_{52}	21	S_{2221}	630
S_{211}	12	S_{311}	20	S_{33}	20	S_{111111}	720	S_{43}	35	S_{31111}	840
S_{1111}	24	S_{221}	30	S_{411}	30			S_{511}	42	S_{22111}	1260
		S_{2111}	60	S_{321}	60			S_{421}	105	S_{211111}	2520
		S_{11111}	120	S_{222}	90			S_{331}	140	$S_{1111111}$	5040
								S_{322}	210		

After substituting expansion (8) into the variational functional (5) and minimizing it [3, 20], we obtain the generalized eigenvalue problem

$$\mathbf{A}^p \boldsymbol{\Phi}^h = \varepsilon^h \mathbf{B}^p \boldsymbol{\Phi}^h. \tag{9}$$

Here \mathbf{A}^p is the symmetric stiffness matrix; \mathbf{B}^p is the symmetric positive definite mass matrix; $\boldsymbol{\Phi}^h$ is the vector approximating the solution on the finite-element grid; and ε^h is the corresponding eigenvalue. The matrices \mathbf{A}^p and \mathbf{B}^p have the form:

$$\mathbf{A}^p = \{a_{ll'}^p\}_{ll'=1}^N, \mathbf{B}^p = \{b_{ll'}^p\}_{ll'=1}^N, \tag{10}$$

where the matrix elements $a_{ll'}^p$ and $b_{ll'}^p$ are calculated for simplex elements as

$$\begin{aligned} a_{ll'}^p &= \sum_{ij=1}^d \int_{\Delta_q} g_{ij}(z) \frac{\partial \varphi_l^p(z)}{\partial z_i} \frac{\partial \varphi_{l'}^p(z)}{\partial z_j} dz + \int_{\Delta_q} g_0(z) \varphi_l^p(z) \varphi_{l'}^p(z) V(z) dz, \\ b_{ll'}^p &= \int_{\Delta_q} g_0(z) \varphi_l^p(z) \varphi_{l'}^p(z) dz. \end{aligned} \tag{11}$$

The economical implementation of FEM is the following.

The calculations, including those of FEM integrals for mass and stiffness matrices at each subdomain Δ_q are performed in the local (reference) system of coordinates x , in which the coordinates of the simplex vertices are the following: $\hat{x}_j = (\hat{x}_{j1}, \dots, \hat{x}_{jd})$, $\hat{x}_{jk} = \delta_{jk}$, $j = 0, \dots, d$, $k = 1, \dots, d$.

Let us construct the Lagrange interpolation polynomial (LIP) on an arbitrary d -dimensional simplex Δ_q with vertices $\hat{z}_i = (\hat{z}_{i1}, \hat{z}_{i2}, \dots, \hat{z}_{id})$, $i = 0, \dots, d$. For this purpose, we introduce the local system of coordinates $x = (x_1, x_2, \dots, x_d) \in \mathcal{R}^d$, in which the coordinates of the simplex vertices are \hat{x}_i . The relation between the coordinates is given by the formula:

$$z_i = \hat{z}_{0i} + \sum_{j=1}^d \hat{J}_{ij} x_j, \quad \hat{J}_{ij} = \hat{z}_{ji} - \hat{z}_{0i}, \quad i = 1, \dots, d. \tag{12}$$

Table 2. Quadrature rule on tetrahedra.

Orbit	Weight	Abcissas
14-points 4-order rule		
S_{31}	0.0801186758957551214557967806191	0.0963721076152827180679867982109
S_{31}	0.1243674424942431317471251193937	0.3123064218132941261147265437508
S_{22}	0.0303425877400011645313853999915	0.0274707886853344957750132954191
14-points 5-order rule		
S_{31}	0.0734930431163619495437102054863	0.0927352503108912264023239137370
S_{31}	0.1126879257180158507991856523333	0.3108859192633006097973457337635
S_{22}	0.0425460207770814664380694281203	0.0455037041256496494918805262793
24-points 6-order rule		
S_{31}	0.0399227502581674920996906275575	0.2146028712591520292888392193863
S_{31}	0.0100772110553206429480132374459	0.0406739585346113531155794489564
S_{31}	0.0553571815436547220951532778537	0.3223378901422755103439944707625
S_{211}	0.0482142857142857142857142857143	0.0636610018750175252992355276057 0.6030056647916491413674311390609
35-points 7-order rule		
S_4	0.0954852894641308488605784361172	0.25000000000000000000000000000000
S_{31}	0.0423295812099670290762861707986	0.3157011497782027994234299995933
S_{22}	0.0318969278328575799342748240829	0.0504898225983963687630538229866
S_{211}	0.0372071307283346213696155611915	0.1888338310260010477364311038546 0.5751716375870000234832415770223
S_{211}	0.0081107708299033415661034334911	0.0212654725414832459888361014998 0.8108302410985485611181053798482
46-points 8-order rule		
S_{31}	0.0063972777406656176515049738764	0.0396757518582111225277078936298
S_{31}	0.0401906214382288067038698161802	0.3144877686588789672386516888007
S_{31}	0.0243081692121760770795396363192	0.1019873469010702748038937565346
S_{31}	0.0548586277637264928464254253584	0.1842037697228154771186065671874
S_{22}	0.0357196747563309013579348149829	0.0634363951662790318385035375295
S_{211}	0.0071831862652404057248973769332	0.0216901288123494021982001218658 0.7199316530057482532021892796203
S_{211}	0.0163720776383284788356885983306	0.2044800362678728018101543629799 0.5805775568740886759781950895673

The inverse transformation and the relation between the differentiation operators are given by the formulas

$$x_i = \sum_{j=1}^d (\hat{J}^{-1})_{ij} (z_j - \hat{z}_{0j}), \frac{\partial}{\partial x_i} = \sum_{j=1}^d \hat{J}_{ji} \frac{\partial}{\partial z_j}, \frac{\partial}{\partial z_i} = \sum_{j=1}^d (\hat{J}^{-1})_{ji} \frac{\partial}{\partial x_j}.$$

Table 3. Quadrature rule on $d = 4$ dimensional simplex.

Orbit	Weight	Abscissas
20-points 4-order rule		
S_{41}	0.0379539224206539610831511760634	0.0784224645320084412701860095372
S_{41}	0.0681384495140965073072374189421	0.2449925002516506829747267241998
S_{32}	0.0469538140326247658048057024973	0.0657807054017604429326659923627
30-points 5-order rule		
S_{41}	0.0492516801753157409383956672833	0.0853466308308594082516329452526
S_{41}	0.0325114606587393649369493738878	0.2369600116614607056460832163398
S_{32}	0.0175327109958004508766635908927	0.0412980141318484010482052159450
S_{32}	0.0415857185871719961856638885218	0.2997443384790352862963354895649
56-points 6-order rule		
S_5	0.0732792367435547721884408088550	0.20000000000000000000000000000000
S_{41}	0.0047429121713183739117905941798	0.0417033817484816144703679735243
S_{32}	0.0371671124025330069869448829255	0.2956227971470980491911963343462
S_{311}	0.0133362480184817717166547744056	0.1543949248731168427369921195673 0.5227506462276968325151584695712
S_{311}	0.0132305059002443927025030951440	0.0478156751378274921515148624255 0.2819739419928806028716278777811
76-points 7-order rule		
S_5	0.0282727667597935101461654674137	0.20000000000000000000000000000000
S_{41}	0.0171637920155537955591265968365	0.2494020893093779695674000557470
S_{32}	0.0084262904177368737487641566458	0.0390279956601069690478223468028
S_{32}	0.0151633627560453145809862914879	0.1283114044638121921594658569279
S_{311}	0.0041099348414815560204478025486	0.0338474709865642635279969618386 0.7462624286813390611020624803775
S_{221}	0.0189271014864994836117247005365	0.0448337964557961849763900084527 0.2098710857162324764262981778162
110-points 8-order rule		
S_{41}	0.0209889631062033488284471858741	0.1064160632601420588468274348524
S_{41}	0.0025569304299619087111133529054	0.0405432824126613113549340882657
S_{32}	0.0153364140237452308225281532013	0.0553205204859791157778648564000
S_{32}	0.0143413703554045577679712361587	0.1329849247207488765271172398305
S_{32}	0.0219839063571691797013874119590	0.2921649623679039933512390863408
S_{311}	0.0036998351176104420717284969383	0.0333398788668747287190327986033 0.6960284779140254845117282473257
S_{311}	0.0102875153954967332446050836803	0.1749055465990825034189472406388 0.4713583394803434080155451322627
S_{221}	0.0028635538231280174352219226847	0.2139955562978852147651302856947 0.0055794471455235244097015787040

Table 4. Quadrature rule on $d = 5$ dimensional simplex.

Orbit	Weight	Abcissas
27-points 4-order rule		
S_6	0.2380952380952380952380952380952	0.16666666666666666666666666666667
S_{51}	0.0476190476190476190476190476190	0.08333333333333333333333333333333
S_{33}	0.0238095238095238095238095238095	0.3110042339640731077939538617922
37-points 5-order rule		
S_6	0.1537202203084293617727126367247	0.16666666666666666666666666666667
S_{51}	0.0289106224493151615615928162885	0.07500000000000000000000000000000
S_{42}	0.0272301053298578547025239158396	0.0620931177937680448262436473512
S_{42}	0.0176242976698541232213247818634	0.2494113069849930171206590075161
102-points 6-order rule		
S_{51}	0.0220609777699918416385171809216	0.0936784796657907179507883184494
S_{51}	0.0010288939840293747752001192602	0.0270566434340766625713558698570
S_{42}	0.0156264172618719457418380080610	0.0653950986037339179722692404805
S_{42}	0.0278282494445825546266341924031	0.2298844181626658901051213339390
S_{321}	0.0034940128146509199331768865324	0.0182868036924305667708203585711 0.1963426392615138866458359282858
137-points 7-order rule		
S_{51}	0.0251079912995851246690568379932	0.1962505998027202386302784835916
S_{51}	0.0268181773072546325688248594140	0.1073064529494792948889112833415
S_{42}	0.0088856106397381008037487732556	0.0499693465734168548516130660759
S_{33}	0.0155965105537609568596496409074	0.2812294050576655725449341659515
S_{411}	0.0013215130252633881273492640567	0.0287356582492413683812555969369 0.7243025794534749187969716773294
S_{321}	0.0033930537821628193917167912812	0.1573270862326151676898601262299 0.0036548286115748769147071291765
257-points 8-order rule		
S_{51}	0.0176303711895221798359615170829	0.1062079269440531427851821818230
S_{51}	0.0022261212103870366035563829745	0.0445128753938546747539305403018
S_{42}	0.0166747305797216127029493671085	0.2215271654487921945556436076078
S_{33}	0.0039660204626209654516270279365	0.0287362439702382298273521354305
S_{411}	0.0013712761289024193505102030670	0.0302807316628161184245512327246 0.5742625240747101119061964222732
S_{321}	0.0009261971752463936292941257741	0.0178653742410041824343316617132 0.1599485035546596050768099856676
S_{321}	0.0048311921097760693226621205033	0.0971175464224689537586197747871 0.3509135920039025566598219642999
S_{321}	0.0027473006113980140692238444274	0.1542598417836536904457879818959 0.0175301902661063495789625995714

Table 5. Quadrature rule on $d = 6$ dimensional simplex.

Orbit	Weight	Abscissas
43-points 4-order rule		
S_7	0.1668996242406426424065553487802	0.1428571428571428571428571428571
S_{61}	0.0271661981514270076903673620086	0.0712015434701090173255254362504
S_{43}	0.0183696282485533801074535176331	0.0378762710421960021962053657298
64-points 5-order rule		
S_7	0.1055608940320069322326417879346	0.1428571428571428571428571428571
S_{61}	0.0242990419532018650013794612051	0.0715539250843990305857473101707
S_{52}	0.0117134616879203157617441588591	0.0506772832103077178123184150643
S_{43}	0.0136675176242643823360307042168	0.2304358521244036512024566237956
175-points 6-order rule		
S_{61}	0.0004610493156525528548408228337	0.0250990960487081544700908516534
S_{61}	0.0130199167458605046501306895616	0.1640882485030238802990581503886
S_{52}	0.0020306497109021799567911952305	0.0278440785001665193354091212251
S_{43}	0.0162220926263431272900952737070	0.0542711738847223476721544566326
S_{421}	0.0028115843020805082211357117490	0.1203196589728741910526848418155 0.0037549817118180216976885119286
266-points 7-order rule		
S_{61}	0.0103583726453788825261551030659	0.1655537069170340713573624387430
S_{52}	0.0127946542771734405339991326892	0.0800416917413849453828158790868
S_{52}	0.0038665797691560684680540249746	0.0462060207372654835707639356206
S_{43}	0.0068482273738159415062980403942	0.2251626772370571673652419443913
S_{43}	0.0013006546667652760792540506406	0.0140208383611713481747343760562
S_{511}	0.0005321899098570485728489000218	0.0246678063639990490447074776734 0.1759636130065151239491183217936
S_{421}	0.0025718345607151378830459140997	0.1242831811867119456481842408470 0.0063723131014287473559192490677
553-points 8-order rule		
S_{61}	0.0119576998439189095322140668380	0.1646768753323421340942870425551
S_{61}	0.0170033855208889021739988777538	0.1010702610627718250051913258275
S_{61}	0.0015763271020889357220309420300	0.0445013301458845571180677283528
S_{43}	0.0029960134851163901478666677698	0.0444259533505434743654069329655
S_{43}	0.0057810264432097073309950803359	0.2211051271607452660739567583653
S_{511}	0.0007096981072933306194796057518	0.0303842211182356803799849235650 0.2575978419615841769164822870809
S_{421}	0.0003172772160146728270743668040	0.0126686383758556644736172343255 0.2101770124793451029895811597503
S_{421}	0.0015276586289853906949163952851	0.1232675348992300327954722629436 0.0050316009864769548591929730662
S_{322}	0.0012167434809951561924521816620	0.0955868297374816410778226310866 0.3377885686906383657970155568362

The integrals that enter the variational functional (5) on the domain $\Omega_h(z) = \bigcup_{q=1}^Q \Delta_q$, are expressed via the integrals, calculated on the element Δ_q , and recalculated to the local coordinates x on the element Δ ,

$$\begin{aligned} \int_{\Delta_q} dz g_0(z) \varphi_r^p(z) \varphi_{r'}^p(z) V(z) &= J \int_{\Delta} dx g_0(z(x)) \varphi_r^p(x) \varphi_{r'}^p(x) V(z(x)), \quad (13) \\ \int_{\Delta_q} dz g_{s_1 s_2}(z) \frac{\partial \varphi_r^p(z)}{\partial z_{s_1}} \frac{\partial \varphi_{r'}^p(z)}{\partial z_{s_2}} \\ &= J \sum_{t_1, t_2=1}^d \hat{J}_{s_1 s_2; t_1 t_2}^{-1} \int_{\Delta} dx g_{s_1 s_2}(z(x)) \frac{\partial \varphi_r^p(x)}{\partial x_{t_1}} \frac{\partial \varphi_{r'}^p(x)}{\partial x_{t_2}}, \end{aligned}$$

where $J = \det \hat{J} > 0$ is the determinant of the matrix \hat{J} from Eq. (12), $\hat{J}_{s_1 s_2; t_1 t_2}^{-1} = (\hat{J}^{-1})_{t_1 s_1} (\hat{J}^{-1})_{t_2 s_2}$, $dx = dx_1 \dots dx_d$.

In the local coordinates, the LIP $\varphi_r^p(x)$ is equal to one at the node point ξ_r characterized by the numbers $[n_0, n_1, \dots, n_d]$, and zero at the remaining node points ξ_r' , i.e., $\varphi_r(\xi_r) = \delta_{rr}$ are determined by Eq. (7) at $H(0; x) = 1 - x_1 - \dots - x_d$, $H(i; z) = x_i$, $i = 1, \dots, d$:

$$\varphi_r(x) = \prod_{i=1}^d \prod_{n_i=0}^{n_i-1} \frac{x_i - n_i/p}{n_i/p - n_i/p} \prod_{n_0=0}^{n_0-1} \frac{1 - x_1 - \dots - x_d - n_0/p}{n_0/p - n_0/p}. \quad (14)$$

Integrals (13) are evaluated using the Gaussian quadrature of the order $2p$.

Let ε_m and $\Phi_m(z)$ be exact solutions of Eq. (9) and ε_m^h and Φ_m^h be the corresponding numerical solutions. Then the following estimations are valid [20]

$$|\varepsilon_m - \varepsilon_m^h| \leq c_1 |\varepsilon_m| h^{2p}, \quad \|\Phi_m(z) - \Phi_m^h\|_0 \leq c_2 |E_m| h^{p+1}, \quad (15)$$

where $\|a(z)\|_0^2 = \langle a(z) | a(z) \rangle$, h is the maximal step of the finite-element grid, m is the number of the corresponding solution, and the positive constants c_1 and c_2 do not depend on the step h .

To solve the generalized eigenvalue problem (9), we choose the subspace iteration method [3, 20] elaborated by Bathe [3] for the solution of large symmetric banded-matrix eigenvalue problems. This method uses the skyline storage mode which stores the components of the matrix column vectors within the banded region of the matrix, and is ideally suited for banded finite-element matrices.

4 Construction of the d -dimensional Quadrature Formulas

Let us construct the d -dimensional p -ordered quadrature formula

$$\int_{\Delta_q} dz V(z) = |\Delta_q| \sum_{j=1}^{n_t} w_j V(z_j), \quad z = (z_1, \dots, z_d), \quad dz = dz_1 \dots dz_d, \quad (16)$$

for integration over the d -dimensional simplex Δ_q with vertices $\hat{z}_i = (\hat{z}_{i1}, \hat{z}_{i2}, \dots, \hat{z}_{id})$, $i = 0, \dots, d$, which is exact for all polynomials of the variables z_1, \dots, z_d of degree not exceeding p , where n_t is the number of nodes that is determined during the calculation process. In Eq. (16), w_j , $j = 1, \dots, n_t$ are the weights and $z_j = (z_{j1}, z_{j2}, \dots, z_{jd})$ are the coordinates of nodes. $|\Delta_q|$ denotes the volume of Δ_q . For each node z_j , instead of sets of d coordinates we use the sets of $d + 1$ barycentric coordinates (BC) $(x_{j0}, x_{j1}, \dots, x_{jd})$:

$$z_j = x_{j0}\hat{z}_0 + \dots + x_{jd}\hat{z}_d, \quad x_{j0} + \dots + x_{jd} = 1. \tag{17}$$

For this purpose, we introduce the local coordinate system $x = (x_1, x_2, \dots, x_d)$ and (12). Therefore, without loss of generality, we construct the d -dimensional p -ordered quadrature formula (16) on the standard simplex Δ with vertices $\hat{x}_j = (\hat{x}_{j1}, \dots, \hat{x}_{jd})$, $\hat{x}_{jk} = \delta_{jk}$, $j = 0, \dots, d$, $k = 1, \dots, d$, which is exact for all polynomials of the variables x_1, \dots, x_d of degree not exceeding p :

$$\int_{\Delta} dx V(x) = \frac{1}{d!} \sum_{j=1}^{n_t} w_j V(x_{j0}, \dots, x_{jd}). \tag{18}$$

Since the following formula is valid for all permutations (l_0, \dots, l_d) of (k_0, \dots, k_d) :

$$\int_{\Delta} dx x_1^{l_1} \dots x_d^{l_d} (1 - x_1 - \dots - x_d)^{l_0} = \frac{\prod_{i=0}^d k_i!}{(d + \sum_{i=0}^d k_i)!},$$

we consider the fully symmetric Gaussian quadratures

$$\int_{\Delta} dx V(x) = \frac{1}{d!} \sum_{j=1}^a w_j \sum_{j_0, \dots, j_d} V(x_{j_0 0}, x_{j_1 1}, \dots, x_{j_d d}), \tag{19}$$

where the internal summation by j_0, \dots, j_d is carried out over the different permutations of $(x_{j_0}, x_{j_1}, \dots, x_{j_d})$. Table 1 presents the orbits and the corresponding number of different permutations for $d = 3, 4, 5, 6$. Here, for example, the orbit S_{331} at $d = 6$ contains BC $(\alpha, \alpha, \alpha, \beta, \beta, \beta, \gamma)$, $\alpha \neq \beta \neq \gamma$, $\alpha \neq \gamma$, $3\alpha + 3\beta + \gamma = 1$ and their different 140 permutations.

Substituting a monomial of the order not exceeding p in Eq. (19) instead of $V(x)$, we arrive at a system of nonlinear algebraic equations, that using the Vieta theorem reduces to the form:

$$\int_{\Delta} dx s_2^{l_2} s_3^{l_3} \times \dots \times s_{d+1}^{l_{d+1}} = \frac{1}{d!} \sum_{j=1}^a w_j Q_j s_{j_2}^{l_2} s_{j_3}^{l_3} \times \dots \times s_{j_{d+1}}^{l_{d+1}}, \tag{20}$$

$$2l_2 + 3l_3 + \dots + (d + 1)l_{d+1} \leq p, \tag{21}$$

where

$$s_2 = \sum_{i=0, j \neq i}^d x_i x_j, \quad \dots, \quad s_{d+1} = \prod_{i=0}^d x_i, \tag{22}$$

s_{ji} , $i = 2, \dots, d+1$, are their values in the BC $(x_{j0}, x_{j1}, \dots, x_{jd})$, and Q_j is the number of different permutation of the BC. As in Ref. [15], instead of Eq. (22), we can use

$$s_j = \sum_{i=0}^d x_i^j, \quad j = 2, \dots, d+1. \quad (23)$$

The number of all $l_j \geq 0$ solutions of Eq. (21) provides the minimal number of independent nonlinear equations for the quadrature formula of the order p . It means that we can obtain a set of independent polynomials by adding new polynomials when increasing the order p . Below the first few independent polynomials of the order not exceeding $p \leq 6$ for $d \geq 5$ are presented:

$$\begin{aligned} V_1(x) &= s_1, & \text{for } p = 1, \\ V_2(x) &= s_2, & \text{for } p = 2, \\ V_3(x) &= s_3, & \text{for } p = 3, \\ V_4(x) &= s_2^2, V_5(x) = s_4, & \text{for } p = 4, \\ V_6(x) &= s_2 s_3, V_7(x) = s_5, & \text{for } p = 5, \\ V_8(x) &= s_2^3, V_9(x) = s_3^2, V_{10}(x) = s_2 s_4, V_{11}(x) = s_6, & \text{for } p = 6. \end{aligned} \quad (24)$$

We consider fully symmetric rules with positive weights, and no points are outside the simplex (the so-called PI-type).

The n_p -points p -order quadrature rules are constructed with Algorithm 1 [21] implemented by us in Maple and Fortran:

- **for** each decomposition n_p **do**
 - repeat**
 1. Randomly choose an initial guess for the unknowns n_t .
 2. Find a least square solution to Eqs. (20), (21) using a quasi-Newton algorithm.
 3. If a PI-type solution is found satisfying Eqs. (20), (21), with sufficient accuracy, go to Step 4.
 - until** maximum number of initial guesses tried.
- **end for**
- **Stop.**
- **4.** Minimize the nonlinear equation for unknowns n_t using the Levenberg–Marquardt algorithm with high accuracy [12, 14].

The Levenberg–Marquardt Algorithm 2:

Let $f(\mathbf{x})$ be twice differentiable with respect to the variable $\mathbf{x} = (x_1, \dots, x_n)$. We consider the minimization

$$\min_{\mathbf{x} \in R^n} f(\mathbf{x}). \quad (25)$$

1. Start with an initial value \mathbf{x}_0 , in S , an initial damping parameter λ_0 , and a scaling parameter ρ . For $k \geq 0$ do the following:

2. Determine a trial iterate \mathbf{y} , using

$$\mathbf{y} = \mathbf{x}_k - (H_f(\mathbf{x}_k) + \lambda \operatorname{diag}(H_f(\mathbf{x}_k)))^{-1} \nabla f(\mathbf{x}_k), \quad (26)$$

with $\lambda = \lambda_k \rho^{-1}$.

3. If $f(\mathbf{y}) < f(\mathbf{x}_k)$, where \mathbf{y} is determined in Step 2, then set $\mathbf{x}_{k+1} = \mathbf{y}$ and $\lambda_{k+1} = \lambda_k \rho^{-1}$. Return to Step 2, replace k with $k + 1$, and compute a new trial iterate.
4. If $f(\mathbf{y}) \geq f(\mathbf{x}_k)$ in Step 3, determine a new trial iterate, \mathbf{y} , using (26) with $\lambda = \lambda_k$.
5. If $f(\mathbf{y}) < f(\mathbf{x}_k)$, where \mathbf{y} is determined in Step 4, then set $\mathbf{x}_{k+1} = \mathbf{y}$ and $\lambda_{k+1} = \lambda_k$. Return to Step 2, replace k with $k + 1$, and compute a new trial iterate.
6. If $f(\mathbf{y}) \geq f(\mathbf{x}_k)$ in Step 5, then determine the smallest value of m so that when a trial iterate \mathbf{y} is computed using (26) with $\lambda = \lambda_k \rho^m$, then $f(\mathbf{y}) < f(\mathbf{x}_k)$. Set $\mathbf{x}_{k+1} = \mathbf{y}$ and $\lambda_{k+1} = \lambda_k \rho^m$. Return to Step 2, replace k with $k + 1$, and compute a new trial iterate.
7. Terminate the algorithm when $\|\nabla f(\mathbf{x}_k)\| < \epsilon$, where ϵ is the specified tolerance.

In the above Algorithm 2, $\nabla f(\mathbf{x})$, $H_f(\mathbf{x})$ are the gradient vector and the Hessian matrix functions of $f(\mathbf{x})$, respectively. $\operatorname{diag}(H_f(\mathbf{x}))$ is the diagonal matrix of the Hessian matrix function $H_f(\mathbf{x})$.

The weights (W) and the BC of PI-type rules of order p are presented in Tables 2, 3, 4 and 5. Here, for example, for the orbit S_{421} at $d = 6$ contains the BC $(\alpha, \alpha, \alpha, \alpha, \beta, \beta, \gamma)$, $\alpha \neq \beta \neq \gamma$, $\alpha \neq \gamma$ and their different 105 permutations. We present α in the first line and β in the second line, since γ is expressed in terms of α , β , i.e., $\gamma = 1 - 4\alpha - 2\beta$. The rules of the fifth and sixth order on tetrahedra coincide with the results of Ref. [2]. We believe that at least some of the rules presented in this paper are new. But we can not guarantee that the presented numbers of points of high-order quadrature rules are minimal. Note that up to the order $p = 6$ W and BC were calculated using Maple with 32 significant digits. For $p > 6$, W and BC were calculated using Fortran with 10 significant digits (the first three steps of Algorithm 1). These calculations were performed using the Central Information and Computer Complex, and HybriLIT heterogeneous computing cluster at JINR. Starting from the approximate values found with the Fortran code, W and BC were then calculated in Maple with 32 significant digits.

5 BVP for Helmholtz Equation in a d -dimensional Hypercube

For benchmark calculations, we use the BVP for the Helmholtz equation (HEQ) with the boundary condition (II) in a d -dimensional hypercube with the edge length π . Since the variables are separated, the eigenvalues $E_m = E_{m_1, \dots, m_d}$ are sums of squared integers, $E_m = E_{m_1, \dots, m_d} = m_1^2 + \dots + m_d^2$, $m_k = 0, 1, \dots$, $k = 1, \dots, d$.

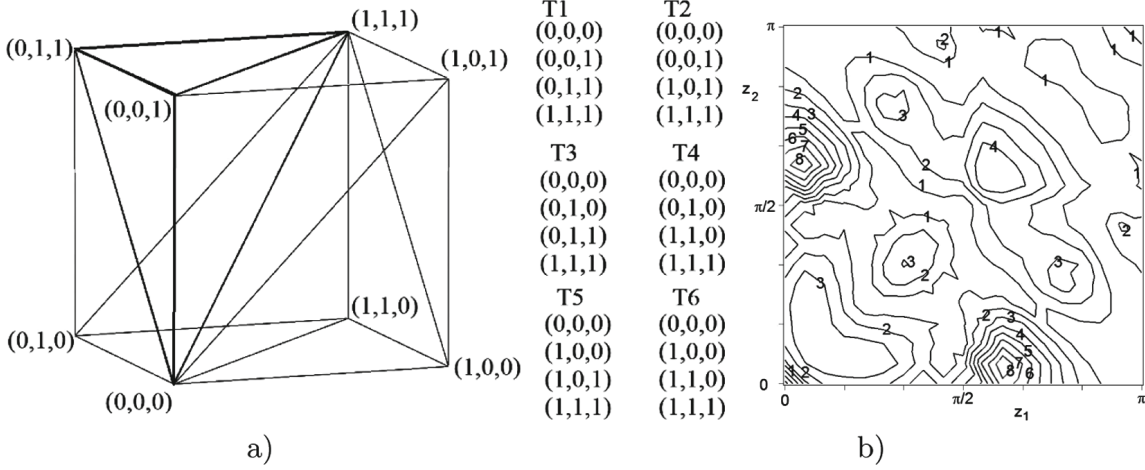


Fig. 1. (a) Division of a 3D cube into $3! = 6$ equal tetrahedrons (T1, ..., T6). (b) The error $\Delta\Phi_8(z_1, z_2, z_3) = |\Phi_8^h(z_1, z_2, z_3) - \Phi_8(z_1, z_2, z_3)|$ for the eighth eigenfunction $\Phi_8^h(z_1, z_2, z_3)$ at fixed $z_3 = \pi/9$, calculated using FEM with third-order LIPs versus the exact eigenfunction $\Phi_8(z_1, z_2, z_3)$ corresponding to the eigenvalue $E_8 = 3$. Here the cube is divided into 2^3 cubes, each comprised of 6 tetrahedrons. The isolines marked 1 correspond to the values of $\Delta\Phi_8(z_1, z_2, z_3) = \Delta\Phi_8^{\max}/10$, the isolines marked 2 correspond to the values of $\Delta\Phi_8(z_1, z_2, z_3) = 2\Delta\Phi_8^{\max}/10, \dots$, at $\Delta\Phi_8^{\max} \approx 0.018$.

Assertion (see also [16]). The hypercube is divided into $d!$ equal simplices. The vertices of each simplex are located on broken lines composed of d mutually perpendicular edges, and the extreme vertices of all polygons are located on one of the diagonals of the hypercube (for $d = 3$ see Fig. 1a).

Algorithm 3.

Input. A single d -dimensional hypercube with vertices the coordinates of which are either 0 or 1 in the Euclidean space \mathcal{R}^d . The chosen diagonal of the hypercube connects the vertices with the coordinates $(0, \dots, 0)$ and $(1, \dots, 1)$.

Output. $z_k^{(i)} = (z_{k1}^{(i)}, \dots, z_{kd}^{(i)})$, the coordinates of the i th simplex.

Local. The coordinates of the vertices of the polygonal line are $z_k = (z_{k1}, \dots, z_{kd})$, $k = 0, \dots, d$.

1. For all $i = (i_1, \dots, i_d)$, the permutations of the numbers $(1, \dots, d)$:

1.1. For all $k = 0, \dots, d$ and $s = 1, \dots, d$: $z_{k,s}^{(i)} = \{1, i_s \leq k; 0, i_s > k\}$

1.2. If $\det(z_{ks}^{(i)})_{ks=1}^d = -1$ then $z_{kd}^{(i)} \leftrightarrow z_{kd-1}^{(i)}$.

3D HEQ for the cube. In Fig. 1b, we show the error $\Delta\Phi_8(z_1, z_2, z_3)$ for the eighth eigenfunction $\Phi_8^h(z_1, z_2, z_3)$ at fixed $z_3 = \pi/9$, calculated using FEM with third-order LIPs versus the exact eigenfunction $\Phi_8(z_1, z_2, z_3)$ corresponding to the eigenvalue $E_8 = 3$. In Fig. 2a, we also show the maximal error $\Delta\Phi_8^{\max}$ for the exact eighth eigenfunction $\Phi_8(z_1, z_2, z_3)$ calculated using FEM with LIPs of the orders $p = 3, 4, 5$ versus the number N of piecewise basis functions $N_l^p(z)$ in the expansion (8). In Fig. 2b, we show the error of eigenvalues of the 3D BVP for the HEQ at $d = 3$ with the boundary condition (II) using the FEM scheme with 3D LIP of the order $p = 6$. As seen from Fig. 2, the errors of the eigenfunctions and eigenvalues lie on parallel lines in the double logarithmic scale

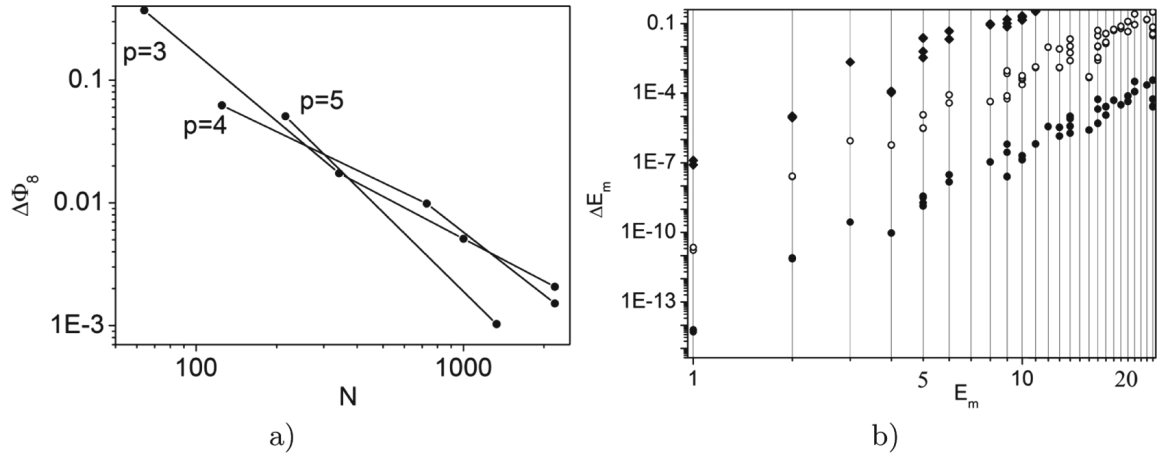


Fig. 2. (a) The maximal error $\Delta\Phi_8^{\max} = \max_{z_1 \in (0, \pi), z_2 \in (0, \pi), z_3 \in (0, \pi)} |\Phi_8^h(z_1, z_2, z_3) - \Phi_8(z_1, z_2, z_3)|$ for the exact eighth eigenfunction $\Phi_8(z_1, z_2, z_3)$ calculated using FEM with LIPs of the orders $p = 3, 4, 5$ versus the number N of piecewise basis functions $N_i^p(z)$ in the expansion (8). (b) The error $\Delta E_m = E_m^h - E_m$ calculated using FEM with sixth-order LIPs versus the exact eigenvalue E_m . Squares: the cube divided into 6 tetrahedrons. Circles: the cube divided into 2^3 cubes, each comprised of 6 tetrahedrons. Solid circles: the cube divided into 4^3 cubes, each comprised of 6 tetrahedrons.

Table 6. The lower part of the exact spectrum E_m and the calculated spectrum E_m^h for the 6D hypercube.

E_m	E_m^h
0	0.183360983479286 e-10
1	1.00023, 1.00034, 1.00034, 1.00034, 1.00034, 1.00034
2	2.04760, 2.04760, 2.04760, 2.04760, 2.04760, 2.04760, 2.04760, 2.04760, 2.04760, 2.07391, 2.08478, 2.08478, 2.08478, 2.08478
3	3.15060, 3.15196, 3.15196, 3.15196, 3.15196, 3.15196, 3.15196, 3.15780, 3.15780, 3.15780, 3.15780, 3.15780, 3.16319, 3.16319, 3.16319, 3.16319, 3.16319, 3.16319, 3.16319, 3.16319

which agrees with the theoretical error estimates (15) for the eigenfunctions and eigenvalues depending on the maximal size of the finite element. For a cube with the edge π divided into 4^3 cubes, each of them comprising 6 tetrahedrons, the matrices \mathbf{A} and \mathbf{B} had the dimension 15625×15625 . The matrices \mathbf{A} and \mathbf{B} were calculated in two ways: analytically or with Gaussian quadratures from Sect. 4 using Maple 2015, 2x 8-core Xeon E5-2667 v2 3.3 GHz, 512 GB RAM, GPU Tesla 2075. For the considered task, the values of matrix elements agree with Gaussian quadratures up to the order 10 with given accuracy. The generalized algebraic eigenvalue problem (9) was solved during 20 min using Intel Fortran.

6D HEQ for the hypercube. We solved HEQ at $d = 6$ with the boundary condition (II) using FEM scheme with 6D LIP of the order $p = 3$. The 6D hypercube having the edge π was divided into $n = d! = 6! = 720$ simplexes

(the size of the finite element being equal to π). On each of them $N_1(p) = (p+d)!/(d!p!) = 84$ third-order LIPs were used. The matrices \mathbf{A} and \mathbf{B} had the dimension 4096×4096 . The lower part of the spectrum E_m is shown in Table 6. The errors of the second, the third, and the fourth degenerate eigenvalue are equal to 0.0003, 0.05, and 0.15, respectively. Note that applying the third-order scheme for solving the BVPs of smaller dimension d , we obtained errors of the same order. The calculation time was 9234.46 s using Maple 2015.

6 Conclusion

We have elaborated new calculation schemes, algorithms, and programs for solving the multidimensional elliptic BVP using the high-accuracy FEM with simplex elements. The elaborated symbolic-numerical algorithms and programs implemented in Maple-Fortran environment calculate multivariate finite elements in the simplex and the fully symmetric PI Gaussian quadrature rules. We demonstrated the efficiency of the proposed finite element schemes, algorithms, and codes by benchmark calculations of BVPs for Helmholtz equation of cube and hypercube. The developed approach is aimed at calculations of the spectral characteristics of nuclei models and electromagnetic transitions [7, 11]. This will be done in our next publications.

Acknowledgment. The work was partially supported by the RFBR (grant No. 16-01-00080 and 18-51-18005), the MES RK (Grant No. 0333/GF4), the Bogoliubov-Infeld program, the Hulubei–Meshcheryakov program, the RUDN University Program 5-100 and grant of Plenipotentiary of the Republic of Kazakhstan in JINR. The authors are grateful to prof. R. Enkhbat for useful discussions.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1965)
2. Akishin, P.G., Zhidkov, E.P.: Some symmetrical numerical integration formulas for simplexes. Communications of the JINR 11–81-395, Dubna (1981). (in Russian)
3. Bathe, K.J.: Finite Element Procedures in Engineering Analysis. Prentice Hall, Englewood Cliffs (1982)
4. Bériot, H., Prinn, A., Gabard, G.: Efficient implementation of high-order finite elements for Helmholtz problems. Int. J. Numer. Meth. Eng. **106**, 213–240 (2016)
5. Ciarlet, P.: The Finite Element Method for Elliptic Problems. North-Holland Publishing Company, Amsterdam (1978)
6. Cui, T., Leng, W., Lin, D., Ma, S., Zhang, L.: High order mass-lumping finite elements on simplexes. Numer. Math. Theor. Meth. Appl. **10**(2), 331–350 (2017)
7. Dobrowolski, A., Mazurek, K., Gózdź, A.: Consistent quadrupole-octupole collective model. Phys. Rev. C **94**, 054322-1–054322-20 (2017)
8. Dunavant, D.A.: High degree efficient symmetrical Gaussian quadrature rules for the triangle. Int. J. Numer. Meth. Eng. **21**, 1129–1148 (1985)

9. Gusev, A.A., et al.: Symbolic-numerical algorithm for generating interpolation multivariate hermite polynomials of high-accuracy finite element method. In: Gerdt, V.P., Koepf, W., Seiler, W.M., Vorozhtsov, E.V. (eds.) CASC 2017. LNCS, vol. 10490, pp. 134–150. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66320-3_11
10. Gusev, A.A., et al.: Symbolic-numerical algorithms for solving the parametric self-adjoint 2D elliptic boundary-value problem using high-accuracy finite element method. In: Gerdt, V.P., Koepf, W., Seiler, W.M., Vorozhtsov, E.V. (eds.) CASC 2017. LNCS, vol. 10490, pp. 151–166. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66320-3_12
11. Gusev, A.A., et al.: Symbolic algorithm for generating irreducible rotational-vibrational bases of point groups. In: Gerdt, V.P., Koepf, W., Seiler, W.M., Vorozhtsov, E.V. (eds.) CASC 2016. LNCS, vol. 9890, pp. 228–242. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45641-6_15
12. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* **2**, 164–168 (1944)
13. www.maplesoft.com
14. Marquardt, D.: An algorithm for least squares estimation of parameters. *J. Soc. Ind. Appl. Math.* **11**, 431–441 (1963)
15. Maeztu, J.I., Sainz de la Maza, E.: Consistent structures of invariant quadrature rules for the n -simplex. *Math. Comput.* **64**, 1171–1192 (1995)
16. Mead, D.G.: Dissection of the hypercube into simplexes. *Proc. Am. Math. Soc.* **76**, 302–304 (1979)
17. Mysovskikh, I.P.: Interpolation Cubature Formulas. Nauka, Moscow (1981). (in Russian)
18. Papanicolopoulos, S.-A.: Analytical computation of moderate-degree fully-symmetric quadrature rules on the triangle. [arXiv:1111.3827v1](https://arxiv.org/abs/1111.3827v1) [math.NA] (2011)
19. Sainz de la Maza, E.: Fórmulas de cuadratura invariantes de grado 8 para el simplex 4-dimensional. *Revista internacional de métodos numéricos para cálculo y diseño en ingeniería* **15**(3), 375–379 (1999)
20. Strang, G., Fix, G.J.: *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs (1973)
21. Zhang, L., Cui, T.: Liu, H.: A set of symmetric quadrature rules on triangles and tetrahedra. *J. Comput. Math.* **27**, 89–96 (2009)

Kinematically complete experimental study of Compton scattering at helium atoms near the threshold

Max Kircher¹✉, Florian Trinter^{2,3}, Sven Grundmann¹, Isabel Vela-Perez¹, Simon Brennecke⁴, Nicolas Eicke⁴, Jonas Rist¹, Sebastian Eckart¹, Salim Houamer⁵, Ochbadrakh Chuluunbaatar^{6,7,8}, Yuri V. Popov^{6,9}, Igor P. Volobuev⁹, Kai Bagschik², M. Novella Piancastelli^{10,11}, Manfred Lein⁴, Till Jahnke¹, Markus S. Schöffler¹ and Reinhard Dörner¹✉

Compton scattering is one of the fundamental interaction processes of light with matter. When discovered¹, it was described as a billiard-type collision of a photon ‘kicking’ a quasi-free electron. With decreasing photon energy, the maximum possible momentum transfer becomes so small that the corresponding energy falls below the binding energy of the electron. In this regime, ionization by Compton scattering becomes an intriguing quantum phenomenon. Here, we report on a kinematically complete experiment studying Compton scattering off helium atoms in that regime. We determine the momentum correlations of the electron, the recoiling ion and the scattered photon in a coincidence experiment based on cold target recoil ion momentum spectroscopy, finding that electrons are not only emitted in the direction of the momentum transfer, but that there is a second peak of ejection to the backward direction. This finding links Compton scattering to processes such as ionization by ultrashort optical pulses², electron impact ionization^{3,4}, ion impact ionization^{5,6} and neutron scattering⁷, where similar momentum patterns occur.

Doubts about energy conservation in Compton scattering at the single-event level motivated the invention, by Bothe and Geiger⁸, of coincidence measurement techniques. This historic experiment settled the dispute about the validity of conservation laws in quantum physics by showing that, for each scattered photon, there is an electron ejected in coincidence. Surprisingly, however, even 95 years after this pioneering work, coincidence experiments on the Compton effect are extremely scarce and are restricted to solid-state systems^{9,10}. To a large extent, this lack of detailed experiments left further progress in the field of Compton scattering to theory. Due to missing experimental techniques, much of the potential of using Compton scattering as a tool in molecular physics remained untapped¹¹. The small cross-section of 10^{-24} cm² (six orders of magnitude below typical photoabsorption cross-sections at the respective thresholds), together with the small collection solid angle of typical photon detectors, has so far prohibited coincidence experiments on free atoms and molecules. In the present work, we have

solved this problem by using the highly efficient cold target recoil ion momentum spectroscopy (COLTRIMS) technique¹² to detect the electron and ion momentum in coincidence. The He⁺ ion and electrons with an energy smaller than 25 eV are detected with 4π collection solid angle. The momentum vector of the scattered photon can be obtained using momentum conservation, thereby circumventing the need for a photon detector. This allows us to obtain a kinematically complete dataset of ionization by Compton scattering of atoms, addressing the intriguing low-energy, near-threshold regime. It has often been pointed out in the theoretical literature that such complete measurements of the process—as opposed to detection of the emitted electron or scattered photon only—are the essential key to sensitive testing of theories¹³ as well as allowing for a clean physics interpretation of the results¹⁴.

For the case of Compton scattering at a quasi-free electron, the angular distribution of the scattered photon is given by the Thomson cross-section (Fig. 1a). Binding of the electron modifies the binary scattering scenario by adding the ion as a third particle. The often invoked impulse approximation accounts for one of the effects of that binding, namely the electron's initial momentum distribution. According to this approximation, the initial electron momentum is added to the momentum balance, while the binding energy is neglected. In this model, the ion momentum is defined such that it compensates only for the electron's initial momentum. The impulse approximation works well when the binding energy is negligible compared to the energy of the electron carrying the momentum Q transferred by the photon. The maximum value of Q is reached for photon backscattering, and is twice the photon momentum E_1/c , where E_1 is the incoming photon energy. For helium with a binding energy of 24.6 eV, this gives a threshold of $E_1 \approx 2.5$ keV, below which photon backscattering at an electron at rest does not provide enough energy to overcome the ionization threshold. In the present experiment, we use a photon energy of $E_1 = 2.1$ keV, well below that threshold. Accordingly, the cross-section for ionization by Compton scattering has dropped to $\sim 20\%$ of its maximum value of $\sim 10^{-24}$ cm² (ref. ¹⁵). As expected, we observe that the photon scattering angular

¹Institut für Kernphysik, J. W. Goethe Universität, Frankfurt, Germany. ²FS-PETRA-S, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany.

³Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany. ⁴Institut für Theoretische Physik, Leibniz Universität Hannover, Hannover, Germany. ⁵LPQSD, Department of Physics, Faculty of Science, University Sétif-1, Setif, Algeria. ⁶Joint Institute for Nuclear Research, Dubna, Moscow, Russia. ⁷Institute of Mathematics and Digital Technologies, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia. ⁸Peoples' Friendship University of Russia (RUDN University), Moscow, Russia. ⁹Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia. ¹⁰Sorbonne Universités, CNRS, UMR 7614, Laboratoire de Chimie Physique Matière et Rayonnement, Paris, France. ¹¹Department of Physics and Astronomy, Uppsala University, Uppsala, Sweden. ✉e-mail: kircher@atom.uni-frankfurt.de; doerner@atom.uni-frankfurt.de

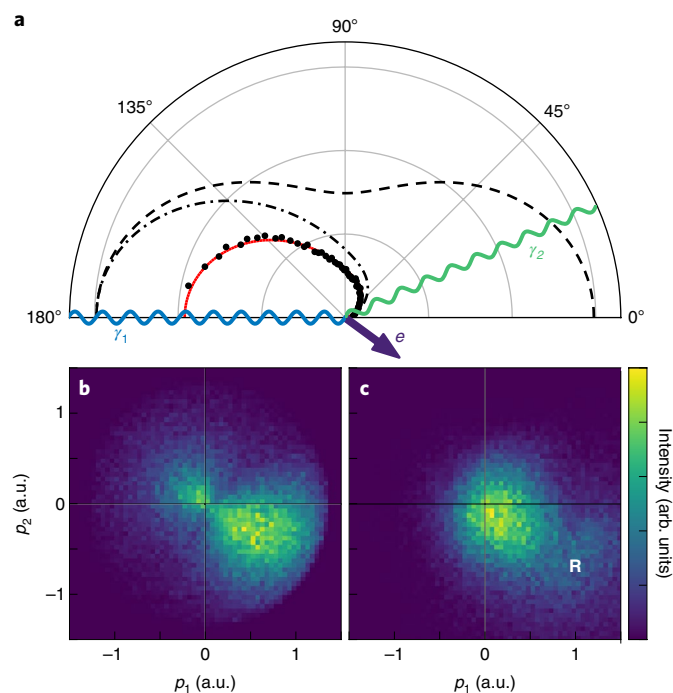


Fig. 1 | Scheme of ionization by Compton scattering at $h\nu = 2.1$ keV.

a, The wavy lines indicate the incoming and outgoing photon, and the purple arrow depicts the momentum vector of the emitted electron. The dashed line shows the Thomson cross-section, that is, the angular distribution of a photon scattering at a free electron. Black dots show the experimental photon angular distribution for ionization of He by Compton scattering, integrated over all electron emission angles and energies below 25 eV. The photon momenta are determined using the electron and ion momenta, as well as momentum conservation. The statistical error is smaller than the dot size. The dash-dotted line shows the A^2 approximation for all electron energies and the solid red line shows the A^2 approximation for electron energies below 25 eV. The calculations were done using Approach I (see Methods). The solid and dash-dotted lines are multiplied by a factor of 1.9. **b**, Momentum distribution of electrons emitted by Compton scattering of 2.1 keV photons at He. The coordinate frame is the same as in **a**: the scattering plane is defined by the incoming (horizontal) and scattered photon (upper half plane); that is, p_1 is the electron momentum component in the \mathbf{k}_i direction and p_2 is the component perpendicular to \mathbf{k}_i , within the scattering plane. The momentum transfer points to the forward lower half plane. The data are integrated over the out-of-plane electron momentum components. **c**, He⁺ ion momentum distribution for the same conditions as in **b**. See main text for an explanation of the feature **R**.

distribution differs significantly from the Thomson cross-section (Fig. 1a). The most striking difference is that all forward angles of photon emission are suppressed and it is almost only backscattered photons that lead to ionization. This measured cross-section shows excellent agreement with our theoretical model, which is described in detail in the Methods.

What is the mechanism facilitating ionization at these low photon energies and small momentum transfers? Our coincidence experiment can answer this question by providing the momentum vectors of all particles, that is, the incoming (\mathbf{k}_i) and outgoing (\mathbf{k}_s) photon, electron (\mathbf{p}_e) and ion (\mathbf{p}_{ion}) momentum vectors for each individual Compton ionization event. This event-by-event momentum correlation gives access to the various particles' momentum distributions in the intrinsic coordinate frame of the process, which is a plane spanned by the wavevectors of the incoming and scattered photon

(Fig. 1). This plane also contains the momentum transfer vector $\mathbf{Q} = \mathbf{k}_i - \mathbf{k}_s$. In Fig. 1b,c, by definition, the photon is scattered to the upper half plane and the momentum transfer \mathbf{Q} (that is, the 'kick' by the photon) points forward and into the lower half plane. The electron momentum distribution visualized in this intrinsic coordinate frame shows two distinct islands, one in the direction of the momentum transfer and a second smaller one to the backward direction, that is, opposite to the momentum transfer direction. These two maxima are separated by a minimum. The He⁺ ions (Fig. 1c) are also emitted to the forward direction. In addition to a main island close to the origin, ions are also emitted strongly in the forward direction, towards the region indicated by **R** (Recoil) in Fig. 1c. This ion momentum distribution shows strikingly that in the below-threshold regime, the situation is very different from the quasi-free electron scattering considered in the standard high-energy Compton process. In the latter case, the ion is only a passive spectator to the photon–electron interaction and, consequently, the ion momenta are centred at the origin of the coordinate frame used in Fig. 1b,c^{15–18}.

The observed bimodal electron momentum distribution becomes even clearer when we examine a subset of the data for which the photon is scattered to a certain direction (Fig. 2). This shows that the momentum distribution follows the direction of momentum transfer and the nodal plane is perpendicular to \mathbf{Q} . Such bimodal distributions are known from different contexts. For example, for ionization by electron impact ($e, 2e$)⁴ and ion impact⁵, the forward lobe has been termed a binary lobe, for obvious reasons, while the backward peak is referred to as the recoil peak. This latter name alludes to the fact that, for the electron to be emitted in a direction opposite the momentum transfer, momentum conservation dictates that the ion recoils in the opposite direction. Mechanistically, this would occur if the electron was initially kicked in the forward direction but then back-reflected at its own parent ion. Such a classical picture would suggest that the ion receives the momentum originally imparted to the electron (that is, \mathbf{Q}) minus the final momentum, \mathbf{p}_e , of the electron. This expectation is verified by our measured ion momentum distributions (Fig. 2g–i). The ions also show a bimodal momentum distribution, with the main island slightly forward shifted and a minor island significantly forward shifted in the momentum transfer direction, in nice agreement with the back-reflection scheme.

The observations suggest a two-step model for below-threshold Compton scattering, which is referred to as the A^2 approximation (see Methods). The first step is the scattering of the photon at an electron being described by the Thomson cross-section. This step sets the direction and magnitude of the approximate momentum transfer. The second step is the response of the electron wavefunction to this sudden kick, which displaces the bound wavefunction in momentum space. This momentum-shifted electron wavefunction then relaxes to the electronic eigenstates of the ion, where it has some overlap with its initial state and with the bound excited states. However, the fraction that overlaps with the Coulomb continuum leads to ionization and is observed experimentally. The bimodal electron momentum distribution for small momentum transfer follows naturally from such a scenario. The leading ionizing term in the Taylor expansion of the momentum transfer operator $e^{i\mathbf{Q}\cdot\mathbf{r}_e}$ is the dipole operator, with the momentum transfer replacing the direction of polarization. This dipolar contribution, resembling the shape of a p orbital, is the origin of the bimodal electron momentum distribution.

The observed electron momentum distributions are in excellent agreement with the prediction of the A^2 approximation shown in Fig. 2a–c. Note that these theoretical distributions are calculated without any reference to Compton scattering. What is shown is the overlap of the ground state with the continuum (altered by the momentum transfer). Exactly the same distributions are predicted for an attosecond half-cycle pulse (see fig. 2 in ref. 2) and identical

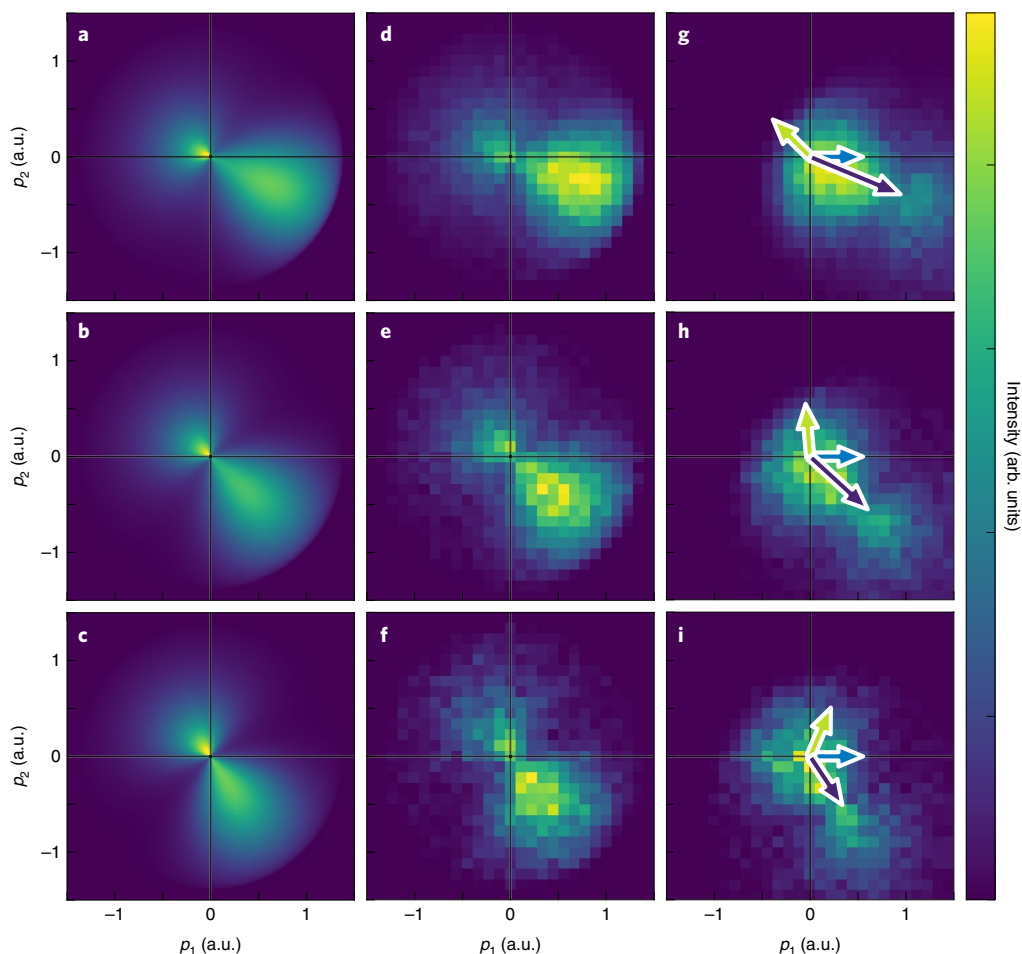


Fig. 2 | Electron and ion momentum distributions for different momentum transfer gates. In all panels, p_1 is the momentum component in the \mathbf{k}_i direction, p_2 is the component perpendicular to \mathbf{k}_i within the scattering plane. **a–c**, Electron momentum distributions obtained from modelling within the A^2 approximation using Approach II (see Methods). **d–f**, Electron momentum distributions measured by our experiment. **g–i**, Measured momentum distributions of the ions. From top to bottom, the rows correspond to different momentum transfers $Q=1.0$ (**a,d,g**), 0.8 (**b,e,h**) and 0.6 (**c,f,i**) a.u., respectively. Arrows in the third column indicate the photon momentum configuration for each row. Blue arrows represent the momentum of the incoming photon, light green arrows the momentum of the scattered photon and dark purple arrows the momentum transfer. A video of the electron and ion momentum distributions for different photon scattering directions is provided in the Supplementary Information.

results are expected for a momentum transfer to the nucleus by neutron scattering⁷.

Within the A^2 approximation, the magnitude of the energy transfer is determined by energy conservation. It is worth mentioning that, under the present conditions, the photon loses only a few percent of its primary energy. Thus the momentum transfer is largely a consequence of the angular deflection of the photon and not a consequence of its change in energy. This can be seen by inspecting the energy distribution of the ejected electron in Fig. 3a. The electron energy distribution peaks at zero and falls off exponentially. For electron forward emission (Fig. 3b) it peaks at 11 eV for photon backscattering, while the backward-emitted electrons for the same conditions are much lower in energy (Fig. 3c). This also manifests itself in the fully differential cross-section (FDCS) showing the electron angular distribution for fixed electron energy and a fixed photon scattering angle of $150 \pm 20^\circ$. These angular distributions (Fig. 4) show that the intensity in the backward-directed recoil lobe drops strongly with increasing electron energy compared to the intensity in the forward-directed binary lobe. The physics governing the relative strength of the binary and recoil lobes is unveiled by two sets of calculations by comparing theoretical calculations for different initial electron wavefunctions and different final states.

First, we use a correlated two-electron wavefunction in the initial state, with outgoing Coulomb waves with charge 1 as the final state. Second, we use a single-active-electron model for the initial state, with a final scattering state in an effective potential (Figs. 3 and 4). We find that the binary peak is similar in all cases. However, the recoil peak is enhanced by more than a factor of two when scattering states in an effective He^+ potential are used instead of Coulomb states. This directly supports the mechanistic argument that the recoil peak originates from backscattering of forward-kicked electrons at the parent ion. This backscattering is enhanced due to the increased depth of the effective potential compared to the Coulomb potential close to the origin. The intensity of the recoil peak of both approaches deviates from our experimental data, whereas the shape is predicted correctly by theory. This hints towards the importance of both theoretical approaches (a more detailed discussion is provided in the Methods).

In conclusion, we have shown the first FDCSs for Compton scattering at a gas-phase atom, unveiling the mechanism of near-threshold Compton scattering. Our experimental work shows good agreement with our theoretical models, but further studies with more sophisticated theoretical models are necessary. This work can function as a benchmark measurement for such studies.

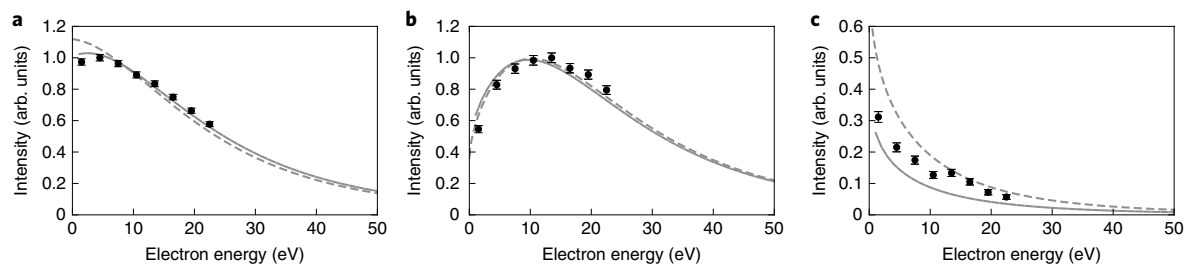


Fig. 3 | Electron energy distribution. The scattering angle between the incoming and outgoing photon for the outgoing photon is restricted to $140 < \theta < 180^\circ$ in all panels. **a**, The electron energy spectrum is shown independent of the electron emission direction. **b**, The electron emission angle is restricted to forward scattering ($0 < \theta_e < 40^\circ$). **c**, The electron emission angle is restricted to backward scattering ($140 < \theta_e < 180^\circ$). The black dots are the experimental data. The error bars represent the standard statistical error. The solid lines are the theoretical results of Approach I and the dashed lines are the results of Approach II (see Methods). The experimental data in **a** and **b** are normalized such that the maximum intensity is 1; the theory is normalized such that the integrals of the experimental data and the theoretical curves are equal. The normalization factors in **c** are identical to those in **b**, because here we depict the forward/backward direction of the same distribution.

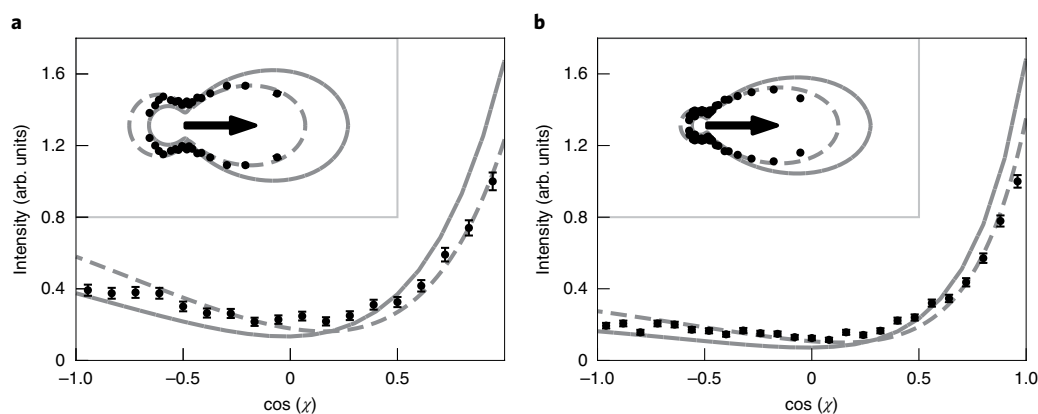


Fig. 4 | Fully differential electron angular distributions. **a, b**, The photon scattering angle is $130 < \theta < 170^\circ$. Displayed is the cosine of the angle χ between the outgoing electron and the momentum transfer \mathbf{Q} for electron energies of $1.0 < E_e < 3.5 \text{ eV}$ (**a**) and $3.5 < E_e < 8.5 \text{ eV}$ (**b**). Insets show the same data in polar representation, where the arrow indicates the direction of momentum transfer. Black dots are the experimental data, normalized such that the maximum is 1. Error bars represent the standard statistical error. The solid and dashed lines are the theoretical curves resulting from Approach I and Approach II, respectively. The theoretical curves are normalized such that the integral of experiment and theory are equal.

Coincidence detection of ions and electrons, as demonstrated here, paves the road to exploit Compton scattering for imaging of molecular wavefunctions not only averaged over the molecular axis but also in the body-fixed frame of the molecule. For slightly higher momentum transfers \mathbf{Q} , that is, photon energies of $\sim 6 \text{ keV}$, one can expect the significance of correlations in the scattering states to diminish, simplifying the theoretical description. As has been pointed out recently, measuring the momentum transfer to the nucleus in this case will give access to the Dyson orbitals¹¹.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41567-020-0880-2>.

Received: 12 November 2019; Accepted: 13 March 2020;

Published online: 13 April 2020

References

- Compton, A. H. in *Bulletin of the National Research Council No. 20* Vol. 4, Pt. 2 (National Research Council of the National Academy of Sciences, 1922)

- Arbó, D. G., Tökési, K. & Miraglia, J. E. Atomic ionization by a sudden momentum transfer. *Nucl. Instr. Methods Phys. Res. B* **267**, 382–385 (2009).
- Dürr, M. et al. Single ionization of helium by 102 eV electron impact: three dimensional images for electron emission. *J. Phys. B* **39**, 4097–4111 (2006).
- Ehrhardt, H., Jung, K., Knoth, G. & Schlemmer, P. Differential cross section of direct single electron impact ionization. *Z. Phys. D Atoms Mol. Clusters* **1**, 3–32 (1986).
- Fischer, D., Moshhammer, R., Schulz, M., Voitkiv, A. & Ullrich, J. Fully differential cross sections for the single ionization of helium by ion impact. *J. Phys. B* **36**, 3555–3567 (2003).
- Schulz, M. et al. Three-dimensional imaging of atomic four-body processes. *Nature* **422**, 48–50 (2003).
- Pindzola, M. S. et al. Neutron-impact ionization of He. *J. Phys. B* **47**, 195202 (2014).
- Bothe, W. & Geiger, H. Über das Wesen des Comptoneffekts; ein experimenteller Beitrag zur Theorie der Strahlung. *Z. Phys.* **32**, 639–663 (1925).
- Bell, E., Tschentscher, T. H., Schneider, J. R. & Rollason, A. J. The triple differential cross section for deep inelastic photon scattering: a $(\gamma, e\gamma)$ experiment. *J. Phys. B* **24**, L533–L538 (1991).
- Metz, C. et al. Three-dimensional electron momentum density of aluminum by $(\gamma, e\gamma)$ spectroscopy. *Phys. Rev. B* **59**, 10512–10520 (1999).
- Hopersky, A. N., Nadolinsky, A. M., Novikov, S. A., Yavna, V. A. & Ikoeva, K. K. H. X-ray-photon Compton scattering by a linear molecule. *J. Phys. B* **48**, 175203 (2015).
- Ullrich, J. et al. Recoil-ion and electron momentum spectroscopy: reaction-microscopes. *Rep. Prog. Phys.* **66**, 1463–1545 (2003).

13. Roy, S. C. & Pratt, R. H. Need for further inelastic scattering measurements at X-ray energies. *Radiat. Phys. Chem.* **69**, 193–197 (2004).
14. Kaliman, Z., Surić, T., Pisk, K. & Pratt, R. H. Triply differential cross section for Compton scattering. *Phys. Rev. A* **57**, 2683–2691 (1998).
15. Samson, J. A. R., He, Z. X., Bartlett, R. J. & Sagurton, M. Direct measurement of He⁺ ions produced by Compton scattering between 2.5 and 5.5 keV. *Phys. Rev. Lett.* **72**, 3329–3331 (1994).
16. Spielberger, L. et al. Separation of photoabsorption and Compton scattering contributions to He single and double ionization. *Phys. Rev. Lett.* **74**, 4615–4618 (1995).
17. Dunford, R. W., Kanter, E. P., Krässig, B., Southworth, S. H. & Young, L. Higher-order processes in X-ray photoionization and decay. *Radiat. Phys. Chem.* **70**, 149–172 (2004).
18. Kaliman, Z. & Pisk, K. Compton cross-section calculations in terms of recoil-ion momentum observables. *Rad. Phys. Chem.* **71**, 633–635 (2004).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Experimental methods. The experiment was performed at beamline P04 of synchrotron PETRA III, DESY in Hamburg, with 40-bunch timing mode; that is, the photon bunches were spaced 192 ns apart. A circularly polarized pink beam was used; that is, the monochromator was set to zero order. To effectively remove low-energy photons from the beam, we put foil filters in the photon beam, namely 980 nm of aluminium, 144 nm of copper and 153 nm of iron. With this set-up, we suppressed photons of <100 eV by at least a factor of 10^{-9} and photons <15 eV by at least a factor of 10^{-25} (data based on ref. ¹⁹, 9 October 2019, obtained from http://henke.lbl.gov/optical_constants/filter2.html). The beam was crossed at a 90° angle with a supersonic gas jet, expanding through a 30 μm nozzle at 30 bar driving pressure and room temperature within a COLTRIMS spectrometer. The supersonic gas jet passed two skimmers (0.3 mm diameter), so the reaction region had approximate dimensions of $0.2 \times 1.0 \times 0.1$ mm³. The electron side of the spectrometer had 5.8 cm of acceleration. To increase the resolution, an electrostatic lens and time-of-flight focusing geometry were used for the ion side to effectively compensate for the finite size of the reaction region. The total length of the ion side was 97.4 cm. The electric field in the spectrometer was 18.3 V cm⁻¹ and the magnetic field was 9.1 G. The charged particles were detected using two position-sensitive microchannel plate detectors with delay-line anodes²⁰.

Theoretical methods. In general, Compton scattering is a relativistic process. In the special case of an initially bound electron, this process may be described by the second-order quantum electrodynamics perturbation terms with exchange in the presence of an external classical electromagnetic field due to the residual ion (see for example ref. ²¹). In the low-energy limit of small incoming photon energy E_i , compared to the remaining energy of an electron, $m_e c^2$, we can apply a non-relativistic quantum-mechanical description^{22,23}. A modern presentation of this approach is provided in ref. ²⁴. (In the following, we use atomic units unless stated otherwise; that is, $e = m_e = \hbar = 1$.) The energy and momentum conservation laws are of the form

$$E_1 = E_2 + I_p + E_e + E_{\text{ion}}, \quad \mathbf{k}_1 = \mathbf{k}_2 + \mathbf{p}_e + \mathbf{p}_{\text{ion}} \quad (1)$$

where I_p is the ionization potential, $E_e(\mathbf{p}_e)$ is the energy (momentum) of the escaped electron, $E_{\text{ion}}(\mathbf{p}_{\text{ion}})$ is the energy (momentum) of the residual ion and $E_{i/2}(\mathbf{k}_{i/2})$ are the energies (momenta) of the incoming and outgoing photons, respectively. For the given keV photon energy range, the momenta are of the order $k_i = E_i/c \sim 1$ a.u. with the speed of light $c = \alpha^{-1}$, so that the energy of the escaped electron is only a few eV. Given that $M_{\text{ion}} \gg 1$, the ionic kinetic energy $E_{\text{ion}} = \mathbf{p}_{\text{ion}}^2/(2M_{\text{ion}})$ can be neglected. Hence, the photon energy is nearly unchanged and the ratio of photon energy after and before the collision is

$$t = \frac{E_2}{E_1} = 1 - \frac{I_p + E_e + E_{\text{ion}}}{E_1} \approx 1 \quad (2)$$

The transferred momentum from the photon to the atomic system is given by $\mathbf{Q} = \mathbf{k}_1 - \mathbf{k}_2 = \mathbf{p}_e + \mathbf{p}_{\text{ion}}$. The magnitude and direction of the transferred momentum \mathbf{Q} may be expressed as a function of the scattering angle θ between the incoming and outgoing photon.

Under the above kinematic conditions, the FDSC may be written as

$$\frac{d\sigma}{dE_e d\Omega_e d\Omega_2} = r_e^2 p_e t |M|^2 \quad (3)$$

with the classical electron radius r_e . In this Letter, we use only the so-called A^2 (seagull) term from the total second-order Kramers–Heisenberg–Waller matrix element, as is presented, for example, in ref. ²⁴:

$$M(\mathbf{Q}, \mathbf{p}_e) = (\mathbf{e}_1 \cdot \mathbf{e}_2) \langle \Psi_{\mathbf{p}_e}^{(-)} | \sum_{j=1}^N e^{i\mathbf{Q} \cdot \mathbf{r}_j} | \Psi_0 \rangle \quad (4)$$

Here, $\mathbf{e}_{i/2}$ are the polarization vectors of the incoming and outgoing photons. Initially, the N electrons of the system with positions \mathbf{r}_j are in the bound state Ψ_0 . Given that, in the detection scheme, we select singly ionized helium ions, the final state of the electronic system is a scattering state $\Psi_{\mathbf{p}_e}^{(-)}$ with one electron in the continuum (corresponding to an asymptotic electron momentum \mathbf{p}_e) and the other electron remaining bound.

Assuming an unpolarized incoming photon beam and that we do not detect the final polarization state of the outgoing photon, we also average over the initial polarization and sum the probabilities corresponding to both possible orthogonal polarization states. Under these assumptions, the FDSC can be written as

$$\frac{d\sigma}{dE_e d\Omega_e d\Omega_2} = \left(\frac{d\sigma}{d\Omega_2} \right)_{\text{Th}} p_e t |M_e|^2 \quad (5)$$

with the Thompson cross-section

$$\left(\frac{d\sigma}{d\Omega_2} \right)_{\text{Th}} = \frac{1}{2} r_e^2 (1 + \cos^2 \theta) \quad (6)$$

for photons scattered off a single free electron and the electronic matrix element

$$M_e(\mathbf{Q}, \mathbf{p}_e) = \langle \Psi_{\mathbf{p}_e}^{(-)} | \sum_{j=1}^N e^{i\mathbf{Q} \cdot \mathbf{r}_j} | \Psi_0 \rangle \quad (7)$$

From the FDSC, the different observables shown in the main text can be calculated. The A^2 approximation resembles the first Born approximation for scattering of a fast particle on an atom, for example ($e, 2e$) ionization by electron impact¹. Therefore, the observed effects have an analogous interpretation and can be described in familiar terms. However, the Compton ionization has some advantages compared to traditional methods such as ($e, 2e$) ionization: (1) the contribution of other second-order terms is very small, so the A^2 approximation is often accurate; (2) the photon has no charge, so we only need to consider the evolution of the field-free system of charged particles; (3) the transferred momentum \mathbf{Q} can vary in a wide range, so different regimes are accessible.

Compton scattering by a bound electron is a sequential process and may be divided into two steps. In the first, the incoming photon is captured by a bound electron. Afterwards, this dressed system evolves in time so that a photon is emitted and an electron escapes. In the A^2 approximation, the second photon is emitted immediately after absorption, so this short photon scattering process can be effectively interpreted as a ‘kick’ of the electronic bound-state distribution by the transferred momentum \mathbf{Q} . The corresponding scattering probability is described by the Thompson formula. The ‘kicked’, field-free atomic system evolves in time. One part of the boosted wavefunction remains bound, while the other part is set free in the continuum and causes ionization. In principle, the time evolution, including the interaction between electrons and their possible correlation, is implicitly contained in the scattering state $\Psi_{\mathbf{p}_e}^{(-)}$ in equation (7). However, the calculation of fully correlated scattering states is beyond the scope of this work.

To calculate the electronic matrix elements, complementary approaches have been used: the first model (Approach I) describes both electrons and takes into account correlation in the ground state, but uses Coulomb waves as scattering states. In contrast, the second model (Approach II) uses a single-active-electron description, but includes accurate one-electron scattering states.

Approach I: model with correlated ground state. In the first approach, both electrons of the helium atom are explicitly treated such that the ‘direct’ ionization of the ‘kicked’ electron as well as the ‘shake-off’ (that is, ejection of the unknicked electron) are considered. In equation (7), the initial state is given by a correlated symmetric two-electron ground state $\Psi_0(\mathbf{r}_1, \mathbf{r}_2)$, obtained from ref. ²³. To approximate the final state, the main idea is that one electron remains bound in the ionic ground state given by

$$\psi_0^{\text{He}^+}(\mathbf{r}) = \sqrt{\frac{8}{\pi}} e^{-2r} \quad (8)$$

and the free electron may be approximated by Coulomb wavefunctions

$$\psi_{\mathbf{p}_e}^{\text{C}}(\mathbf{r}) = \sqrt{\frac{e^{-\pi\zeta}}{(2\pi)^3}} \Gamma(1 - i\zeta) e^{i\mathbf{p}_e \cdot \mathbf{r}} {}_1F_1(i\zeta, 1 - i\mathbf{p}_e r - i\mathbf{p}_e \cdot \mathbf{r}) \quad (9)$$

with $\zeta = -1/p_e$ and ${}_1F_1$ being the confluent hypergeometric function. Because the correct scattering states $\Psi_{\mathbf{p}_e}^{(-)}(\mathbf{r}_1, \mathbf{r}_2)$ have to be orthogonal to the initial bound states, the resulting symmetrized final state

$$\tilde{\Psi}_{\mathbf{p}_e}^{(-)}(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}} \left[\psi_{\mathbf{p}_e}^{\text{C}}(\mathbf{r}_1) \psi_0^{\text{He}^+}(\mathbf{r}_2) + \psi_{\mathbf{p}_e}^{\text{C}}(\mathbf{r}_2) \psi_0^{\text{He}^+}(\mathbf{r}_1) \right] \quad (10)$$

is afterwards explicitly orthogonalized with respect to the initial state Ψ_0 such that the electronic matrix elements of equation (7) read

$$\begin{aligned} M_e(\mathbf{Q}, \mathbf{p}_e) &= \langle \Psi_{\mathbf{p}_e}^{(-)} | e^{i\mathbf{Q} \cdot \mathbf{r}_1} + e^{i\mathbf{Q} \cdot \mathbf{r}_2} | \Psi_0 \rangle \\ &= \langle \tilde{\Psi}_{\mathbf{p}_e}^{(-)} | e^{i\mathbf{Q} \cdot \mathbf{r}_1} + e^{i\mathbf{Q} \cdot \mathbf{r}_2} | \Psi_0 \rangle - \\ &\quad \langle \tilde{\Psi}_{\mathbf{p}_e}^{(-)} | \Psi_0 \rangle \langle \Psi_0 | e^{i\mathbf{Q} \cdot \mathbf{r}_1} + e^{i\mathbf{Q} \cdot \mathbf{r}_2} | \Psi_0 \rangle \end{aligned} \quad (11)$$

Approach II: single-active-electron model. In the second approach only the ‘kicked’ electron may escape, while the other electron stays frozen at the core. To model the influence of the remaining electron on the escaping electron, we use a single-active-electron effective potential²⁴. This potential has an asymptotic charge of $Z=2$ for $r \rightarrow 0$, which is screened by the second electron such that, asymptotically for large r , it reaches $Z=1$. The one-electron ground state ψ_0 and the one-electron continuum state $\psi_{\mathbf{p}_e}^{(-)}$ with incoming boundary conditions are calculated numerically by solving the radial Schrödinger equation. Hence, the electronic matrix element in equation (7) is approximated as

$$M_e(\mathbf{Q}, \mathbf{p}_e) = \sqrt{2} \langle \psi_{\mathbf{p}_e}^{(-)} | e^{i\mathbf{Q} \cdot \mathbf{r}} | \psi_0 \rangle \quad (12)$$

This expression is calculated using a plane wave expansion of $e^{i\mathbf{Q}\cdot\mathbf{r}}$ and an expansion of the scattering states $\psi_{\mathbf{p}_e}^{(-)}$ in terms of spherical harmonics.

Both approaches use two main approximations. (1) The final scattering states are not the exact fully correlated states. This leads to deviations in the low-energy region at the recoil peak. In particular, 'shake-off' and 'shake-up' processes are not fully included. To some extent, correlations are included due to the orthogonalization in Approach I and the effective potential that is used in Approach II. However, we believe that including correlations in the final state in a more systematic way is more important than in the ground state. (2) The state of the residual ion has not been resolved in the experiment. In Approach I, we assume that the bound electron remains in the ground state of the ion, whereas it is simply frozen in the ground state of the atom in Approach II. We expect that this works well for the binary peak (forward direction), but not for the recoil peak (backward direction). To improve the calculations, ionization in different channels corresponding to excited states of the residual ion need to be considered.

Data availability

The data that support the plots within this Letter are available from the corresponding authors upon reasonable request.

Code availability

The code that supports the theoretical plots within this Letter is available from the corresponding authors upon reasonable request.

References

- Henke, B. L., Gullikson, E. M. & Davis, J. C. X-ray interactions: photoabsorption, scattering, transmission and reflection at $E=50\text{--}30,000\text{ eV}$, $Z=1\text{--}92$. *At. Data Nucl. Data Tables* **54**, 181–342 (1993).
- Jagutzki, O. et al. Multiple hit readout of a microchannel plate detector with a three-layer delay-line anode. *IEEE Trans. Nucl. Sci.* **49**, 2477–2483 (2002).
- Akhiezer, A. I. & Berestetskii, V. B. *Quantum Electrodynamics* (Wiley, 1965).
- Bergstrom, P. M. Jr, Surić, T., Pisk, K. & Pratt, R. H. Compton scattering of photons from bound electrons: full relativistic independent-particle-approximation calculations. *Phys. Rev. A* **48**, 1134–1162 (1993).
- Chuluunbaatar, O. et al. Role of the cusp conditions in electron–helium double ionization. *Phys. Rev. A* **74**, 014703 (2006).
- Tong, X. M. & Lin, C. D. Empirical formula for static field ionization rates of atoms and molecules by lasers in the barrier-suppression regime. *J. Phys. B* **38**, 2593–2600 (2005).

Acknowledgements

This work was supported by DFG and BMBF. O.C. acknowledges support from the Hulubei-Meshcheryakov programme JINR-Romania and the RUDN University Program 5-100. Y.V.P. is grateful to the Russian Foundation of Basic Research (RFBR) for financial support under grant no. 19-02-00014a. S.H. thanks the Direction Generale de la Recherche Scientifique et du Developpement Technologique (DGRSDT-Algeria) for financial support. We are grateful to the staff of PETRA III for excellent support during the beam time. Calculations were performed on the Central Information and Computer Complex and heterogeneous computing platform HybriLIT through supercomputer 'Govorun' of JINR.

Author contributions

M.K., F.T., S.G., I.V.-P., J.R., S.E., K.B., M.N.P., T.J., M.S.S. and R.D. contributed to the experimental work. S.B., N.E., S.H., O.C., Y.V.P., I.P.V. and M.L. contributed to theory and numerical simulations. All authors contributed to the manuscript.

Competing interests

The authors declare no competing interests.



Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41567-020-0880-2>.

Correspondence and requests for materials should be addressed to M.K. or R.D.

Peer review information *Nature Physics* thanks Steven Manson, Andre Staudte and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Near-barrier heavy-ion fusion: Role of boundary conditions in coupling of channelsP. W. Wen *Joint Institute for Nuclear Research, 141980 Dubna, Russia
and China Institute of Atomic Energy, 102413 Beijing, China*O. Chuluunbaatar *Joint Institute for Nuclear Research, 141980 Dubna, Russia
and Institute of Mathematics and Digital Technologies, Mongolian Academy of Sciences, 13330 Ulaanbaatar, Mongolia*A. A. Gusev *Joint Institute for Nuclear Research, 141980 Dubna, Russia*R. G. Nazmitdinov *Joint Institute for Nuclear Research, 141980 Dubna, Russia
and Dubna State University, 141982 Dubna, Russia*A. K. Nasirov *Joint Institute for Nuclear Research, 141980 Dubna, Russia
and Institute of Nuclear Physics, Ulugbek, 100214, Tashkent, Uzbekistan*S. I. Vinitsky *Joint Institute for Nuclear Research, 141980 Dubna, Russia
and Peoples' Friendship University of Russia (RUDN University), 117198 Moscow, Russia*C. J. Lin *China Institute of Atomic Energy, 102413 Beijing, China
and Department of Physics, Guangxi Normal University, 541004 Guilin, China*

H. M. Jia

China Institute of Atomic Energy, 102413 Beijing, China

(Received 24 July 2019; published 23 January 2020)

The problem of a quantum-mechanical description of a near-barrier fusion of heavy nuclei, that occurs at strong coupling of their relative motion to surface vibrations, is analyzed. To this end, an efficient finite-element method is proposed for numerically solving coupled Schrödinger equations with boundary conditions corresponding to total absorption. The method allows us to eliminate the instabilities in the numerical solutions that appear at a large number of coupled channels in some reactions. To illustrate the validity of our approach, the results of fusion cross section of the $^{64}\text{Ni} + ^{100}\text{Mo}$ and $^{36}\text{S} + ^{48}\text{Ca}$ reactions have been re-examined. The obtained results demonstrate a remarkable agreement with the available experimental data. It is found that experimental data can be reproduced with the use of the Woods-Saxon potential, without introducing the repulsive cores. It appears that the fusion cross sections at deep sub-barrier energies are sensitive to the potential pocket profile.

DOI: [10.1103/PhysRevC.101.014618](https://doi.org/10.1103/PhysRevC.101.014618)**I. INTRODUCTION**

Nuclear fusion phenomena have attracted considerable theoretical and experimental attention over several decades [1–7]. Although basic notions of this phenomenon are relatively well understood, there are still many hidden details that require clarification. This is especially important, for example, in light of synthesis of superheavy nuclei and eval-

uation of boundaries of the nuclear drip line. According to general wisdom, the latest problems are closely related to intimate knowledge of various stages of astrophysical nucleogenesis, from the Big Bang to creation of life on our Earth.

Thanks to significant improvement of experimental sensitivity in view of the remarkable development of semiconductor detectors and computational capability, it becomes

possible to systematically investigate stable and exotic nuclei at energies well below the Coulomb barrier. In particular, the fusion experimental data at deep sub-barrier energy have been measured down to 10^{-5} mb [8]. Evidently, already available experimental data require reliable qualitative and quantitative interpretation. For example, the threshold anomaly problem [9–11], diffuseness parameter anomaly problem [12,13], deep sub-barrier fusion hindrance and its associated impact on stellar evolution [14–16], subbarrier positive Q -value fusion enhancement [17–22], and above barrier fusion suppression phenomena, to name just a few, are challenges to the theory. It is at once apparent that the degree of accuracy of theoretical calculations may lead to different conclusions on the same phenomenon [8,23–29].

The cross sections of near-barrier and especially sub-barrier fusion of nuclei can be described within the coupled-channels models that are based on various approximations (e.g., Refs. [30–33]). In particular, the approach of directly constructing a numerical solution to the set of coupled Schrödinger equations (see for details Refs. [6,34]) provides a convenient ground from which to calculate the fusion cross sections. Note that colliding nuclei may develop large dynamical deformations. Consequently, this problem requires the consideration of large number of coupled channels (see, e.g., for discussion Refs. [35–37]). As a result, one needs to preserve the numerical accuracy of the calculations, and this requires carefully treating boundary conditions.

There are generally two approaches to construct the fusion cross sections based on the solving of the coupled-channels equations. The first one is to use the regular boundary condition and the complex potential [5]. The fusion is defined as the absorption of the incident flux due to the imaginary part of the potential. The fusion cross section can be predicted accurately by the explicit integration of the imaginary potential over the radial wave functions. The other approach adopts the incoming wave boundary condition (IWBC). It assumes that there is a strong absorption in the inner region such that the incoming flux never returns. In this case, it is enough to consider the real potential only [6]. Following the same theoretical ideas as in Ref. [6], we develop a new algorithm for solving a set of second-order differential equations. We consider the boundary conditions with a strict requirement of a complete absence of the reflected waves from the intrinsic region behind the barrier. We calculate the matrix elements of the interaction between colliding nuclei explicitly.

The structure of the paper is the following. The theoretical framework is briefly discussed in Sec. II. Results of numerical calculations on two fusion reaction systems within our approach are presented in Sec. III. A summary of our work is given in Sec. IV.

II. THEORETICAL FRAMEWORK

In this section, for the sake of completeness, we review the basic notions of the coupled-channels model (see for details, e.g., Refs. [6,30,34–36,38]).

A. Basic equations

The fusion cross sections, decomposed over partial waves, have the following form:

$$\sigma_f(E) = \frac{\pi \hbar^2}{2\mu E} \sum_{l=0}^{\infty} (2l+1) P_l(E). \quad (1)$$

Here, E is the center-of-mass energy, $\mu = A_P A_T / (A_P + A_T)$ is the reduced mass, $A_{P(T)}$ is the mass of the projectile (target) nucleus, l is the orbital angular momentum, and $P_l(E)$ is the barrier penetration probability. Our task is to find the coefficients $P_l(E)$.

Consider a collision between two nuclei, taking into account the coupling of the relative motion between the centers of mass of the colliding nuclei, $\mathbf{r} = (r, \hat{\mathbf{r}})$ to a nuclear intrinsic motion ξ . The system Hamiltonian has the following form:

$$H(\mathbf{r}, \xi) = -\frac{\hbar^2}{2\mu} \nabla_{\mathbf{r}}^2 + V(\mathbf{r}) + H_0(\xi) + V_{\text{coup}}(\mathbf{r}, \xi), \quad (2)$$

where $H_0(\xi)$ describes the intrinsic structure, while the term $V_{\text{coup}}(\mathbf{r}, \xi)$ describes the coupling between the relative motion and the intrinsic structure. Note that the intrinsic degree of freedom ξ may have a finite spin I . In this case, for a fixed total angular momentum J and its z component M of the system, the channel wave function can be chosen in the following form:

$$\langle \hat{\mathbf{r}} \xi | (\alpha I) J M \rangle = \sum_{m_l, m_i} \langle l m_l I m_i | J M \rangle Y_{l m_l}(\hat{\mathbf{r}}) \varphi_{\alpha I m_i}(\xi), \quad (3)$$

where $Y_{l m_l}(\hat{\mathbf{r}})$ is the spherical harmonics. The wave functions of the internal motion $\varphi_{\alpha I m_i}(\xi)$ is subject to the equation

$$H_0(\xi) \varphi_{\alpha I m_i}(\xi) = \epsilon_{\alpha I} \varphi_{\alpha I m_i}(\xi), \quad (4)$$

where α stands for quantum numbers associated with the intrinsic motion and $\epsilon_{\alpha I}$ is the corresponding eigenenergy. Expanding the total wave function with the channel wave functions as

$$\Psi_J(\mathbf{r}, \xi) = \sum_{\alpha, l, I} \frac{u_{\alpha I}^J(r)}{r} \langle \hat{\mathbf{r}} \xi | (\alpha I) J M \rangle, \quad (5)$$

one obtains the coupled-channels equations for $u_{\alpha I}^J(r)$

$$\left[-\frac{\hbar^2}{2\mu} \frac{d^2}{dr^2} + \frac{l(l+1)\hbar^2}{2\mu r^2} + V(r) - E + \epsilon_{\alpha, I} \right] u_{\alpha I}^J(r) + \sum_{\alpha', l', I'} V_{\alpha I, \alpha' l' I'}^J(r) u_{\alpha' l' I'}^J(r) = 0, \quad (6)$$

where the coupling matrix elements $V_{\alpha, l, I, \alpha', l', I'}^J(r)$ are given as

$$V_{\alpha, l, I, \alpha', l', I'}^J(r) = \langle (\alpha I) J M | V_{\text{coup}}(\vec{r}, \xi) | (\alpha' l' I') J M \rangle. \quad (7)$$

In solving the quantum problem in question, we employ the so-called isocentrifugal approximation (see details in Ref. [6]). In this approximation, the angular momentum of the relative motion in each channel is replaced by the total angular momentum. In this case, one ignores the change of the orbital angular momentum due to intrinsic excitations. Such approximation allows us to reduce several-fold the dimensionality of the set of differential equations that should be solved.

B. Vibrational coupling

To demonstrate all pros and cons of our approach we analyze couplings of the relative motion to surface vibrations of a target nucleus only, comparing our results with those well known from literature. Hereafter, for the sake of convenience, we consider the potential between the projectile and the target as a function of the relative distance r between them:

$$V(r) = V_N(r) + V_C(r). \quad (8)$$

The potential contains the Coulomb term $V_C = Z_P Z_T e^2 / r$ and a phenomenological nuclear potential $V_N(r)$, that is chosen in the Woods-Saxon form:

$$V_N(r) = -\frac{V_0}{1 + \exp[(r - R_0)/a_0]}. \quad (9)$$

Here, the parameters V_0 , R_0 , a_0 are the potential depth, potential radius, and diffuseness, respectively.

The nuclear coupling term of the Hamiltonian (2) can be generated by changing the target radius in the potential to a dynamical operator $R_0 + \hat{O}$ [39]. The surface operator \hat{O} is defined as

$$\hat{O} = \frac{\beta_\lambda}{\sqrt{4\pi}} r_{\text{coup}} A_T^{1/3} (a_{\lambda 0}^\dagger + a_{\lambda 0}), \quad (10)$$

where $a_{\lambda 0}^\dagger$ ($a_{\lambda 0}$) is the creation (annihilation) operator of the vibrational mode of the multipolarity λ . In this representation, the matrix element of the operator \hat{O} has the following form:

$$\hat{O}_{nm} = \frac{\beta_\lambda}{\sqrt{4\pi}} r_{\text{coup}} A_T^{1/3} (\sqrt{m} \delta_{n,m-1} + \sqrt{n} \delta_{n,m+1}), \quad (11)$$

where the n -phonon state of the multipolarity λ is defined as

$$|n\rangle = \frac{1}{\sqrt{n!}} (a_{\lambda 0}^\dagger)^n |0\rangle.$$

The deformation parameter β_λ , that defines the amplitude of the zero-point motion, can be determined from the

experimental transition probability

$$\beta_\lambda = \frac{4\pi}{3Z_T R_T^\lambda} \sqrt{\frac{B(E\lambda) \uparrow}{e^2}}, \quad (12)$$

where R_T^λ is the radius of the spherical nucleus. In our consideration, the variable r_{coup} is a free parameter, being slightly varied around the mean value 1.2 fm. The nuclear coupling matrix elements are then determined as

$$V_{nm}^{(N)}(r) = \langle n | V_N(r, \hat{O}) | m \rangle - V_N^{(0)}(r) \delta_{n,m}, \quad (13)$$

where $V_N(r, \hat{O}) \iff V_N(r, R_0 + \hat{O})$, $V_N^{(0)}(r) \equiv V_N(r)$ (see Eqs. (3.52)–(3.59) in Ref. [6] for more details). The latter term in Eq. (13) is introduced to counteract the coupling interaction in the entrance channel [40].

Thus, in the isocentrifugal approximation, the coupled-channels Schrödinger equation has the following form:

$$\left[-\frac{\hbar^2}{2\mu} \frac{d^2}{dr^2} + \frac{l(l+1)\hbar^2}{2\mu r^2} + V_N^{(0)}(r) + \frac{Z_P Z_T e^2}{r} + \epsilon_n - E \right] \psi_{nm_o} + \sum_{n'=1}^N V_{nn'}(r) \psi_{n'n_o}(r) = 0. \quad (14)$$

In the above equation, ϵ_n is the excitation energy of the n th channel or threshold energy, $n = 1, \dots, N$, that is defined by Eq. (4). The number n_o is a number of the open entrance channel with a positive relative energy $E_{n_o} = E - \epsilon_{n_o} > 0$, $n_o = 1, \dots, N_o \leq N$, and the wave functions $\{\psi_{nm_o}(r)\}_{n=1}^N$ are components of a desirable matrix solution. The coupling matrix elements (7) are transformed to the matrix element $V_{nm}(r)$ that consists of the Coulomb and the nuclear potentials $V_N^{(0)}(r)$ in each entrance channel.

The solution of Eq. (14) is obtained under the IWBC. Namely, it is assumed that there is a strong absorption inside the potential pocket. The asymptotic boundary conditions of such type are determined conventionally for components of matrix solutions $\{\psi_{nm_o}(r)\}_{n=1}^N$ in the open entrance channels n_o with a positive relative energy E_{n_o} by the following relations:

$$\psi_{m_o}^{as}(r) = \begin{cases} \exp(-ik_n(r_{\min})r) T_{m_o}, & r \leq r_{\min}, \quad k_n(r_{\min}) > 0, \\ H_l^-(k_n r) \delta_{n,n_o} - H_l^+(k_n r) R_{m_o}, & r > r_{\max}. \end{cases} \quad (15)$$

The functions $H_l^\pm(k_n r) = \pm i F_l(\eta_n, k_n r) + G_l(\eta_n, k_n r)$ are the outgoing and the incoming Coulomb partial wave functions, respectively. They are determined by means of the regular $F_l(\eta_n, k_n r)$ and the irregular $G_l(\eta_n, k_n r)$ Coulomb partial wave functions [41]. Here, $k_n(r)$ is the local wave number for the n th channel

$$k_n(r) = \sqrt{\frac{2\mu}{\hbar^2} \left[E - \epsilon_n - \frac{l(l+1)\hbar^2}{2\mu r^2} - V_N^{(0)}(r) - \frac{Z_P Z_T e^2}{r} - V_{nm}(r) \right]} \quad (16)$$

that depends on the excitation energy ϵ_n of the n th channel. The asymptotic behaviors of the functions $H_l^\pm(k_n r)$ are defined as

$$H_l^\pm(k_n r) \rightarrow \exp \left[\pm i \left(k_n r - \eta_n \ln(2k_n r) + \sigma_{ln} - \frac{l\pi}{2} \right) \right], \quad (17)$$

where $\eta_n = k_n Z_T Z_P e^2 / (2E_n)$ is the Sommerfeld parameter; $\sigma_{ln} = \arg \Gamma(l+1 + i\eta_n)$ is the Coulomb phase shift in open channels at $k_n = \sqrt{2\mu(E - \epsilon_n)/\hbar^2} > 0$.

On the other hand, for the components of $\psi_{m_o}^{as}(r)$ with elements $n = N_o + 1, \dots, N$, where n is restricted by the condition $E_n = E - \epsilon_n \leq 0$, we have

$$\psi_{m_o}^{as}(r) = \begin{cases} \exp(|k_n(r_{\min})|r), & r \leq r_{\min}, \\ 0, & r \geq r_{\max}. \end{cases} \quad (18)$$

The conventional partial fusion probability $P_l(E)$ for the incident channel n_o is determined by summation over all open

channels of intrinsic states at $k_n(r_{\min}) > 0$ for $n = 1, \dots, N_o$:

$$P_l(E) \equiv T_{n_o n_o}^{(l)}(E) = \sum_{n=1}^{N_o} \frac{k_n(r_{\min})}{k_{n_o}} |T_{nn_o}|^2, \quad (19)$$

where the incident wave number $k_{n_o} = \sqrt{2\mu(E - \epsilon_{n_o})/\hbar^2}$. Finally, the total fusion cross section is expressed as a sum over partial waves at the center of mass energy E , which is

$$\sigma_f(E) = \sum_{l=0}^L \sigma_f^{(l)}(E) = \frac{\pi}{k_{n_o}^2} \sum_{l=0}^L (2l+1)P_l(E). \quad (20)$$

C. Boundary conditions

Prior to proceeding to the numerical analysis, a few comments are in order. In Eq. (15), r_{\max} is set as a large enough distance where the interaction is weak, and the off-diagonal elements of the coupled matrix tend to be zero. The minimal point r_{\min} is taken as the minimum of the potential pocket. The plane wave boundary condition at the left boundary r_{\min} involves only the diagonal part of the coupling matrix element $k_n(r_{\min})$ from Eq. (16). This requires that the off-diagonal matrix elements tend to be zero. However, at r_{\min} , the distance between two nuclei is so short that the off-diagonal matrix elements are usually not zero. As addressed in Ref. [35], there can be sudden noncontinuous changes in the left boundary, and this will cause the distortion for the total wave function in the barrier region. To resolve this problem, we further develop the approach proposed in a series of papers [37,42–48].

First, it is reasonable to assume that at r_{\max} the contribution of closed channels is negligible small. Consequently, we can use the conventional Dirichlet boundary condition at r_{\max} for components of matrix solutions $\psi_{nn_o}(r_{\max}) = 0$ of Eq. (14) for $[n = 1, \dots, N; n_o = N_o + 1, \dots, N]$ in closed channels [see also Eq. (18)]. Second, at the left boundary we adopt the linear transformation method [37]. The essence of this method is the following:

Let consider the matrix \mathbf{W} to be a symmetric matrix of our problem [see Eq. (14)] of the dimension $N \times N$

$$\begin{aligned} W_{nm} &= W_{mn} \\ &= \frac{2\mu}{\hbar^2} \left[\left(\frac{l(l+1)\hbar^2}{2\mu r^2} + V_N^{(0)}(r) \right. \right. \\ &\quad \left. \left. + \frac{Z_P Z_T e^2}{r} + \epsilon_n \right) \delta_{nm} + V_{nm}(r) \right], \quad (21) \end{aligned}$$

and the constant matrix in the vicinity of the left boundary point $r = r_{\min}$. In the above equation, $V_{nm}(r) = V_{nm}^{(N)}(r) + V_{nm}^{(C)}(r)$, where $V_{nm}^{(N)}(r)$ is obtained by Eq. (13) and $V_{nm}^{(C)}(r)$ is the Coulomb coupling matrix elements (see Eq. (28) in

Ref. [34] for more details). Here, the matrices \mathbf{A} and $\tilde{\mathbf{W}}$ are the matrix of eigenvectors and the diagonal matrix of eigenvalues of the eigenvalue problem, respectively. Namely, we have

$$\begin{aligned} \mathbf{W}\mathbf{A} &= \mathbf{A}\tilde{\mathbf{W}}, \quad \{\tilde{\mathbf{W}}\}_{nm} = \delta_{nm}\tilde{W}_{mm}, \\ \tilde{W}_{11} &\leq \tilde{W}_{22} \leq \dots \leq \tilde{W}_{NN}. \quad (22) \end{aligned}$$

In this case, the linear independent matrix solution $\{\phi_{nm}(r)\}_{n,m=1}^N$ of Eq. (14) can be written in the form

$$\phi_{nm}(r) = A_{nm}y_m(r), \quad (23)$$

where functions $y_m(r)$ are solutions of the uncoupled equations

$$y_m''(r) + K_m^2 y_m(r) = 0, \quad K_m^2 = \frac{2\mu}{\hbar^2} E - \tilde{W}_{mm}. \quad (24)$$

In open channels at $K_m^2 > 0$, $m = 1, \dots, M_o \leq N$, the solutions $y_m(r)$ have the form

$$y_m(r) = \frac{\exp(-iK_m r)}{\sqrt{K_m}}, \quad (25)$$

while in closed channels at $K_m^2 \leq 0$, $m = M_o + 1, \dots, N$,

$$y_m(r) = \frac{\exp(|K_m| r)}{\sqrt{|K_m|}}. \quad (26)$$

In this case, $\psi_{nn_o}(r)$ is expressed by the linear combinations of the linear independent solutions $\phi_{nm}(r)$

$$\begin{aligned} \psi_{nn_o}(r) &= \sum_{m=1}^{M_o} \phi_{nm}(r) \hat{T}_{mn_o} \equiv \sum_{m=1}^{M_o} A_{nm} y_m(r) \hat{T}_{mn_o}, \\ r &= r_{\min}. \quad (27) \end{aligned}$$

In this way, the off-diagonal matrix elements have been considered in our calculations (see Sec. III). We consider the following boundary conditions in two endpoints. At $r = r_{\min}$, we have in terms of corresponding solutions A_{nm} $n, m = 1, \dots, N$ at $K_m \geq 0$, $m = 1, \dots, M_o \leq N$ for open exit channels and pure imaginary $K_m < 0$, $m = M_o + 1, \dots, N$ for closed exit channels

$$\begin{aligned} \psi_{nn_o}^{as}(r) &= \sum_{m=1}^{M_o} A_{nm} \frac{\exp(-iK_m r)}{\sqrt{K_m}} \hat{T}_{mn_o} \\ &\quad + \sum_{m=M_o+1}^N A_{nm} \frac{\exp(|K_m| r)}{\sqrt{|K_m|}} \hat{T}_{mn_o}^c, \quad r = r_{\min}. \quad (28) \end{aligned}$$

At $r = r_{\max}$, the asymptotic solutions are given in the terms of normalized Coulomb functions $\hat{H}_l^\pm(k_n r) = H_l^\pm(k_n r)/\sqrt{k_n}$, $k_n \geq 0$, $n = 1, \dots, N_o \leq N$, and for components of $\psi_{nn_o}^{as}(r_{\max}) = o(1)$ with elements $n = N_o + 1, \dots, N$ for closed channels,

$$\psi_{nn_o}^{as}(r) = \begin{cases} \hat{H}_l^-(k_n r) \delta_{n,n_o} - \hat{H}_l^+(k_n r) \hat{R}_{nn_o}, & r = r_{\max}, \\ 2|k_n|^{1/2} r^{l+1} \exp(-|k_n| r) U(l+1+\eta_n, 2l+2, 2|k_n| r), & r = r_{\max}. \end{cases} \quad (29)$$

Here $U(l+1+\eta_n, 2l+2, 2|k_n| r)$ is Whittaker function [41], \hat{T}_{nn_o} and \hat{R}_{nn_o} are desirable partial transmission and reflection

amplitudes, and they are at $n_o = 1$ desirable—from a ground state $|i_o\rangle = |n_o - 1\rangle = |0\rangle$ of the intrinsic motion before the

collision, $\hat{T}_{mn_o}^c$ are transmission amplitudes in closed channels $m = M_o + 1, \dots, N$.

The third type or Robin boundary conditions for solutions $\psi_{mn_o}(r)$ of Eq. (14) follow from their asymptotic expansion $\psi_{mn_o}(r)$

$$\left(\frac{d\psi_{mn_o}(r)}{dr} - \sum_{n'=1}^N G_{nn'}(r)\psi_{n'n_o}(r) \right)_{r=r_{\min}, r_{\max}} = 0, \quad (30)$$

where $G_{nn'}(r)$ are solutions of algebraic problem

$$\left(\frac{d\psi_{mn_o}^{as}(r)}{dr} - \sum_{n'=1}^N G_{nn'}(r)\psi_{n'n_o}^{as}(r) \right)_{r=r_{\min}, r_{\max}} = 0. \quad (31)$$

In this case, at fixed orbital momentum l the partial fusion probability

$$P_l(E) \equiv T_{n_o n_o}^{(l)}(E) \quad (32)$$

is given by summation over all possible intrinsic states:

$$\begin{aligned} T_{n_o n_o}^{(l)}(E) &= \sum_{m=1}^{M_o} |\hat{T}_{mn_o}^{(l)}|^2, \\ R_{n_o n_o}^{(l)}(E) &= \sum_{n=1}^{N_o} |\hat{R}_{nn_o}^{(l)}|^2, \\ T_{n_o n_o}^{(l)}(E) &= 1 - R_{n_o n_o}^{(l)}(E), \end{aligned} \quad (33)$$

that we used $P_{n_o n_o}^{(l)}(E) \equiv T_{n_o n_o}^{(l)}(E)$ in the conventional formula for total fusion cross section (20). The above discussed ideas have been transformed to the improved version of the program KANTBP used in our calculations. This program is based on the finite-element method and will be presented in the forthcoming paper.

It is noteworthy that the condition $T_{n_o n_o}^{(l)}(E) + R_{n_o n_o}^{(l)}(E) - 1 = 0$ fulfills in below calculations with ten significant digits. This means that the calculated scattering \mathbf{S} matrix is symmetric and unitary with an accuracy of the same order [45]. The reader can find details of the preceding version of the program KANTBP in Refs. [43,44].

III. RESULTS AND DISCUSSION

In order to validate our approach, we calculate the tunneling probability and fusion cross sections for $^{16}\text{O} + ^{144}\text{Sm}$ by means of KANTBP. We consider one incident channel and one coupled channel. Only the low-lying collective 3^- vibrational state of ^{144}Sm with the excitation energy 1.81 MeV is taken into account. Our results demonstrates a remarkable agreement with those obtained by the modified Numerov (MNumerov) method (employed in CCFULL) (see Fig. 1). The potential parameters, used in this calculation, produce a very steep potential pocket, with barrier height at 61.25 MeV and pocket minimum at 8.94 MeV. The lowest incident energy is 55 MeV, which is far higher than the potential minimum.

For most fusion reaction systems, the results predicted with the use of KANTBP and CCFULL are almost identical when there are few coupled channels at near-barrier incident energy. For example, we observe such the agreement as well

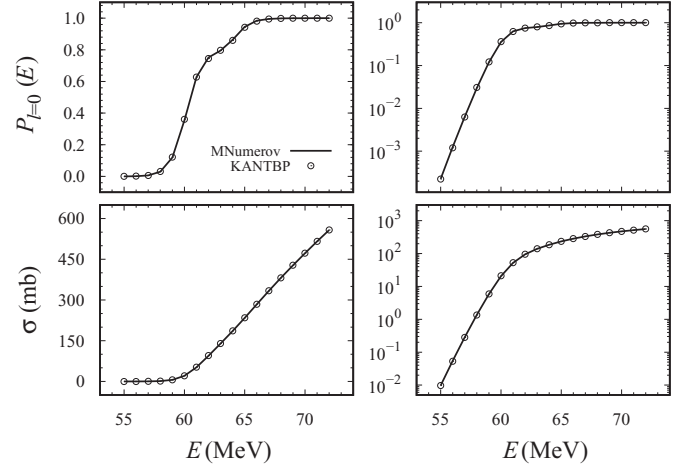


FIG. 1. The tunneling probability and fusion cross sections for $^{16}\text{O} + ^{144}\text{Sm}$ at linearization and logarithmic scale. The results obtained with the use of CCFULL are connected by solid line; also labeled as MNumerov. The results obtained by means of KANTBP are denoted by open circles. The parameters used in both calculations are taken from Ref. [34].

for ^{32}S , $^{40}\text{Ca} + ^{90,94,96}\text{Zr}$ reactions. The method introduced in these cases does not gain so much. However, when the number of coupled channels is increased considerably, the differences become more evident. Besides that, at the deep sub-barrier energy region, when the incident energy is close to the potential minimum, the fusion cross sections are very low and quite sensitive to the theoretical scheme.

In the following, we will consider $^{64}\text{Ni} + ^{100}\text{Mo}$ and $^{36}\text{S} + ^{48}\text{Ca}$ reactions and compare the results obtained by means of our approach and with the use of CCFULL. These two reactions have been both measured down to the deep sub-barrier energy region. Because of the instability of the modified Numerov method used in the CCFULL calculations, the shapes of the cross section lines can be different by connecting fusion cross section points at different incident energies. In order to avoid the shape uncertainty, we perform calculations at available experimental data except where there is no experimental point at the lowest energy.

In Table I, the adopted structure properties including excitation energies and deformation parameters for the nuclei used in this study are listed [49,50]. The low-lying collective 2^+ and 3^- vibrational states are considered. The radius parameter r_{coup} in the coupling interactions of Eq. (10) is assumed as

TABLE I. Adopted excitation energies E_x , spins and parities λ^π , $\pi = (-1)^\lambda$, and deformation parameters β_λ of the low-lying collective excited states for the indicated nuclei. The units of the excitation energies are in MeV.

Nucleus	^{36}S	^{48}Ca	^{64}Ni	^{100}Mo
E_{2^+} [49]	3.291	3.832	1.346	0.536
β_2 [49]	0.168	0.106	0.179	0.231
E_{3^-} [50]	4.193	4.507	3.560	1.908
β_3 [50]	0.376	0.230	0.201	0.218

1.2 fm for both target and projectile in all the following calculations. The numbers of target 3^- phonon, target 2^+ phonon, and projectile 2^+ phonon are denoted as $N_{T_{3^-}}$, $N_{T_{2^+}}$, and $N_{P_{2^+}}$ respectively. The total coupled-channels number will be $N_{\text{coup}} = (N_{T_{3^-}} + 1)(N_{T_{2^+}} + 1)(N_{P_{2^+}} + 1) - 1$ when all mutual excitations are included. It means that number of coupled equations in Eq. (14) is $N = N_{\text{coup}} + 1$.

The Woods-Saxon potential parameters in Eq. (9) derived from Akyüz-Winther (AW) parametrization [53,54] are used in the next step of calculations. This potential is obtained by means of fitting large-scale experimental scattering data, and has been successfully used in describing different kinds of reactions. It is written as

$$V_{\mathcal{N}}^{(0)}(r) = -\frac{V_0}{1 + \exp[(r - R_P - R_T)/a_0]}, \quad (34)$$

where

$$V_0 = (16\pi\gamma a_0 \bar{R}) \text{ MeV},$$

$$\frac{1}{a_0} = 1.17[1 + 0.53(A_P^{-1/3} + A_T^{-1/3})] \text{ fm}^{-1},$$

$$\bar{R} = \frac{R_P R_T}{R_P + R_T},$$

$$R_i = (1.2A_i^{1/3} - 0.09) \text{ fm}, \quad i = P, T,$$

$$\gamma = 0.95 \left(1 - 1.8 \frac{(N_P - Z_P)(N_T - Z_T)}{A_P A_T} \right) \text{ MeV fm}^{-2}.$$

Here $N_{P(T)}$ is the neutron number of the projectile (target) nucleus. The fusion reaction $^{64}\text{Ni} + ^{100}\text{Mo}$ had been measured at the superconducting linear accelerator ATLAS of Argonne National Laboratory [51]. The coupled-channel calculations that adopted different vibrational properties and nuclear radii were unable to reproduce the experimental fusion cross sections at deep sub-barrier energies. Hence, it was concluded that this system exhibits a hindrance for fusion. In Ref. [29], the CC calculations with the M3Y+ repulsion potential were used to describe the experimental data. Later, it was reported in Ref. [52] that the coupled-channel calculations, including a much deeper well potential than the standard AW potential [53,54] and a small radius parameter, will fit the experimental data well. In Ref. [19], the authors reproduced the major part of the experimental data by means of the coupled-channel method [35,36]. In these calculations, the standard AW potential, and the phonon numbers $N_{T_{3^-}} = 2$, $N_{T_{2^+}} = 2$, and $N_{P_{2^+}} = 2$, $N_{P_{3^-}} = 1$ are adopted. However, the slopes are deeper, and the predicted cross sections are generally smaller than the experimental data for energies $E \leq 125$ MeV.

The results obtained by means of KANTBP reproduce the experimental data well (see Fig. 2), without any special settings on the potential. In these calculations, 26 coupled channels are considered in the calculations, taking into account the number of excited states of the target: $N_{T_{3^-}} = 2$, $N_{T_{2^+}} = 2$, and $N_{P_{2^+}} = 2$. The detailed channels are listed in Table II. The standard AW potential [53,54] is adopted. Note that the above-barrier and below-barrier fusion cross sections are described within the experimental errors quite well. In contrast, the CCFULL results fluctuate when the incident energy $E < 130$ MeV, and far

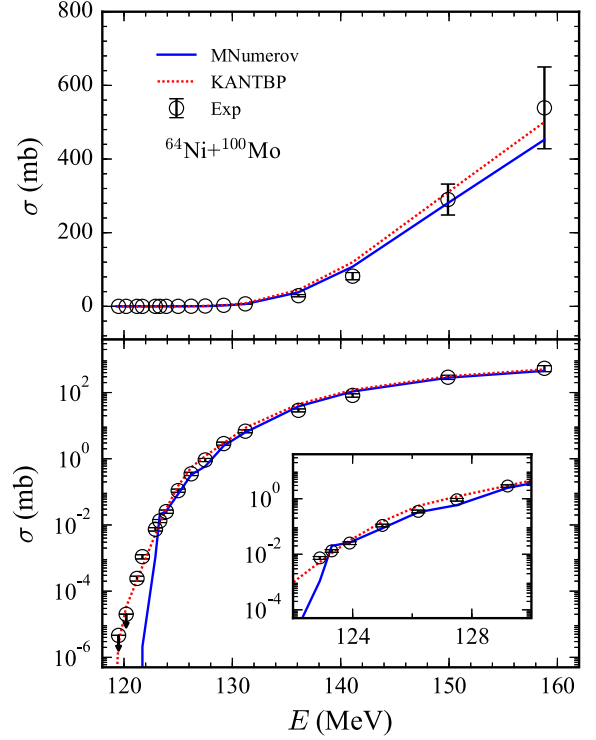


FIG. 2. Fusion cross sections for $^{64}\text{Ni} + ^{100}\text{Mo}$. The experimental data (open circles) are from Ref. [51]. The fusion cross sections at the lowest two energies are the upper limits, which are indicated with arrows. The comparison of results obtained by means of CCFULL (solid line, also labeled as MNumerov), and by means of KANTBP (dotted line). All calculations are performed at the experimental incident energies except where there is no experimental point at the lowest energy. The insert is an enlargement of the sub-barrier fusion cross sections.

from the experimental data at deep sub-barrier energy region. We have also tested the predictions obtained with the use of CCFULL + stabilization method (see Ref. [55]), which is the same as the solid lines shown in Fig. 2. It should also be noted that when the Coulomb potential is changed to a spherical one, the instability at the low-energy tail will be shifted downward according to the radius parameters because the spherical Coulomb potential produces a deeper potential pocket and lower threshold energy.

In these calculations for $l = 0$, the largest diagonal matrix elements of the matrix $\hbar^2 \mathbf{W}/2\mu$ in Eq. (21) is 130.98 MeV. At the incident energy $E < 130.98$ MeV, the results of

TABLE II. The list of the 26 coupled channel for $N_{T_{3^-}} = 2$, $N_{T_{2^+}} = 2$, $N_{P_{2^+}} = 2$ in the form of $|T_{3^-} T_{2^+} P_{2^+}\rangle$ excluding the ground-state channel $|000\rangle$.

Configuration	Channels
Projectile	$ 001\rangle, 002\rangle$
Target	$ 100\rangle, 200\rangle, 010\rangle, 020\rangle, 110\rangle, 120\rangle, 210\rangle, 220\rangle$
Mutual	$ 101\rangle, 201\rangle, 011\rangle, 021\rangle, 111\rangle, 121\rangle, 211\rangle, 221\rangle$ $ 102\rangle, 202\rangle, 012\rangle, 022\rangle, 112\rangle, 122\rangle, 212\rangle, 222\rangle$

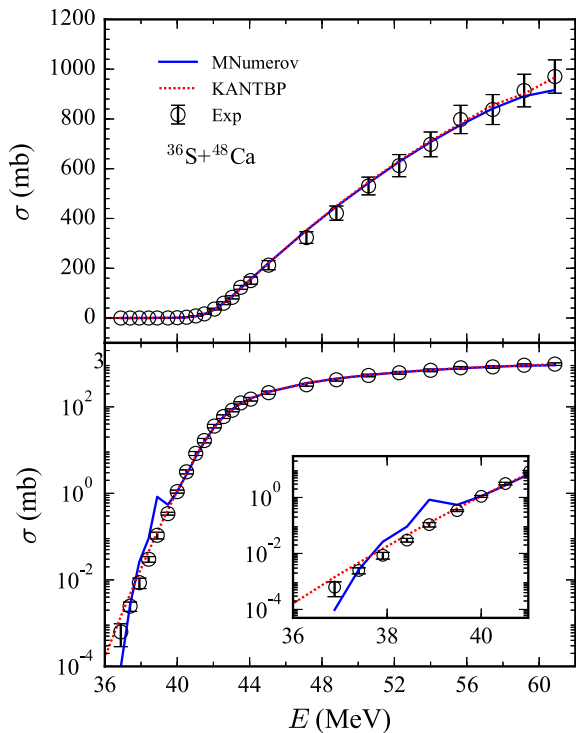


FIG. 3. Fusion cross sections for $^{36}\text{S} + ^{48}\text{Ca}$. The notations are the same as in Fig. 2. The experimental data (open circles) on fusion cross sections are from Ref. [15]. The calculations are performed at the experimental incident energies except where there is no experimental point at the lowest energy. The insert is an enlargement of the sub-barrier fusion cross sections.

calculations should be influenced heavily by the nondiagonal elements. However, this effect is not considered in CCFULL. This observation explains the reason why the calculation start to fluctuate below this energy. The linear transform procedure introduced in this study changes not only the threshold energy by diagonalization, but also the number of open channels and closed channels. As a result, the final transmission matrix, and the cross sections will be affected. By considering the nondiagonal element in the \mathbf{W} matrix at the left boundary, the calculation by KANTBP produces more stable results below about 130 MeV.

The fusion reaction $^{36}\text{S} + ^{48}\text{Ca}$ was performed at the accelerator of the Laboratori Nazionali di Legnaro of INFN [15]. A deep sub-barrier fusion hindrance feature of this reaction system had been reported. A large diffuseness parameter of $a = 0.95$ fm was used to reproduce the data above and below the barrier, which may actually mimic the presence of the deep inelastic reactions [12]. In Refs. [56] and [16], the double-folding ion-ion potential from different parametrization plus a repulsive contact term was adopted to describe the experimental cross sections. A weak and short-ranged imaginary potential is also used in Ref. [16] in order to remove some unwanted fluctuations in the theoretical calculation.

The results of fusion cross section calculations for $^{36}\text{S} + ^{48}\text{Ca}$ are presented in Fig. 3. In KANTBP, we use the same standard AW nuclear potential and the 26 coupled channels are considered, since there are the following excitations:

TABLE III. Woods-Saxon potential parameters V_0 (MeV), a_0 (fm), R_0 (fm), fitted at different combinations of the vibration phonon numbers $N_{T_{3-}}$, $N_{P_{2+}}$, $N_{T_{2+}}$ for fusion excitation function of the $^{36}\text{S} + ^{48}\text{Ca}$ reaction. The standard AW-type potential parameters are listed in the second column for comparison.

	AW	Ch-0	Ch-1	Ch-17
$N_{T_{3-}}$		0	0	1
$N_{T_{2+}}$		0	0	2
$N_{P_{2+}}$		0	1	2
V_0 (MeV)	61.338	72.325	61.355	55.911
a_0 (fm)	0.654	0.636	0.652	0.676
R_0 (fm)	8.143	8.272	8.298	8.167
V_B (MeV)	42.706	41.885	42.041	42.617
R_B (fm)	10.052	10.296	10.228	10.042
$\hbar\omega$ (MeV)	3.285	3.315	3.249	3.196

$N_{T_{3-}} = 2$, $N_{T_{2+}} = 2$, and $N_{P_{2+}} = 2$. The results demonstrate good agreement with the experimental data near the barrier energy region. At the deep sub-barrier energy region, KANTBP results are slightly higher than the experiments, which may indicate the fusion hindrance feature for this reaction system. In contrast, CCFULL results manifest large fluctuations at the deep sub-barrier energy region. The reason of this fluctuation is the same as for the above discussed case (see Fig. 2). Namely, at $l = 0$ the largest diagonal matrix elements of the matrix $\hbar^2\mathbf{W}/2\mu$ in Eq. (21) is 38.13 MeV. For the incident energy $E < 38.13$ MeV, there is a contribution of nondiagonal elements that are non-negligible. They are missing in CCFULL calculations.

Despite the many previous calculations mentioned above, it is of great interest to see whether the experimental fusion data can be explained by a simple Woods-Saxon-type potential model. In the following, we will try to find out that whether it is possible to describe well the experimental data by fitting the three parameters of the Woods-Saxon potential. The stability of the numerical method KANTBP is advantageous for fitting under some extreme parameters. The three Woods-Saxon potential parameters V_0 , R_0 , and a_0 are fitted to reproduce the fusion cross sections of $^{36}\text{S} + ^{48}\text{Ca}$ under different kinds of collective vibrations. The fitting parameters are shown in Table III, and the corresponding calculations under different conditions are given in Fig. 4. Different lines in the figure are denoted by the number of the coupled channels used. Three cases are examined here: 0, 1, and 17 coupled channels. The results of fitting demonstrate a good agreement with the experimental data for all three cases, in the above barrier energy region and below barrier energy region. The calculations under different collective vibrations are also almost overlapped.

The fitted Coulomb barrier properties, including barrier height V_B , barrier radius R_B , and the barrier curvature $\hbar\omega$, are listed in Table III. It can be seen that all fitted parameters in the last three columns are not very far from the stand AW parameters in the second column. It is not necessary to use very deep sub-barrier depths or very large diffuseness parameters to agree with the experimental data, like $V_0 = 165$ MeV and $a_0 = 0.95$ fm in previous study [15].

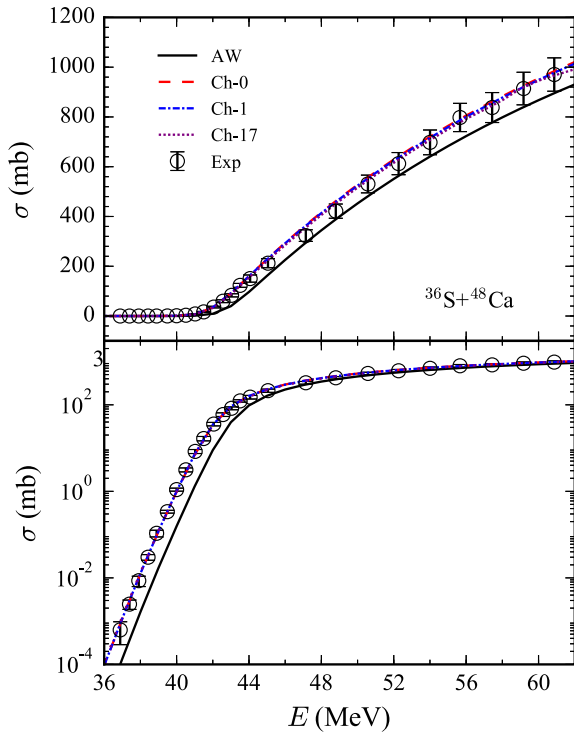


FIG. 4. Fusion cross sections for $^{36}\text{S} + ^{48}\text{Ca}$. The experimental data (open circles) on fusion cross sections are from Ref. [15]. The calculations with standard AW potential and 0 coupled channels are denoted by the solid lines. The fitted calculations performed with 0, 1, 17 coupled channels are represented by the dashed lines (Ch-0), dash-dotted lines (Ch-1), and dotted lines (Ch-17), respectively.

Corresponding fitted Woods-Saxon potentials are plotted in Fig. 5. The results obtained by the different methods show the different fitting the experimental data. However, the changing trends of the potential barrier can be seen from this figure. The fitted potentials reflect two tendencies in order to describe well the experimental fusion data theoretically, especially at the deep sub-barrier energy region. On the one hand, one can include fewer reaction channels and deeper potential inside the barrier pocket. On the other hand, one can make the

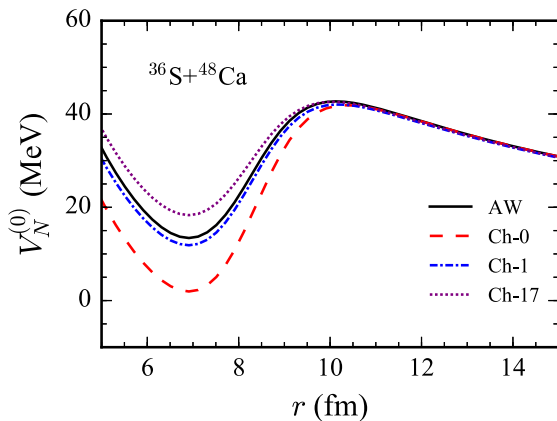


FIG. 5. The Woods-Saxon potentials calculated by different parameters listed in Table III. The notations are the same as in Fig. 4.

potential shallower and use more reaction channels. However, this reaction system can still be described well by using the simple Woods-Saxon potential. Comparing the lines AW and Ch-17, the Coulomb barriers are almost not changed, but the shapes of the potential well are quite different. This demonstrates that the deep sub-barrier fusion cross sections are very sensitive to the inner shape of the potential well, which have also been discussed in Ref. [25].

IV. SUMMARY

One of the standard methods to predict the fusion cross sections of light nuclei and capture cross section of the massive nuclei is to solve the set of coupled-channel differential equations with the use of the Numerov method. It is the heart of the full order coupling code CCFULL [34]. However, the CCFULL code, taking into account a large number of coupled channels, exhibits characteristics instabilities of the fusion excitation functions for some reactions. In the present paper, we developed a new algorithm for solving a set of second-order differential equations with the use of the finite-element method. To attack this problem, we further developed the approach proposed in series of papers [37,42–46,48]. Guided by our approach, we constructed the program KANTBP that was used for analysis of the fusion cross section of the $^{64}\text{Ni} + ^{100}\text{Mo}$ and $^{36}\text{S} + ^{48}\text{Ca}$ reactions. We demonstrated that our approach allows us to eliminate successfully the instabilities in the numerical solutions of the coupled-channels differential equations for these reactions.

In previous studies, special treatments of the potential, such as a large diffuseness parameter [15], a very deep potential plus a small radius parameter [52], or a repulsive core are needed to explain these experimental data related to the considered nuclei [16,29,56]. By means of our approach, we found that the fusion cross sections can be still described well with a simple Woods-Saxon-type potential. In particular, the fusion excitation function of the $^{64}\text{Ni} + ^{100}\text{Mo}$ reaction is remarkably well described with the use of the standard AW potential. On the other hand, the fusion excitation function of the $^{36}\text{S} + ^{48}\text{Ca}$ is described well by fitting of the Woods-Saxon potential parameters, without introducing the repulsive cores. It is demonstrated that the deep sub-barrier fusion cross sections are very sensitive to the potential pocket profile. The deep sub-barrier fusion cross sections can be used as a sensitive probe to explore the inner shape of the potential pockets.

ACKNOWLEDGMENTS

The work was supported in part by the Bogoliubov-Infeld and Hulubei-Meshcheryakov programs and by a grant from the Plenipotentiary Representative of the Government of the Republic of Kazakhstan in the framework of collaboration program JINR–RK. The publication has been prepared with the support of the “RUDN University Program 5-100.” The work of P.W.W., C.J.L., and H.M.J. is supported by the National Key R&D Program of China (Contract No. 2018YFA0404404), the National Natural Science Foundation of China (Grants No.11635015, No. 11805120, No. 11635003, No. 11805280, No. 11811530071, No. U1867212,

and No. U1732145), the project funded by China Postdoctoral Science Foundation (Grant No. 2017M621035), and the Continuous Basic Scientific Research Project (Grant No. WDJC-2019-13). A.K.N. thanks the Russian Foundation for

Basic Research for the partial support of the Project No. 17-52-45037. The present research benefited from computational resources of the HybriLIT heterogeneous platform of the JINR.

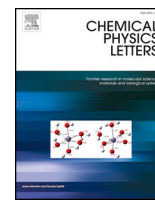
-
- [1] G. Montagnoli and A. M. Stefanini, *Eur. Phys. J. A* **53**, 169 (2017).
- [2] B. B. Back, H. Esbensen, C. L. Jiang, and K. E. Rehm, *Rev. Mod. Phys.* **86**, 317 (2014).
- [3] C. J. Lin, *Heavy-Ion Nuclear Reactions* (Harbin Engineering University Press, Harbin, 2015).
- [4] J. J. Kolata, V. Guimarães, and E. F. Aguilera, *Eur. Phys. J* **52**, 123 (2016).
- [5] L. F. Canto, P. R. S. Gomes, R. Donangelo, J. Lubian, and M. S. Hussein, *Phys. Rep.* **596**, 1 (2015).
- [6] K. Hagino and N. Takigawa, *Prog. Theor. Phys.* **128**, 1061 (2012).
- [7] N. Keeley, R. Raabe, N. Alamanos, and J. L. Sida, *Prog. Part. Nucl. Phys.* **59**, 579 (2007).
- [8] M. Dasgupta, D. J. Hinde, A. Diaz-Torres, B. Bouriquet, C. I. Low, G. J. Milburn, and J. O. Newton, *Phys. Rev. Lett.* **99**, 192701 (2007).
- [9] G. R. Satchler, *Phys. Rep.* **199**, 147 (1991).
- [10] C. J. Lin, J. C. Xu, H. Q. Zhang, Z. H. Liu, F. Yang, and L. X. Lu, *Phys. Rev. C* **63**, 064606 (2001).
- [11] L. Yang, C. J. Lin, H. M. Jia, D. X. Wang, N. R. Ma, L. J. Sun, F. Yang, X. X. Xu, Z. D. Wu, H. Q. Zhang, and Z. H. Liu, *Phys. Rev. Lett.* **119**, 042503 (2017).
- [12] J. O. Newton, R. D. Butt, M. Dasgupta, D. J. Hinde, I. I. Gontchar, C. R. Morton, and K. Hagino, *Phys. Rev. C* **70**, 024605 (2004).
- [13] J. O. Newton, R. D. Butt, M. Dasgupta, D. J. Hinde, I. Gontchar, C. R. Morton, and K. Hagino, *Phys. Lett. B* **586**, 219 (2004).
- [14] C. L. Jiang, K. E. Rehm, R. V. F. Janssens, H. Esbensen, I. Ahmad, B. B. Back, P. Collon, C. N. Davids, J. P. Greene, D. J. Henderson, G. Mukherjee, R. C. Pardo, M. Paul, T. O. Pennington, D. Seweryniak, S. Sinha, and Z. Zhou, *Phys. Rev. Lett.* **93**, 012701 (2004).
- [15] A. M. Stefanini, G. Montagnoli, R. Silvestri, S. Beghini, L. Corradi, S. Courtin, E. Fioretto, B. Guiot, F. Haas, D. Lebhertz, N. Mărginean, P. Mason, F. Scarlassara, R. N. Sagaidak, and S. Szilner, *Phys. Rev. C* **78**, 044607 (2008).
- [16] G. Montagnoli, A. M. Stefanini, H. Esbensen, C. L. Jiang, L. Corradi, S. Courtin, E. Fioretto, A. Goasduff, J. Grebosz, F. Haas, M. Mazzocco, C. Michelagnoli, T. Mijatovic, D. Montanari, C. Parascandolo, K. E. Rehm, F. Scarlassara, S. Szilner, X. D. Tang, and C. A. Ur, *Phys. Rev. C* **87**, 014611 (2013).
- [17] H. M. Jia, C. J. Lin, F. Yang, X. X. Xu, H. Q. Zhang, Z. H. Liu, L. Yang, S. T. Zhang, P. F. Bao, and L. J. Sun, *Phys. Rev. C* **86**, 044621 (2012).
- [18] H. M. Jia, C. J. Lin, F. Yang, X. X. Xu, H. Q. Zhang, Z. H. Liu, Z. D. Wu, L. Yang, N. R. Ma, P. F. Bao, and L. J. Sun, *Phys. Rev. C* **89**, 064605 (2014).
- [19] A. V. Karpov, V. A. Rachkov, and V. V. Samarin, *Phys. Rev. C* **92**, 064603 (2015).
- [20] H. M. Jia, C. J. Lin, L. Yang, X. X. Xu, N. R. Ma, L. J. Sun, F. Yang, Z. D. Wu, H. Q. Zhang, Z. H. Liu, and D. X. Wang, *Phys. Lett. B* **755**, 43 (2016).
- [21] P. W. Wen, Z. Q. Feng, C. Li, C. J. Lin, and F. S. Zhang, *Chinese. Phys. Lett.* **34**, 042501 (2017).
- [22] P. W. Wen, Z. Q. Feng, C. Li, C. J. Lin, and F. S. Zhang, *Chin. Phys. C* **41**, 064102 (2017).
- [23] C. J. Lin, *Phys. Rev. Lett.* **91**, 229201 (2003).
- [24] K. Hagino, N. Rowley, and M. Dasgupta, *Phys. Rev. C* **67**, 054603 (2003).
- [25] C. H. Dasso and G. Pollarolo, *Phys. Rev. C* **68**, 054604 (2003).
- [26] C. J. Lin, *Prog. Theor. Phys. Supp* **154**, 184 (2004).
- [27] Ş. Mişicu and H. Esbensen, *Phys. Rev. Lett.* **96**, 112701 (2006).
- [28] T. Ichikawa, K. Hagino, and A. Iwamoto, *Phys. Rev. Lett.* **103**, 202701 (2009).
- [29] Ş. Mişicu and H. Esbensen, *Phys. Rev. C* **75**, 034606 (2007).
- [30] C. H. Dasso and S. Landowne, *Comput. Phys. Commun.* **46**, 187 (1987).
- [31] J. Fernandez-Niello, C. H. Dasso, and S. Landowne, *Comput. Phys. Commun.* **54**, 409 (1989).
- [32] M. Dasgupta, A. Navin, Y. K. Agarwal, C. V. K. Baba, H. C. Jain, M. L. Jhingan, and A. Roy, *Nucl. Phys. A* **539**, 351 (1992).
- [33] J. O. Newton, C. R. Morton, M. Dasgupta, J. R. Leigh, J. C. Mein, D. J. Hinde, H. Timmers, and K. Hagino, *Phys. Rev. C* **64**, 064608 (2001).
- [34] K. Hagino, N. Rowley, and A. T. Kruppa, *Comput. Phys. Commun.* **123**, 143 (1999).
- [35] V. I. Zagrebaev, *Prog. Theor. Phys. Supp* **154**, 122 (2004).
- [36] V. V. Samarin and V. I. Zagrebaev, *Nucl. Phys. A* **734**, E9 (2004).
- [37] A. A. Gusev, O. Chuluunbaatar, S. I. Vinitzky, L. L. Hai, V. L. Derbov, and A. Gózdź, *Bull. Peoples' Friendship Univ. Russia. Ser. Math. Inf. Sci. Phys.* **3**, 38 (2016).
- [38] T. Tamura, *Rev. Mod. Phys.* **37**, 679 (1965).
- [39] A. Bohr and B. R. Mottelson, *Nuclear Structure* (Amsterdam, New York, 1974), Vol. II.
- [40] K. Hagino, N. Takigawa, M. Dasgupta, D. J. Hinde, and J. R. Leigh, *Phys. Rev. C* **55**, 276 (1997).
- [41] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
- [42] O. Chuluunbaatar, A. A. Gusev, A. G. Abrashkevich, A. Amaya-Tapia, M. S. Kaschiev, S. Y. Larsen, and S. I. Vinitzky, *Comput. Phys. Commun.* **177**, 649 (2007).
- [43] O. Chuluunbaatar, A. A. Gusev, V. P. Gerdt, V. A. Rostovtsev, S. I. Vinitzky, A. G. Abrashkevich, M. S. Kaschiev, and V. V. Serov, *Comput. Phys. Commun.* **178**, 301 (2008).
- [44] A. A. Gusev, O. Chuluunbaatar, S. I. Vinitzky, and A. G. Abrashkevich, *Comput. Phys. Commun.* **185**, 3341 (2014).
- [45] A. A. Gusev, O. Chuluunbaatar, S. I. Vinitzky, and A. G. Abrashkevich, *Math. Mod. Geom.* **3**, 22 (2015).

- [46] A. A. Gusev, O. Chuluunbaatar, S. I. Vinitzky, A. G. Abrashkevich, A. Amaya-Tapia, M. S. Kaschiev, S. Y. Larsen, V. P. Gerdt, V. A. Rostovtsev, and V. V. Serov, <https://www1.jinr.ru/programs/jinrlib/kantbp/indexe.html>.
- [47] A. A. Gusev, L. L. Hai, O. Chuluunbaatar, and S. I. Vinitzky, <http://wwwinfo.jinr.ru/programs/jinrlib/kantbp4m>.
- [48] A. A. Gusev, S. I. Vinitzky, O. Chuluunbaatar, V. P. Gerdt, and V. A. Rostovtsev, *Lect. Notes Comp. Sci.* **6885**, 175 (2011).
- [49] S. Raman, C. W. Nestor, and P. Tikkanen, *Atom. Data. Nucl. Data.* **78**, 1 (2001).
- [50] T. Kibédi and R. H. Spear, *Atom. Data. Nucl. Data.* **80**, 35 (2002).
- [51] C. L. Jiang, K. E. Rehm, H. Esbensen, R. V. F. Janssens, B. B. Back, C. N. Davids, J. P. Greene, D. J. Henderson, C. J. Lister, R. C. Pardo, T. Pennington, D. Peterson, D. Seweriyak, B. Shumard, S. Sinha, X. D. Tang, I. Tanihata, S. Zhu, P. Collon, S. Kurtz, and M. Paul, *Phys. Rev. C* **71**, 044613 (2005).
- [52] A. M. Stefanini, G. Montagnoli, F. Scarlassara, C. L. Jiang, H. Esbensen, E. Fioretto, L. Corradi, B. B. Back, C. M. Deibel, B. Di Giovine, J. P. Greene, H. D. Henderson, S. T. Marley, M. Notani, N. Patel, K. E. Rehm, D. Seweriyak, X. D. Tang, C. Ugalde, and S. Zhu, *Eur. Phys. J. A* **49**, 1 (2013).
- [53] R. A. Broglia, R. A. Ricci, C. H. Dasso, and S. I. D. Fisica, *Nuclear Structure and Heavy-Ion Collisions: Varenna on Lake Como, Villa Monastero, 9th–21st July 1979* (North-Holland, Amsterdam, 1981).
- [54] A. Winther, *Nucl. Phys. A* **594**, 203 (1995).
- [55] N. R. K. Hagino and A. Kruppa, <http://www.nucl.phys.tohoku.ac.jp/~hagino/ccfull.html>.
- [56] Ş. Mişicu and F. Carstoiu, *Phys. Rev. C* **83**, 054622 (2011).



Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Research paper

D_{3h} symmetry adapted correlated three center wave functions of the ground and the first five excited states of H_3^+

O. Chuluunbaatar^{a,b}, S. Obeid^c, B.B. Joulakian^{c,*}, A.A. Gusev^{a,d}, P.M. Krassovitskiy^{a,e}, L.A. Sevastianov^{a,f}

^a Joint Institute for Nuclear Research, Dubna, Moscow Region 141980, Russia

^b Institute of Mathematics and Digital Technologies, Mongolian Academy of Sciences, Ulaanbaatar 13330, Mongolia

^c Université de Lorraine, LPCT (UMR CNRS 7019), 1 bld Arago, bat. ICPM 57078, Metz Cedex 3, France

^d Dubna University, Dubna, Moscow Region 141980, Russia

^e Institute of Nuclear Physics, Almaty 050032, Kazakhstan

^f Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russia



HIGHLIGHTS

- Three-center basis for the D_{3h} symmetry of the H_3^+ studied here with electronic correlation.

ARTICLE INFO

Keywords:

Electronic structure of equilateral triangular H_3^+ molecule
Three center wave functions
 D_{3h} symmetry

ABSTRACT

The $^1A_1'$, $^3E'$, $^1E'$, $^3A_1'$, $^1A_2'$, $^3A_2'$ representing the ground and the five excited states, which have the common character of being symmetrical with respect to reflection on the plane of the equilateral triangular H_3^+ molecule, are determined by an original three center wave function constructed by the use of the irreducible representations of the D_{3h} point group. In contrast to past large one center or linear combinations of atomic orbitals functions, our model has the advantage of being well adapted to all internuclear distances, with limited number of basis functions including the electron-electron term. Our functions satisfy, by their nature, the triangular geometry of the molecule and thus permit the study the asymptotic behavior of the potential energy curves of the fundamental and excited levels for which, new experimental and theoretical results are needed to confirm astronomical observations. The results of this work and the implementation of the computational techniques employed opens the way to further studies on complex three center systems.

1. Introduction

H_3^+ is the simplest existing polyatomic molecule, which provokes interest in different fields of chemistry and astronomy. It is most stable in the equilateral triangular configuration [1]. It has the particularity of dissociating both when an electron is attached or detached from it. It is the subject of many studies concerning specially the dissociative recombination with electrons [2,3], or the observations of its vibration-rotation band [4]. It plays also an important role in the domain of the study of magnetic and ionospheric properties of planets [5,6].

It is evident that, like helium and H_2 the smallest two electron systems, for which electron-electron correlation can be evaluated and understood both theoretically and experimentally (e.g. in the determination of the cross sections of the double ionization [7–9], the study of

the electronic structure of H_3^+ is one of the fundamental challenges of molecular physics. As the smallest three center molecule existing naturally, it has been largely studied in the past [10–16]. These calculations employ wave functions constructed by linear combinations of atomic orbitals (LCAO), Gaussians, or one center basis functions and do not include, in contrast to the present work, the electron-electron correlation term separately. We can mention here the original model presented in [17,18] in which the singlet and triplet excited states of H_3^+ are studied by the application of the “diatomic in molecules” method, employing a product of diatomic and atomic orbitals. Although these different types of orbitals succeed in producing comparable results for intermediate internuclear distances by employing very large Gaussian basis functions [16], still some disagreements exist between them and verified potential energy curves are needed specially for large

* Corresponding author.

E-mail address: boghos.joulakian@univ-lorraine.fr (B.B. Joulakian).

<https://doi.org/10.1016/j.cplett.2020.137304>

Received 9 January 2020; Received in revised form 1 March 2020; Accepted 2 March 2020

Available online 05 March 2020

0009-2614/ © 2020 Elsevier B.V. All rights reserved.

internuclear distances.

In recent years, Berencz type functions [19] have shown their efficiency in the treatment of many center electronic structures [20,21]. In this method, the LCAO applied to the diatomic cases, is replaced by a products of atomic orbitals centered on each nucleus. For the three center case, this model was successfully applied to the fundamental electronic level of H_3^+ [22]. Recently, we have introduced, for the three center two electron case, the electron-electron correlation to this type of functions and applied it to the determination of the multiply differential cross section of the simple ionization of H_3^+ by electron impact [23].

The aim of the present paper is to show that, the above mentioned three center model, which contains, by its nature, the equilateral triangular symmetry, presents many advantages, as it permits the application of the group theoretical irreducible representation of the D_{3h} point group to the construction of the wave functions of the first five excited states of H_3^+ , which are symmetrical with respect to reflection on the plane of the molecule, and permits to identify the dissociation limits of each level, and produces, with a small basis, compared to that applied in [16] for example, quite good accuracy for the energy values of the different levels. We believe also, that this wave functions will bring as in [23] theoretical support in the determination of cross sections in electronic excitation and ionization experiments. From a more practical point of view, we can say that the tackling of the analytic and numerical computational difficulties related to the two-electron three center problem in this work, opens the way to further developments in more complex three center systems, that we intend to study.

2. Theory

The Hamiltonian, which describes, for fixed nuclei, the two electrons of the H_3^+ ion (see Fig. 1) is written in atomic units as follows

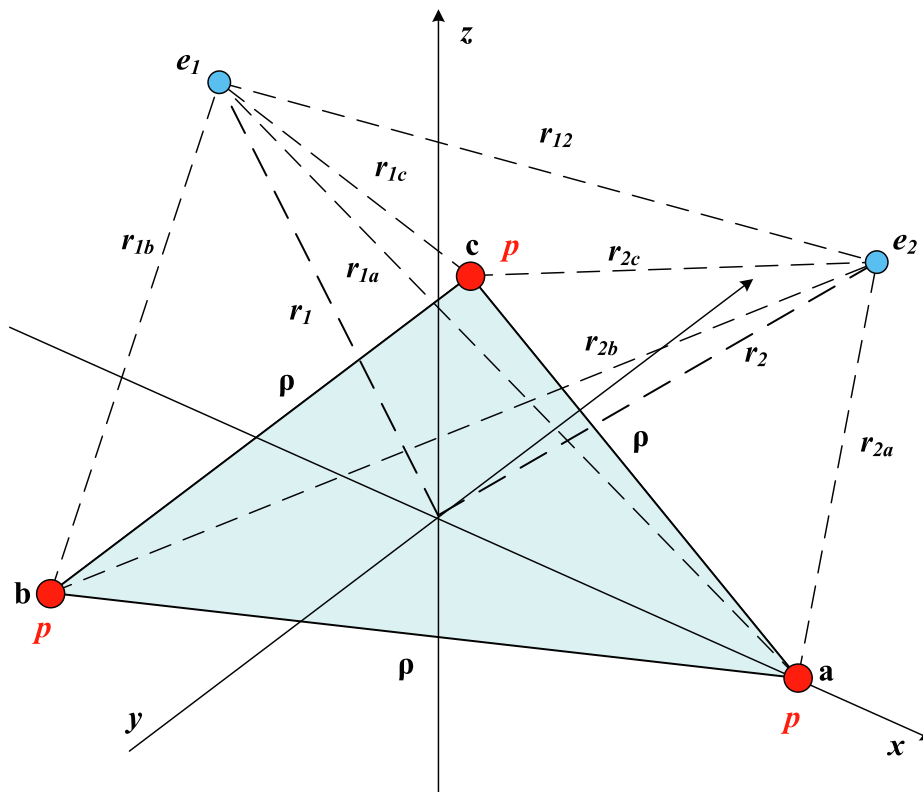


Fig. 1. The positions of the three fixed protons a , b , c and the two electrons e_1 , e_2 in a body fixed frame (x, y, z) with the origin on the barycenter of the equilateral triangle, and the z axis perpendicular to plane of the molecule H_3^+ .

$$\mathcal{H} = \sum_{j=1}^2 \left(-\frac{1}{2} \Delta_{\mathbf{r}_j} - \frac{1}{r_{ja}} - \frac{1}{r_{jb}} - \frac{1}{r_{jc}} \right) + \frac{1}{r_{12}} + \frac{3}{\rho}, \quad (1)$$

with $\mathbf{r}_{ja} = \mathbf{r}_j - \mathbf{a}$, $\mathbf{r}_{jb} = \mathbf{r}_j - \mathbf{b}$, $\mathbf{r}_{jc} = \mathbf{r}_j - \mathbf{c}$, $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$. Here \mathbf{r}_j gives the position of the j -th electron, and \mathbf{a} , \mathbf{b} , \mathbf{c} the position vectors of the three protons in a body fixed system of reference with the following coordinates:

$$\begin{aligned} \mathbf{a} &= \frac{\rho}{\sqrt{3}}(1, 0, 0), \\ \mathbf{b} &= \frac{\rho}{\sqrt{3}}\left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}, 0\right), \\ \mathbf{c} &= \frac{\rho}{\sqrt{3}}\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}, 0\right), \end{aligned} \quad (2)$$

with ρ representing the mutual internuclear distance between the three nuclei.

The computational schemes, which will deliver the wave functions and the energy values of the desired levels are based on the Rayleigh-Ritz variational functional with the electronic energy given by

$$\varepsilon_Q = \frac{\langle \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) | \mathcal{H} | \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) \rangle}{\langle \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) | \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) \rangle}, \quad (3)$$

where Q represents the energy levels ${}^1A_1'$, ${}^1E'$, ${}^3E'$, ${}^3A_1'$, ${}^1A_2'$, ${}^3A_2'$ under consideration and $\Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2)$ the corresponding trial wave function. We admit that these functions must be orthogonal

$$\langle \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) | \Psi_{Q'}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) \rangle = \delta_{QQ'}, \quad (4)$$

and satisfy the symmetry properties of the D_{3h} group.

2.1. The seven-parametric basis functions

Let us first form the bi-electronic correlated basis functions, with which we will construct the variational wave function $\Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2)$

of Eq. (3) for the different levels. We consider, as in [22], a combination of functions constructed by a product of two Berencz type [19] mono-electronic functions, adapted each to the equilateral triangular system and an electron-electron correlation term which fits in very elegantly. This forms the following seven-parametric function

$$\chi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = \exp(-\alpha_1 r_{1a} - \alpha_2 r_{1b} - \alpha_3 r_{1c} - \alpha_4 r_{2a} - \alpha_5 r_{2b} - \alpha_6 r_{2c} - \alpha_7 r_{12}). \quad (5)$$

Here the nonlinear α_i parameters will be determined by the variational method described below. The wave function of a given state $\Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2)$ of Eq. (3) must include all the permutations with respect to the three centers $\mathbf{a}, \mathbf{b}, \mathbf{c}$, and between electrons 1 and 2, (see Fig. 1). We will thus consider the following twelve functions, which have the structure of the one given in (5) representing all the permutation cases:

$$\begin{aligned} \chi_1 &= \chi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2), & \chi_2 &= \chi(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2), \\ \chi_3 &= \chi(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2), & \chi_4 &= \chi(\mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2), \\ \chi_5 &= \chi(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2), & \chi_6 &= \chi(\mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2), \end{aligned} \quad (6)$$

$$\begin{aligned} \chi_7 &= \chi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), & \chi_8 &= \chi(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_2, \mathbf{r}_1), \\ \chi_9 &= \chi(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_2, \mathbf{r}_1), & \chi_{10} &= \chi(\mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), \\ \chi_{11} &= \chi(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_2, \mathbf{r}_1), & \chi_{12} &= \chi(\mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{r}_2, \mathbf{r}_1). \end{aligned}$$

Our task will be to find the appropriate combinations of these twelve functions for each of the states ${}^1A'_1, {}^3E', {}^1E', {}^3A'_1, {}^1A'_2, {}^3A'_2$ defined by the irreducible representations of the D_{3h} point group.

2.2. Method of construction of the ground and excited state wave functions

The spin part of the two electron wave function being eliminated, the space wave functions of the singlet states such as ${}^1A'_1, {}^1A'_2$, etc. must be symmetrical with respect to exchange of electrons. Those of the triplet states, such as ${}^3A'_1, {}^3A'_2$, etc. must be antisymmetrical with respect to this exchange. So we can write the general conditions for the permutation of the two electrons:

$$\Psi_{1Q}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = +\Psi_{1Q}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), \quad (7)$$

$$\Psi_{3Q}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{3Q}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1). \quad (8)$$

Let us define two constants

$$\sigma_1 = \begin{cases} 0; & \text{for even states,} \\ 1; & \text{for odd states,} \end{cases} \quad (9)$$

which will characterize symmetry with respect to exchange of any two centers, and

$$\sigma_2 = \begin{cases} 0; & \text{singlet states,} \\ 1; & \text{triplet states,} \end{cases} \quad (10)$$

which will characterize the spin state.

Using these constants, let us associate the twelve basis functions χ_j , which have in common the exchange between the centers and \mathbf{c} . This will give us the following set of three new functions:

$$\begin{aligned} \psi_1 &= \chi_1 + (-1)^{\sigma_1} \chi_2 + (-1)^{\sigma_2} [\chi_7 + (-1)^{\sigma_1} \chi_8], \\ \psi_2 &= \chi_3 + (-1)^{\sigma_1} \chi_4 + (-1)^{\sigma_2} [\chi_9 + (-1)^{\sigma_1} \chi_{10}], \\ \psi_3 &= \chi_5 + (-1)^{\sigma_1} \chi_6 + (-1)^{\sigma_2} [\chi_{11} + (-1)^{\sigma_1} \chi_{12}]. \end{aligned} \quad (11)$$

We can now define the wave function for a given state Q in the following form

$$\Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = c_{1Q} \psi_1 + c_{2Q} \psi_2 + c_{3Q} \psi_3. \quad (12)$$

Here c_{iQ} -s correspond to the coefficient of ψ_i for the state Q .

Let us apply the circular permutation $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \rightarrow (\mathbf{b}, \mathbf{c}, \mathbf{a})$ to Eq. (12). We will then obtain a new combination for the same c_{iQ} s

$$\Psi_Q(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) = c_{1Q} \tilde{\psi}_1 + c_{2Q} \tilde{\psi}_2 + c_{3Q} \tilde{\psi}_3, \quad (13)$$

where the $\tilde{\psi}_i$ correspond to the association of the twelve basis functions having in common the exchange between \mathbf{a} and \mathbf{c} such that

$$\begin{aligned} \tilde{\psi}_1 &= \chi_5 + (-1)^{\sigma_1} \chi_4 + (-1)^{\sigma_2} [\chi_{11} + (-1)^{\sigma_1} \chi_{10}], \\ \tilde{\psi}_2 &= \chi_1 + (-1)^{\sigma_1} \chi_6 + (-1)^{\sigma_2} [\chi_7 + (-1)^{\sigma_1} \chi_{12}], \\ \tilde{\psi}_3 &= \chi_3 + (-1)^{\sigma_1} \chi_2 + (-1)^{\sigma_2} [\chi_9 + (-1)^{\sigma_1} \chi_8]. \end{aligned} \quad (14)$$

Finally a third circular permutation is possible $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \rightarrow (\mathbf{c}, \mathbf{a}, \mathbf{b})$, which will result into

$$\Psi_Q(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = c_{1Q} \hat{\psi}_1 + c_{2Q} \hat{\psi}_2 + c_{3Q} \hat{\psi}_3, \quad (15)$$

with

$$\begin{aligned} \hat{\psi}_1 &= \chi_3 + (-1)^{\sigma_1} \chi_6 + (-1)^{\sigma_2} [\chi_9 + (-1)^{\sigma_1} \chi_{12}], \\ \hat{\psi}_2 &= \chi_5 + (-1)^{\sigma_1} \chi_2 + (-1)^{\sigma_2} [\chi_{11} + (-1)^{\sigma_1} \chi_8], \\ \hat{\psi}_3 &= \chi_1 + (-1)^{\sigma_1} \chi_4 + (-1)^{\sigma_2} [\chi_7 + (-1)^{\sigma_1} \chi_{10}], \end{aligned} \quad (16)$$

where the exchange is between \mathbf{a} and \mathbf{b} .

By adding the three functions of the Eq. (12), (13) and (15), we obtain the following relation

$$\begin{aligned} \Xi_Q &= \Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) + \Psi_Q(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) + \Psi_Q(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) \\ &= (c_{1Q} + c_{2Q} + c_{3Q}) \\ &\times \sum_{k=0}^2 \{ \chi_{1+2k} + (-1)^{\sigma_1} \chi_{2+2k} + (-1)^{\sigma_2} [\chi_{7+2k} + (-1)^{\sigma_1} \chi_{8+2k}] \} \\ &= (c_{1Q} + c_{2Q} + c_{3Q}) (\psi_1 + \psi_2 + \psi_3). \end{aligned} \quad (17)$$

This is a general relation valid to all the levels. Our aim during the minimization process is to determine the coefficients c_{iQ} and the parameters α_i for each level Q defined above. Before passing to the variational determination, let us first exploit the symmetry properties of equilateral triangular system to simplify the relations between the coefficients c_{iQ} of each state.

2.3. The symmetry properties of the ground and excited states

Let us begin with the A'_1 states, for which any permutation of the centers should leave the wave function invariable, such that

$$\begin{aligned} \Psi_{A'_1}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) &= \text{perm}_{(\mathbf{a}, \mathbf{b}, \mathbf{c})} \Psi_{A'_1}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) \\ &= \frac{1}{3} \Xi_{A'_1} = c_{1A'_1} \psi_1 + c_{2A'_1} \psi_2 + c_{3A'_1} \psi_3. \end{aligned} \quad (18)$$

Here $\text{perm}_{(\mathbf{a}, \mathbf{b}, \mathbf{c})}$ is the permutation operator for the three nuclei ($\mathbf{a} \leftrightarrow \mathbf{b} \leftrightarrow \mathbf{c}$). Introducing this relation in Eq. (17) we can show that the coefficients of the ${}^1A'_1$ and ${}^3A'_1$ states (i.e. for $\sigma_1 = 0$ in Eq. (11)) must satisfy the condition

$$c_{1A'_1} = c_{2A'_1} = c_{3A'_1} \quad (19)$$

For the A'_2 states, the wave functions are antisymmetric with respect to exchange of the two nuclei, such that

$$\begin{aligned} \Psi_{A'_2}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) &= -\Psi_{A'_2}(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = \Psi_{A'_2}(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) \\ &= -\Psi_{A'_2}(\mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) = \Psi_{A'_2}(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{A'_2}(\mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2). \end{aligned} \quad (20)$$

This means that

$$\begin{aligned} \Psi_{A'_2}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) &= \Psi_{A'_2}(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) = \Psi_{A'_2}(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) \\ &= \frac{1}{3} \Xi_{A'_2} = c_{1A'_2} \psi_1 + c_{2A'_2} \psi_2 + c_{3A'_2} \psi_3. \end{aligned} \quad (21)$$

Comparing with Eq. (17) corresponding to ${}^1A'_2$ and ${}^3A'_2$ (i.e. for $\sigma_1 = 1$ in Eq. (11)) we can find that

$$c_{1A'_2} = c_{2A'_2} = c_{3A'_2}. \quad (22)$$

Let us pass to the E' states. The singlet ${}^1E'$ and the triplet ${}^3E'$ states are doubly degenerate, such that they are symmetric or antisymmetric

with respect to the exchange of any two nuclei. We will designate these states by $e^{1E'}$ and $e^{3E'}$ for the symmetrical (even) cases, and by $o^{1E'}$ and $o^{3E'}$ for the antisymmetrical (odd) cases, such that:

$$\Psi_{e^{1E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = +\Psi_{e^{1E'}}(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = +\Psi_{e^{1E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), \quad (23)$$

$$\Psi_{o^{1E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{o^{1E'}}(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = +\Psi_{o^{1E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), \quad (24)$$

$$\Psi_{e^{3E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = +\Psi_{e^{3E'}}(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{e^{3E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1), \quad (25)$$

$$\Psi_{o^{3E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{o^{3E'}}(\mathbf{a}, \mathbf{c}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = -\Psi_{o^{3E'}}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_2, \mathbf{r}_1). \quad (26)$$

We have demonstrated in the A (Eqs. (A.9), (A.13)) that

$$\Xi_{E'} = \Psi_{E'}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) + \Psi_{E'}(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) + \Psi_{E'}(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) = 0. \quad (27)$$

Using Eq. (17) we can write

$$\Xi_{E'} = (c_{1E'} + c_{2E'} + c_{3E'}) (\psi_1 + \psi_2 + \psi_3) = 0, \quad (28)$$

which imposes the condition

$$c_{1E'} + c_{2E'} + c_{3E'} = 0. \quad (29)$$

Till now, we have studied the case, where we had a basis function with 7 parameters Eq. (5). We will now extend our choice to additional 7 nonlinear parameters α_i , with indices $i = 8, \dots, 14$. This will create twelve additional functions χ_j with $j = 13, \dots, 24$, thus three more functions like in Eq. (11)

$$\begin{aligned} \psi_4 &= \chi_{13} + (-1)^{\sigma_1} \chi_{14} + (-1)^{\sigma_2} [\chi_{19} + (-1)^{\sigma_1} \chi_{20}], \\ \psi_5 &= \chi_{15} + (-1)^{\sigma_1} \chi_{16} + (-1)^{\sigma_2} [\chi_{21} + (-1)^{\sigma_1} \chi_{22}], \\ \psi_6 &= \chi_{17} + (-1)^{\sigma_1} \chi_{18} + (-1)^{\sigma_2} [\chi_{23} + (-1)^{\sigma_1} \chi_{24}], \end{aligned} \quad (30)$$

with which we can define our extended wave function for a given state Q in the following form

$$\Psi_Q(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) = \sum_{i=1}^6 c_{iQ} \psi_i. \quad (31)$$

Applying the symmetry conditions to the new function we can show for the complete set of coefficients we must impose the following conditions

$$\begin{aligned} c_{1A_1} &= c_{2A_1} = c_{3A_1}, & c_{4A_1} &= c_{5A_1} = c_{6A_1}, \\ c_{1A_2} &= c_{2A_2} = c_{3A_2}, & c_{4A_2} &= c_{5A_2} = c_{6A_2}, \\ c_{1E'} + c_{2E'} + c_{3E'} &= 0, \\ c_{4E'} + c_{5E'} + c_{6E'} &= 0. \end{aligned} \quad (32)$$

Table 1

The energies values (in au) of the ground $^1A_1'$ and excited $^3E'$, $^1E'$, $^3A_1'$, $^1A_2'$ and $^3A_2'$ states of H_3^+ for internuclear distance $\rho = 1.65$ au. The second row corresponds to results obtained with seven-parameter functions and the third row to those obtained by fourteen-parameter functions. The lower rows show existing results from the references [12,15,16,22,23] (where [12] at $\rho = 1.63332$ au and [23] with 6 parameters function at $\alpha_7 = 0$).

	$^1A_1'$	$^3E'$	$^1E'$	$^3A_1'$	$^1A_2'$	$^3A_2'$
with 7-parameters	-1.340 352	-0.775 600	-0.609 823	-0.498 901	-0.186 083	-0.028 621
with 14-parameters	-1.342 520	-0.792 082	-0.626 730	-0.510 758	-0.209 148	-0.035 843
[12]		-0.776 695	-0.622 277	-0.496 986		
[15]	-1.342 230		-0.632 050			
[16]	-1.343 835		-0.633 512	-0.511 569		
[22]	-1.340 345					
[23]	-1.331 48					

3. Numerical methods

After substituting the expansion Eq. (12) for the simpler seven-parameter case, or Eq. (31) for the extended case into the variational functional Eq. (3) and minimizing, we obtain the generalized eigenvalue problem

$$\mathbf{A}_Q \mathbf{c}_Q = \varepsilon_Q \mathbf{B}_Q \mathbf{c}_Q, \quad \mathbf{c}_Q^T \mathbf{B}_Q \mathbf{c}_Q = 1. \quad (33)$$

Here \mathbf{A}_Q and \mathbf{B}_Q represent symmetric matrices with matrix elements respectively

$$A_{ij} = \langle \psi_i | \mathcal{H} | \psi_j \rangle, \quad B_{ij} = \langle \psi_i | \psi_j \rangle, \quad (34)$$

$\mathbf{c}_Q = (c_{1Q}, c_{2Q}, c_{3Q})^T$ (or $\mathbf{c}_Q = (c_{1Q}, c_{2Q}, c_{3Q}, c_{4Q}, c_{5Q}, c_{6Q})^T$ for the extended case) represent the eigenvectors. Using for each state Q the conditions of Eqs. (19), (22) and (29) (or (32) and applying linear transformation of the Eq. (33), we reduce the matrices and obtain the following problem

$$\tilde{\mathbf{A}}_Q \tilde{\mathbf{c}}_Q = \varepsilon_Q \tilde{\mathbf{B}}_Q \tilde{\mathbf{c}}_Q, \quad \tilde{\mathbf{c}}_Q^T \tilde{\mathbf{B}}_Q \tilde{\mathbf{c}}_Q = 1, \quad (35)$$

with $\tilde{\mathbf{c}}_Q = (c_{1Q})^T$ (or $\tilde{\mathbf{c}}_Q = (c_{1Q}, c_{4Q})^T$) for A' states, and $\tilde{\mathbf{c}}_Q = (c_{1Q}, c_{2Q})^T$ (or $\tilde{\mathbf{c}}_Q = (c_{1Q}, c_{2Q}, c_{4Q}, c_{5Q})^T$) for E' states.

To minimize the energy ε_Q with the variation of the parameters α_i , we have used a sequential quadratic programming method for several variables [24–26] and the code E04UCF from NAG Library [27] with additional constraints on the parameters:

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_7 &> 0, \\ \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 &> 0, \\ \alpha_8 + \alpha_9 + \alpha_{10} + \alpha_{14} &> 0, \\ \alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14} &> 0. \end{aligned} \quad (36)$$

We have also performed the optimization as in our previous paper [23], where we have studied only the fundamental state, by calculating the first derivatives with the parameters of the energy:

$$\frac{\partial \varepsilon_Q}{\partial \alpha_i} = \tilde{\mathbf{c}}_Q^T \left(\frac{\partial \tilde{\mathbf{A}}_Q}{\partial \alpha_i} - \varepsilon_Q \frac{\partial \tilde{\mathbf{B}}_Q}{\partial \alpha_i} \right) \tilde{\mathbf{c}}_Q. \quad (37)$$

All 6D integrals, which appear in the functional of the energy were calculated numerically using a globally adaptive subdivision scheme [28–30] and a code Cuhre [31]. All numerical integrations were done with an absolute accuracy $10^{-6} - 10^{-7}$.

4. Results and discussion

As we mentioned above, the aim of our present work is, among others, to implement the application of a new three center two electron correlated wave function, described above, for the triangular equilateral case represented by the H_3^+ system. The motivations for such a work are multiple. These are, for instance, the need for observation of

Table 2Optimal seven-variational parameters $\alpha_1, \dots, \alpha_7$ for the ground $^1A'_1$ and excited $^3E', ^1E', ^3A'_1, ^1A'_2$ and $^3A'_2$ states of H_3^+ for $\rho = 1.65$ au.

	$^1A'_1$	$^3E'$	$^1E'$	$^3A'_1$	$^1A'_2$	$^3A'_2$
α_1	-0.026 442	0.240 329	0.535 125	0.080 623	-2.028 037	0.436 381
α_2	0.211 880	0.504 412	0.137 175	-0.270 522	1.024 753	-0.229 632
α_3	1.409 787	0.689 107	0.994 871	0.794 331	1.136 674	0.858 861
α_4	1.068 443	0.241 417	0.901 676	1.181 098	0.561 273	0.990 903
α_5	0.122 722	0.773 526	-0.044 017	0.099 545	0.552 654	0.157 249
α_6	0.603 098	0.459 867	0.042 509	0.369 651	0.552 356	-0.193 064
α_7	-0.212 311	-0.249 522	-0.166 339	-0.088 500	-0.014 324	-0.084 081

Table 3The coefficients c_1, c_2 and c_3 corresponding to each energy level.

	c_1	c_2	c_3
$^1A'_1$	0.220 519	c_1	c_1
even $^3E'$	15.23 642	-5.857 385	$-c_1 - c_2$
odd $^3E'$	2.036 528	-14.29 091	$-c_1 - c_2$
even $^1E'$	0.376 844	-0.209 584	$-c_1 - c_2$
odd $^1E'$	0.024 434	0.314 108	$-c_1 - c_2$
$^3A'_1$	0.061 310	c_1	c_1
$^1A'_2$	0.078 908	c_1	c_1
$^3A'_2$	0.170 436	c_1	c_1

electronic transitions in interstellar media, where this ion is present abundantly, more, the fact that, like atomic helium and diatomic H_2 , triatomic H_3^+ is a two electron system in which, electron–electron correlation is identifiable, because it is the principal cause of many photo-excitation effects such as the photo-double-ionization of this type of targets. Our work is also motivated by the possibility that these three center calculations develop benchmark procedures extendable to more complex three center molecules. We seek also to verify our variational procedure and the numerical calculations of six order integrals involving the two bound electrons in the triangular coulomb field.

We begin by presenting, on Table 1, the energy values of the ground $^1A'_1$ and excited $^3E', ^1E', ^3A'_1, ^1A'_2$ and $^3A'_2$ states of H_3^+ for the ground state equilibrium internuclear distance $\rho = 1.65$ au. On this Table the second row shows the results obtained by the wave function of Eq. (12) having seven-parameters, while the third row shows the results obtained by the expression of Eq. (31) having fourteen-parameter functions. We observe that the extended fourteen-parameter (Eq. (31)) wave function improves the results. These $^1A'_1, ^3E', ^1E'$ and $^3A'_1$ are in good agreement (up to three digits after the decimal point) with the results of [12,15,16,22]. The influence of the electron–electron correlation term can be seen by the comparison of the energy values of $^1A'_1$, level of this table with those

Table 4Optimal fourteen-variational parameters $\alpha_1, \dots, \alpha_{14}$ for the ground $^1A'_1$ and excited $^3E', ^1E', ^3A'_1, ^1A'_2$ and $^3A'_2$ states energies of H_3^+ at the for $\rho = 1.65$ au.

	$^1A'_1$	$^3E'$	$^1E'$	$^3A'_1$	$^1A'_2$	$^3A'_2$
α_1	0.115 567	0.083 428	0.208 380	1.260 608	1.091 554	0.210 296
α_2	0.136 146	-0.009 707	0.223 048	0.233 957	-0.906 230	1.241 690
α_3	1.499 119	0.975 277	1.280 699	0.215 901	0.142 430	0.317 057
α_4	1.124 039	0.200 127	0.869 745	0.262 064	0.558 355	0.739 985
α_5	0.093 855	1.216 321	0.042 846	0.509 664	0.550 752	0.372 041
α_6	0.587 165	0.248 488	-0.023 289	0.053 417	0.554 221	0.041 675
α_7	-0.467 799	-0.103 159	-0.117 594	-0.026 991	-0.018 901	-0.057 578
α_8	0.595 668	1.239 064	0.297 544	0.243 724	0.557 903	0.441 187
α_9	0.084 919	0.209 395	0.098 634	0.207 222	0.551 828	-0.065 797
α_{10}	1.142 581	0.298 909	1.314 096	1.251 002	0.554 588	0.808 231
α_{11}	1.507 875	0.606 583	0.619 055	0.538 336	1.063 816	1.0302191
α_{12}	0.126 634	0.645 124	-0.359 666	0.102 680	-0.889 231	0.165 134
α_{13}	0.136 727	-0.167 738	0.659 638	0.311 287	0.130 554	-0.032 221
α_{14}	-0.511 026	-0.084 233	-0.214 367	-0.013 325	-0.019 960	-0.087 467

given on the last row in [23] for $\alpha_7 = 0$.

Now in the aim of verifying the process of convergence and trying to approach the results of [16] which uses a very large basis we additionally calculated the ground state $^1A'_1$, level the energy $\epsilon_{^1A'_1}$ with twenty-one-parameters and obtained $\epsilon_{^1A'_1} = -1.343 083$ au. This shows that our procedure converges and brings the agreement up to four digits. Using the extrapolation formula given in [32]

$$\epsilon_{^1A'_1}^{qs} = \epsilon_{^1A'_1}(N) + \frac{C}{N^\beta}, \quad (38)$$

we will define the convergence rate β with respect to the number of the basis set N . Here $\epsilon_{^1A'_1}^{qs}$ is the extrapolated value of the energy, i.e. $N \rightarrow \infty$, $\epsilon_{^1A'_1}(N)$ is the energy for a given N basis set, and C is the constant. We have for $N = 1$ the result of the seven-parameter case with 12 basis functions, for $N = 2$ that of fourteen-parameter case with 24 basis functions and $N = 3$ that of twenty-one-parameter case with 36 basis functions. $\epsilon_{^1A'_1}^{qs}$, β and C are obtained from the system of nonlinear problem (38) at $N = 1 - 3$:

$$\epsilon_{^1A'_1}^{qs} = -1.343 796 \text{ au}, \quad \beta = 1.429876, \quad C = -0.003447. \quad (39)$$

One can see that the extrapolated value of the energy is very close to the result of [16].

In Tables 2 and 3, respectively, we display the optimal seven-variational parameters $\alpha_i, i = 1, \dots, 7$ for the ground and excited states energies under consideration, and the corresponding coefficients $c_{iQ}, i = 1, \dots, 3$ including odd and even excited states of $^3E', ^1E'$. In Tables 4 and 5, we display the same cases as in Tables 2 and 3, but for the optimal fourteen-variational parameters $\alpha_i, i = 1, \dots, 14$ and corresponding coefficients $c_{iQ}, i = 1, \dots, 6$. We must mention here that some parameters α_i in Tables 2 and 4 have negative values. This is quite normal, as long as the conditions of Eq. (36) are satisfied. The negative values α_7 and α_{14} can be particularly reasonable as the distribution of the two electrons must have relatively higher density for large r_{12} , such that they have higher probability to be apart, because of the two electron repulsion potential $1/r_{12}$.

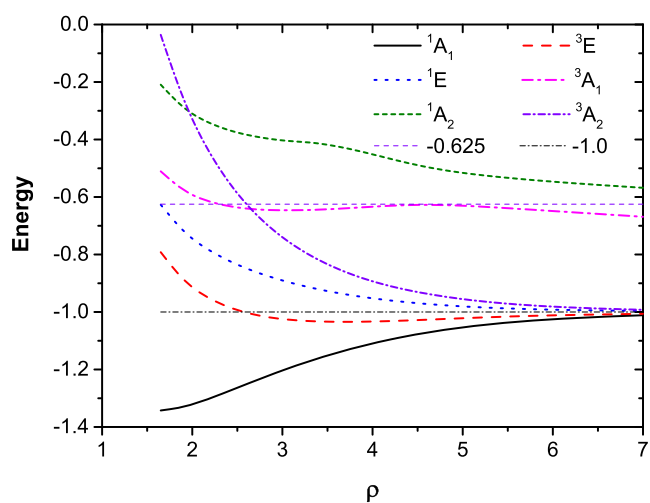
To compare further our results with existing ones. We consider

Table 5The coefficients c_1 , c_2 , c_3 , c_4 , c_5 and c_6 corresponding to each energy level (see the Table 4).

	c_1	c_2	c_3	c_4	c_5	c_6
$^1A'_1$	1.274 131	c_1	c_1	-1.069 841	c_4	c_4
even $^3E'$	0.060 304	0.273 509	$-c_1 - c_2$	0.410 264	-0.035 898	$-c_4 - c_5$
odd $^3E'$	0.349 790	-0.228 055	$-c_1 - c_2$	0.196 415	-0.451 523	$-c_4 - c_5$
even $^1E'$	-0.289 907	0.073 125	$-c_1 - c_2$	0.023 981	0.125 188	$-c_4 - c_5$
odd $^1E'$	0.082 955	-0.292 542	$-c_1 - c_2$	-0.158 406	0.099 972	$-c_4 - c_5$
$^3A'_1$	0.829 633	c_1	c_1	-1.052 157	c_4	c_4
$^1A'_2$	-2.149 799	c_1	c_1	2.236 809	c_4	c_4
$^3A'_2$	1.064 310	c_1	c_1	0.719 721	c_4	c_4

Table 6The ground $^1A'_1$ and excited $^3E'$, $^1E'$, $^3A'_1$, $^1A'_2$ and $^3A'_2$ states energies (in au) of H_3^+ versus the internuclear distance ρ .

ρ	$^1A'_1$	$^3E'$	$^1E'$	$^3A'_1$	$^1A'_2$	$^3A'_2$
1.650	-1.342 520	-0.792 082	-0.626 730	-0.510 758	-0.209 149	-0.035 843
1.900	-1.329 584	-0.885 479	-0.716 543	-0.574 405	-0.287 441	-0.255 484
2.200	-1.299 819	-0.953 837	-0.787 186	-0.616 226	-0.343 498	-0.449 428
2.750	-1.232 880	-1.013 051	-0.865 638	-0.643 901	-0.392 400	-0.673 110
3.052	-1.197 553	-1.026 233	-0.894 834	-0.645 996	-0.404 557	-0.751 866
3.350	-1.165 897	-1.032 265	-0.917 646	-0.644 671	-0.412 183	-0.809 932
4.500	-1.077 293	-1.028 068	-0.969 321	-0.627 978	-0.489 352	-0.930 539
5.500	-1.036 820	-1.016 830	-0.987 624	-0.638 934	-0.533 256	-0.971 017
7.000	-1.011 227	-1.006 228	-0.997 107	-0.668 777	-0.567 851	-0.992 534

**Fig. 2.** Calculated potential-energy curves for the ground $^1A'_1$ and excited $^3E'$, $^1E'$, $^3A'_1$, $^1A'_2$ and $^3A'_2$ states energies of H_3^+ versus the internuclear distance ρ .

particularly the results given in [15], which we think is the only reference that gives numerical results for singlet states of H_3^+ for different internuclear distances ρ . The comparison is made on Table 6. We observe that the singlet $^1A'_1$, $^1E'$ states energies are in good agreement (up to three digits after the decimal point) with results of [15] (see their Table III, concerning the D_{3h} symmetry). What concerns the singlet $^1A'_2$ state energy for which we have some inconsistencies with [15] at $\rho \geq 2.75$ au, we will analyze the situation in more detail below. Also, we can observe that the minimum of the $^1E'$ state energy is comparable with the results of Alijah et al. [33].

Let us now pass to the potential-energy curves. We have two possible asymptotic limits for large ρ shown on Fig. 2. The lower one corresponds to the energy of a system constituted by two separate hydrogens having their electrons on the 1s level. The second higher level corresponds to the same system, but with one hydrogen on the 1s level

and the second on the 2s giving a total energy of $-0.5 - 0.125$ au. Now, we observe on Fig. 2 that the curves of the levels $^1A'_1$, $^3E'$, $^1E'$ and $^3A'_2$, are consistent with the curves of [14] (see their Figures IV and V in [14]), and have a common asymptotic limit -1.0 au. This shows that at large ρ the wave functions of these states which have in our model conserved their D_{3h} character are still valid at the dissociation limit of $H(1s) + H(1s) + H^+$.

What concerns the state $^3A'_1$ which should normally dissociate into $H(1s) + H(2s) + H^+$ has a particular behavior above $\rho = 4$ au. It doesn't continue to the limit $-0.5 - 0.125$ au but it goes down (see please Fig. 2). In fact the potential curve is comparable to that of [12,14] which stops at the distance $\rho = 4$ au. We believe that in the equilateral triangular D_{3h} case, which we are adopting in our calculations even for large ρ , this state possesses two possible dissociation cases which could satisfy this configuration, one being the $H(1s) + H(2s) + H^+$ which could be the most probable, and the second the $H(1s) + H_2^+$ with large ρ , which has a lower energy limit than that of $H(1s) + H(2s) + H^+$ system. The existence of these two dissociation channels can explain the form of the potential energy curve of this level, which decreases after $\rho = 5$ au.

5. Conclusion

In this paper we construct original three center correlated wave functions necessary for the theoretical study of the electronic structure of the ground and the first five excited states of the equilateral triangular H_3^+ , which have the common character of being symmetrical with respect to reflection on the plane of the molecule. Our functions which possess by their nature the triangular symmetry, include electron-electron correlation and respect the irreducible representations of the D_{3h} point group. Our results concerning the electronic energy values of these levels, which are necessary to guide future experimental observations, confirm and complete the existing results which are obtained by large basis functions. Our functions permit also the determination of the asymptotic behavior of the potential energy curves, which show the possible dissociation fragments for this particular

equilateral symmetry. The three center two electron integrals are determined by applying new numerical and analytical approaches which open the way to further applications of this type of functions to more complex three center molecules.

CRedit authorship contribution statement

O. Chuluunbaatar: Methodology, Software. **S. Obeid:** Software. **B.B. Joulakian:** Conceptualization. **A.A. Gusev:** Conceptualization, Software. **P.M. Krassovitskiy:** Conceptualization, Software. **L.A. Sevastianov:** Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial

Appendix A. Demonstration of the Eq. (27) for the E' states

We want to show that

$$\Psi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2) + \Psi(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{r}_1, \mathbf{r}_2) + \Psi(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{r}_1, \mathbf{r}_2) = 0. \quad (\text{A.1})$$

For the sake of simplicity, let's represent the wave function of the two electron equilateral triangular system $\Psi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{r}_1, \mathbf{r}_2)$ in this form $\Psi(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2)$ where we show only the azimuthal angles of the two electrons. The three centers being given in Eq. (2). We can verify that for this equilateral triangular form

$$\begin{aligned} \Psi(\mathbf{c}, \mathbf{a}, \mathbf{b}, \phi_1, \phi_2) &= \Psi\left(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1 - \frac{2\pi}{3}, \phi_2 - \frac{2\pi}{3}\right), \\ \Psi(\mathbf{b}, \mathbf{c}, \mathbf{a}, \phi_1, \phi_2) &= \Psi\left(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1 + \frac{2\pi}{3}, \phi_2 + \frac{2\pi}{3}\right). \end{aligned} \quad (\text{A.2})$$

This permits us to write in Eq. (A.1) in the following more compact form

$$\sum_{s=-1}^1 \Psi\left(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1 + s\frac{2\pi}{3}, \phi_2 + s\frac{2\pi}{3}\right) = 0. \quad (\text{A.3})$$

Let us make use of the fact, that any square-integrable function $f(\theta_1, \phi_1, \theta_2, \phi_2)$ on the unit sphere, can be expressed as a linear combination of the product of the real spherical harmonic functions $Y_{l_1 m_1}(\theta_1, \phi_1)$ and $Y_{l_2 m_2}(\theta_2, \phi_2)$:

$$f(\theta_1, \phi_1, \theta_2, \phi_2) = \sum_{l_1, l_2=0}^{\infty} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} f_{l_1 m_1 l_2 m_2} Y_{l_1 m_1}(\theta_1, \phi_1) Y_{l_2 m_2}(\theta_2, \phi_2). \quad (\text{A.4})$$

This can also be written explicitly

$$\begin{aligned} f(\theta_1, \phi_1, \theta_2, \phi_2) &= \sum_{l_1, l_2=0}^{\infty} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} [\hat{f}_{l_1 m_1 l_2 m_2}(\theta_1, \theta_2) \cos(m_1 \phi_1 - m_2 \phi_2) + \hat{g}_{l_1 m_1 l_2 m_2}(\theta_1, \theta_2) \sin(m_1 \phi_1 - m_2 \phi_2)]. \end{aligned} \quad (\text{A.5})$$

A.1. The case of the even E' state wave function

Let us first consider the even E' state wave function. It should be symmetrical with respect to the following inversion of the signs of the azimuthal angles:

$$\Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) = \Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, -\phi_1, -\phi_2). \quad (\text{A.6})$$

Using the development of Eq. (A.5), we can express the wave function in the following form where the term with $[\sin(m_1 \phi_1 - m_2 \phi_2)]$ does not appear

$$\Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) = \sum_{l_1, l_2=0}^{\infty} \sum_{\substack{|m_1| \leq l_1, |m_2| \leq l_2, \\ \text{mod}(m_1 - m_2, 3) \neq 0}} \hat{f}_{l_1 m_1 l_2 m_2}(r_1, r_2, \theta_1, \theta_2) \cos(m_1 \phi_1 - m_2 \phi_2). \quad (\text{A.7})$$

In this decomposition terms with $\text{mod}(m_1 - m_2, 3) = 0$ should be excluded to insure the orthogonality of the wave functions $\Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2)$ and $\Psi_{A_1'}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2)$.

Now for the terms for which $\text{mod}(m_1 - m_2, 3) \neq 0$ we have the following relation.

$$\sum_{s=-1}^1 \cos\left(m_1\left(\phi_1 + s\frac{2\pi}{3}\right) - m_2\left(\phi_2 + s\frac{2\pi}{3}\right)\right) = 0. \quad (\text{A.8})$$

From here, taking into account the symmetry conditions (A.2), we deduce the relation

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Hulubei-Meshcheryakov JINR-Romania program, RUDN University Program 5-100 and grant of Plenipotentiary of the Republic of Kazakhstan in JINR. Most of the calculations were performed on Central Information and Computer Complex, and heterogeneous computing platform HybriLIT through supercomputer "GOVORUN" of the Joint Institute for Nuclear Research.

$$\sum_{s=-1}^1 \Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1 + s\frac{2\pi}{3}, \phi_2 + s\frac{2\pi}{3}) \equiv \Psi_e(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) + \Psi_e(\mathbf{c}, \mathbf{a}, \mathbf{b}, \phi_1, \phi_2) + \Psi_e(\mathbf{b}, \mathbf{c}, \mathbf{a}, \phi_1, \phi_2) = 0. \quad (\text{A.9})$$

A.2. The odd E' type state wavefunction

Finally we consider the odd E' state wavefunction, antisymmetric with respect to the inversion of the sign of the two azimuthal angles

$$\Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) = -\Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, r, -\phi_1, -\phi_2). \quad (\text{A.10})$$

As in the even case, we can write the odd wavefunction in terms of the development of Eq. (A.5) including now the $\sin(m_1\phi_1 - m_2\phi_2)$ part

$$\Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) = \sum_{l_1, l_2=0}^{\infty} \sum_{\substack{|m_1| \leq l_1, |m_2| \leq l_2; \\ \text{mod}(m_1 - m_2, 3) \neq 0}} \hat{g}_{l_1 m_1 l_2 m_2}(r_1, r_2, \theta_1, \theta_2) \sin(m_1\phi_1 - m_2\phi_2). \quad (\text{A.11})$$

This decomposition also doesn't contain terms with $\text{mod}(m_1 - m_2, 3) = 0$, since the wavefunctions $\Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2)$ and $\Psi_{A_2}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2)$ are orthogonal. Now for $\text{mod}(m_1 - m_2, 3) \neq 0$ we should have

$$\sum_{s=-1}^1 \sin\left(m_1\left(\phi_1 + s\frac{2\pi}{3}\right) - m_2\left(\phi_2 + s\frac{2\pi}{3}\right)\right) = 0. \quad (\text{A.12})$$

From here, taking into account the symmetry conditions (A.2), we have relation

$$\sum_{s=-1}^1 \Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1 + s\frac{2\pi}{3}, \phi_2 + s\frac{2\pi}{3}) \equiv \Psi_0(\mathbf{a}, \mathbf{b}, \mathbf{c}, \phi_1, \phi_2) + \Psi_0(\mathbf{c}, \mathbf{a}, \mathbf{b}, \phi_1, \phi_2) + \Psi_0(\mathbf{b}, \mathbf{c}, \mathbf{a}, \phi_1, \phi_2) = 0. \quad (\text{A.13})$$

References

- [1] P. Allmendinger, J. Deiglmayr, O. Schullian, K. Hoveler, J.A. Agner, H. Schmutz, F. Merkt, *ChemPhysChem* 17 (2016) 3596–3608.
- [2] H. Helm, U. Galster, I. Mistrk, U. Müller, R. Reichle, Steven Guberman (Ed.), *Dissociative Recombination of Molecular Ions with Electrons*, Springer US, 2003, pp. 275–288.
- [3] D. Strasser, J. Levin, H.B. Pedersen, O. Heber, A. Wolf, D. Schwalm, D. Zajfman, *Phys. Rev. A* 65 (2001) 010702-1–010702-4.
- [4] T. Oka, *Phys. Rev. Lett.* 45 (1980) 531–534.
- [5] M. Goto, T.R. Geballe, T. Usuda, *Astrophys. J.* 806 (2015) 57-1–57-8.
- [6] H.A. Lam, S. Miller, R.D. Joseph, T.R. Geballe, L.M. Trafton, J. Tennyson, G.E. Ballester, *Astrophys. J.* 474 (1997) L73–L76.
- [7] F.W. Byron Jr., C.J. Joachain, *Phys. Rev.* 164 (1965) 1–9.
- [8] H. Le Rouzo, C. Dal Cappello, *Phys. Rev. A* 43 (1991) 318–329.
- [9] O. Schwarzkopf, B. Krässig, J. Elniger, V. Schmidt, *Phys. Rev. Lett.* 70 (1993) 3008–3011.
- [10] J. Hirschfelder, H. Eyring, N. Rosen, *J. Chem. Phys.* 4 (1936) 130–133.
- [11] R.E. Christoffersen, *J. Chem. Phys.* 41 (1964) 960–971.
- [12] K. Kawaoka, R.F. Borkman, *J. Chem. Phys.* 54 (1971) 4234–4238.
- [13] K. Kawaoka, R.F. Borkman, *J. Chem. Phys.* 55 (1971) 4637–4641.
- [14] L.J. Schaad, W.V. Hicks, *J. Chem. Phys.* 61 (1974) 1934–1942.
- [15] D. Talbi, R.P. Saxon, *J. Chem. Phys.* 89 (1988) 2235–2241.
- [16] M. Pavanello, L. Adamowicz, *J. Chem. Phys.* 130 (2009) 034104-1–034104-6.
- [17] Ay-ju A. Wu, Frank O. Ellison, *J. Chem. Phys.* 48 (1968) 1491–1496.
- [18] Ay-ju A. Wu, Frank O. Ellison, *J. Chem. Phys.* 48 (1968) 5032–5037.
- [19] F. Berencz, *Acta. Phys. Acad. Sci. Hung.* 6 (1957) 423–441.
- [20] B. Joulakian, J. Hanssen, R. Rivarola, A. Motassim, *Phys. Rev. A* 54 (1996) 1473–1479.
- [21] O. Chuluunbaatar, B.B. Joulakian, Kh. Tsookhuu, S.I. Vinitzky, *J. Phys. B* 37 (2014) 2607–2616.
- [22] J.C. Lopez Vieyra, A.V. Turbiner, H. Medal-Cobaxin, *J. Phys. B* 44 (2011) 195101-1–195101-6.
- [23] S. Obeid, O. Chuluunbaatar, B.B. Joulakian, *J. Phys. B* 50 (2017) 145201-1–145201-9.
- [24] P.E. Gill, W. Murray, M.H. Wright, *Practical Optimization*, Academic Press, 1981.
- [25] M.J.D. Powell, *Mathematical Programming: the State of the Art* (eds A Bachem et al) 288–311 (1983).
- [26] R. Fletcher, *Practical Methods of Optimization*, 2nd Edition, Wiley, 1987.
- [27] <https://www.nag.co.uk>.
- [28] J. Berntsen, T. Espelid, A. Genz, *ACM Trans. Math. Software* 17 (1991) 437–451.
- [29] J. Berntsen, T. Espelid, A. Genz, *ACM Trans. Math. Software* 17 (1991) 452–456.
- [30] A. Genz, A. Malik, *SIAM J. Numer. Anal.* 20 (1983) 580–588.
- [31] <http://www.feynarts.de/cuba/>.
- [32] O. Chuluunbaatar, I.V. Puzynin, S.I. Vinitzky, *J. Phys. B* 34 (2001) L425–L432.
- [33] A. Alijah, A.J.C. Varandas, *Phil. Trans. R. Soc. A* 364 (2006) 2889–2901.



**The 6th International Conference on
Optimization, Simulation and Control
(COSC2019)**

PROGRAM AND ABSTRACTS



**June 21-23, 2019
Ulaanbaatar, Mongolia**



NUMERICAL SOLUTION OF BURGERS' EQUATION BY LOCAL INTEGRO SPLINE

R.Mijiddorj¹, T.Zhanlav²

¹ Institute of Mathematics, National University of Mongolia, Mongolia

² Mongolian National University of Education, Mongolia

E-mail: mijiddorj@msue.edu.mn, tzhanlav@yahoo.com

Abstract: Burgers' equation is an important non-linear parabolic partial differential equation. Increasingly, this model is used in applications such as fluid dynamics, turbulence and others. In this talk, we examine some practical numerical methods to solve initial-boundary value PDE-s based on the local integro splines. Higher-order accurate hybrid schemes for the numerical solution is presented. The accuracy of the proposed schemes is demonstrated by some test problems.

MAXIMIZING THE SUM OF RADII OF BALLS INSCRIBED IN A POLYHEDRAL SET

R. Enkhbat and J. Davaadulam

National University of Mongolia, Mongolia

e-mail: jamsran_dd@yahoo.com

Abstract: The sphere packing problem is one of the most applicable areas in mathematics which finds numerous applications in science and technology. We consider a maximization problem of a sum of radii of non-overlapping balls inscribed in a polyhedral set in Hilbert space. This problem is often formulated as the sphere packing problem. We extend the problem in Hilbert space as an optimal control problem with the terminal functional and phase constraints for each moment. This problem belongs to a class of nonconvex optimal control problem and application of Pontryagin's maximum principle does not always guarantee finding a global solution to the problem. We show that the problem in a finite dimensional case for three balls(spheres) is connected to well-known Malfatti's problem [1]. Malfatti's generalized problem was examined in [2,3,4] as