ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

1 • 2014

УДК 681.142

В. М. Заяць

Два підходи до побудови оптимальних числових методів другого порядку та їх застосування до аналізу нелінійних систем коливної природи

(Представлено членом-кореспондентом НАН України В. В. Грициком)

Запропоновано ітераційний та прямий підходи до мінімізації похибки дискретизації числових методів другого порядку. Ітераційний підхід грунтується на модифікації методу трапецій і встановленні моменту часу, коли внески явного і неявного методів Ейлера мають однаковий внесок до поправки для наступної точки дискретизації динамічної системи. При комбінуванні отриманої формули з методом трапеції показано можливість побудови оптимального за точністю числового методу. Прямий підхід грунтується на встановленні моменту часу, коли дотичні, проведені до сусідніх точок дискретизації неперервної системи, перетинаються, що забезпечує нульову похибку дискретизації. Підтверджено доцільність їх застосування до аналізу нелінійних динамічних систем коливної природи з малим коефіцієнтом загасання, тривалими перехідними процесами та високою добротністю.

Для аналізу складних процесів і явищ у динамічних нелінійних системах коливної природи, що описуються системою неперервних диференційних рівнянь, поданих у нормальній формі Коші

$$\frac{d\mathbf{x}}{dt} = f[\mathbf{x}(t), t)],$$

де $\mathbf{x}$ — $\mathbf{N}$-мірний вектор змінних стану; $\mathbf{f}$ — $\mathbf{N}$-мірна нелінійна вектор-функція, яка описує динаміку фазових траєкторій системи, застосовують чисельні методи для проведення дискретизації. Такі методи повинні бути збіжними та мати малу похибку дискретизації для забезпечення збереження якісної та кількісної відповідності між досліджуваним процесом або явищем та його дискретною моделлю [1–5]. Друга вимога до різницевих методів — це властивість $\mathbf{A}$-стійкості [8–10]. У протилежному випадку наявність незначної локальної похибки обчислень, допущеної на одному кроці, може призвести до нагромадження цієї похибки в процесі руху зображуючої точки вздовж фазової траєкторії і цілковитої непридатності для прикладних застосувань остаточного результату обчислень [6, 8, 9]. Третя

вимога — простота реалізації алгоритму обчислень та мінімальні технічні та часові затрати для досягнення заданої точності.

У програмах комп'ютерного аналізу електронних схем [8], аналізі поведінки систем зі складною динамікою [3, 4], аналізі коливних систем з високою добротністю, для яких перехідні процеси є тривалими [6], виникає проблема між складністю різницевого алгоритму та його точністю. Як правило, використовують методи не вище другого порядку складності або їх комбінації. Зокрема, часто застосовується метод трапецій [6–8]. Різницева формула цього методу має вигляд:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{2}(\mathbf{f}_n + \mathbf{f}_{n+1}). \tag{1}$$

Ця формула є комбінацією двох методів: на першій половині кроку дискретизації застосовується явний метод Ейлера, а на другій — неявний метод Ейлера [8]. В результаті побудови такої комбінації, як засвідчують численні публікації, точність зростає більше, ніж на порядок порівняно з методами Ейлера. Крім того, для цього методу характерна властивість **A**-стійкості, що підтверджено розрахунком генераторних схем з високою добротністю та тривалими перехідними процесами. Однак похибка дискретизації змінних має від'ємний знак і істотно залежить від крутизни характеристики фазових траєкторій, що описують систему, і може перевищувати віддаль між двома точками дискретизації у кілька разів.

Спосіб мінімізації похибки дискретизації. У роботі [7] запропоновано враховувати поправки для наступної точки дискретизації не на середині кроку h, а в той момент часу, коли внески явного і неявного методів Ейлера є еквівалентними. З цією метою різницеву формулу (1) подано у формі, що запропонував Лінігер–Уілаббі:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h(1-\mu)\mathbf{f}_n + h\mu\mathbf{f}_{n+1}, \tag{2}$$

яка при $\mu = 0$ відповідає явному методу Ейлера; $\mu = 0{,}5$ — методу трапецій; $\mu = 1$ — неявному методу Ейлера. Прирівнявши другий і третій члени з правого боку у формулі (2), отримаємо значення параметра μ, при якому явний і неявний методи Ейлера вносять однаковий внесок у поправку до значення $\mathbf{x}_n$:

$$\mu = \frac{\mathbf{f}_n}{\mathbf{f}_n + \mathbf{f}_{n+1}}. \tag{3}$$

Після підстановки (3) в (2) отримано нову різницеву формулу:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{2h\mathbf{f}_n\mathbf{f}_{n+1}}{(\mathbf{f}_n + \mathbf{f}_{n+1})}. \tag{4}$$

Оскільки за побудовою формули (4) внесок кожного з методів Ейлера не перевищує половини віддалі між $\mathbf{x}_n$ і $\mathbf{x}_{n+1}$, то метод (4) дає гарантоване обмеження на величину похибки дискретизації на кожному кроці та забезпечує її додатність.

Геометричну ілюстрацію запропонованого способу зменшення похибки дискретизації проілюстровано на рис. 1. Якщо поправки за явним та неявним методами Ейлера до наступної точки дискретизації враховувати в момент часу, що відповідає точці C, як показано на рис. 1, то отримаємо метод трапеції; в точці B маємо пропонований метод, який зрівноважує внески методів Ейлера; в точці A отримується оптимальна комбінація, яка відповідає точці перетину дотичних до $\mathbf{x}_n$ та $\mathbf{x}_{n+1}$ точок дискретизації.

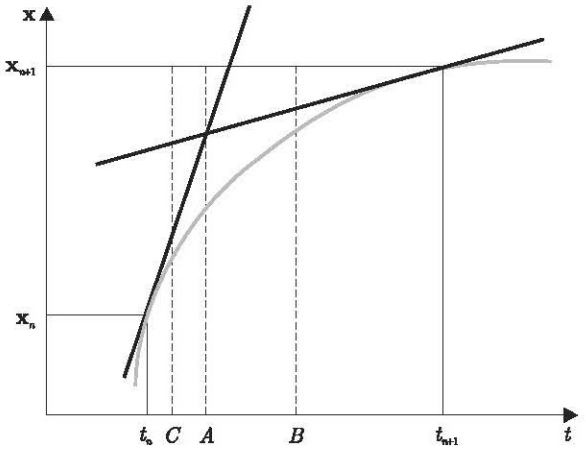

Рис. 1. Геометрична інтерпретація ітераційного та прямого підходів до мінімізації похибки дискретизації

Для оцінки похибки методу (4) проведено аналіз похибки дискретизації на прикладі моделі консервативної системи другого порядку

$$\frac{d^2\mathbf{x}}{dt^2} = -\omega_0^2\mathbf{x},$$

який підтвердив, що похибка дискретизації методу (4) пропорційна до $h^2/24$, як і в методі трапецій, але має протилежний знак і в два рази меншу абсолютну величину. Дослідження показали, що метод (4), як і метод трапецій, має властивість **A**-стійкості.

Ітераційний підхід до мінімізації похибки дискретизації. Враховуючи, що похибка методу (4) і методу трапецій (формула (1)) мають протилежні знаки, можна провести їх арифметичне усереднення, тим самим зменшити величину похибки. Застосовуючи на першій половині кроку формулу (4), а на другій — формулу (1), отримуємо різницеву формулу

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h\mathbf{f}_n\mathbf{f}_{n+1}}{(\mathbf{f}_n + \mathbf{f}_{n+1})} + \frac{h}{4}(\mathbf{f}_n + \mathbf{f}_{n+1}), \tag{5}$$

яку назвемо різницевою комбінацією першого роду (К1Р). Похибка дискретизації при використанні (5) до консервативної системи виявилася у два рази меншою, порівняно з методом (4), і протилежною за знаком відносно методу трапеції. Тепер після усереднення (1) і (5) отримуємо різницеву комбінацію другого роду (К2Р):

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h\mathbf{f}_n\mathbf{f}_{n+1}}{2(\mathbf{f}_n + \mathbf{f}_{n+1})} + \frac{3h}{8}(\mathbf{f}_n + \mathbf{f}_{n+1}). \tag{6}$$

Як засвідчили результати аналізу похибки дискретизації методу (6) при розгляді моделі без втрат, вона виявилася у чотири рази меншою за похибку методу трапецій і в два рази меншою, ніж похибка методу (5). При цьому знак похибки в К2Р збігається зі знаком похибки у методі трапецій і протилежний до похибки, який дає К1Р. Таким чином, можна очікувати подальшого зменшення величини похибки дискретизації комбінації методів (5) і (6), яка приводить до різницевої комбінації третього роду (К3Р):

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{3h\mathbf{f}_n\mathbf{f}_{n+1}}{4(\mathbf{f}_n + \mathbf{f}_{n+1})} + \frac{5h}{16}(\mathbf{f}_n + \mathbf{f}_{n+1}). \tag{7}$$

Зауважимо, що розглядати комбінацію (6) з (4) недоцільно (хоча вона й має право на існування), оскільки (5) має в чотири рази меншу похибку дискретизації порівняно з (4). Крім того, знаки похибки в (4) і (6) збігаються.

Оптимальна комбінація для мінімізації похибки дискретизації. Запропоновані комбінації різницевих схем побудовано таким чином, що в комбінаціях непарного роду (К1Р, К3Р) більш значним є внесок другого члена в отриманих формулах, порівняно з третім, а в комбінаціях парного роду (К2Р) ці внески практично вирівнюються. Така побудова забезпечує зміну знака похибки при отриманні нової комбінації. Отже, можна сконструювати метод другого порядку, який забезпечить з точністю до членів другого порядку малості як завгодно малу похибку дискретизації. Після арифметичного усереднення (6) і (7) приходимо до різницевої схеми четвертого роду (К4Р):

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{5h\mathbf{f}_n\mathbf{f}_{n+1}}{8(\mathbf{f}_n + \mathbf{f}_{n+1})} + \frac{11h}{32}(\mathbf{f}_n + \mathbf{f}_{n+1}). \tag{8}$$

Аналізуючи формули (5)–(8), на k-му кроці, застосовуючи півкроку парну комбінацію, а півкроку — непарну, отримуємо різницеву схему для комбінації k-го роду (ККР):

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{a_k h\mathbf{f}_n\mathbf{f}_{n+1}}{(\mathbf{f}_n + \mathbf{f}_{n+1})} + a_{k+1}h(\mathbf{f}_n + \mathbf{f}_{n+1}), \tag{9}$$

де

$$a_k = \frac{2^k - (-1)^k}{3 \cdot 2^{k-1}}; \qquad a_{k+1} = \frac{2^{k+1} + (-1)^k}{3 \cdot 2^{k+1}}.$$

Очевидно, з ростом k величини коефіцієнтів a_k і a_{k+1} зменшуються, що приводить до зменшення похибки дискретизації. При цьому похибка дискритизації будь-якої k-ї комбінації може бути обчислена за формулою

$$\delta = \frac{(-1)^k}{2^{k+1}}, \tag{10}$$

що засвідчує аналіз консервативних систем другого порядку та систем з високою добротністю високих порядків.

З метою мінімізації похибки дискретизації в (9) здійснимо граничний перехід, спрямувавши k до безмежності. Отримуємо різницеву схему (11), для якої з точністю до членів другого порядку малості похибка дискретизації відсутня:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{2h\mathbf{f}_n\mathbf{f}_{n+1}}{3(\mathbf{f}_n + \mathbf{f}_{n+1})} + \frac{1}{3}h(\mathbf{f}_n + \mathbf{f}_{n+1}). \tag{11}$$

Висновок про відсутність похибки дискретизації різницевої схеми (11) випливає з формули (10), якщо в ній спрямувати k до безмежності. Недолік методу (11) в тому, що він потребує виконання в два рази більшої кількості арифметичних операцій порівняно з (4) і більше, ніж в три рази порівняно з (1).

Безпосередній пошук оптимальної комбінації. Щоб безпосередньо одержати аналітичний вираз, для якого похибка дискретизації в першому наближені відсутня, знайдемо координати точки A (рис. 1), що відповідають перетину дотичних

$$\mathbf{x} = \mathbf{x}_0^n + \mathbf{f}_n t \qquad \text{і} \qquad \mathbf{x} = \mathbf{x}_0^{n+1} + \mathbf{f}_{n+1} t,$$

проведених в двох сусідніх n і $n+1$ точках дискретизації. Прирівнявши праві частини в останніх двох рівняннях, отримуємо

$$t = \frac{\mathbf{x}_0^{n+1} - \mathbf{x}_0^n}{\mathbf{f}_n - \mathbf{f}_{n+1}}, \tag{12}$$

де

$$\mathbf{x}_0^n = \mathbf{x}_n - nh\mathbf{f}_n \qquad \text{і} \qquad \mathbf{x}_0^{n+1} = \mathbf{x}_{n+1} - (n+1)h\mathbf{f}_{n+1},$$

що видно з рис. 1. З іншого боку, моменту перетину дотичних відповідає значення

$$t = h(n + \mu). \tag{13}$$

Прирівнявши праві частини рівнянь (12) і (13), знаходимо значення параметра μ, при виборі якого внески явного і неявного методу Ейлера забезпечують потрапляння фазової точки з n в $n+1$ точку дискретизації:

$$\mu = \frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{(\mathbf{f}_n - \mathbf{f}_{n+1})} + \frac{h\mathbf{f}_{n+1}}{\mathbf{f}_n - \mathbf{f}_{n+1}}. \tag{14}$$

Підставивши значення μ з (14) у формулу (2), отримуємо оптимальну комбінацію числового методу другого порядку, для якої похибка дискретизації відсутня:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{2}(\mathbf{f}_n + h\mathbf{f}_{n+1}). \tag{15}$$

За алгоритмічною складністю метод (15) простіший за (4) і незначно поступається методу (1), забезпечуючи при цьому мінімальну похибку обчислень, пов'язану лише з точністю подання чисел у середовищі обчислень.

Зазначимо, що всі одержані різницеві формули (5)–(9), (11), (15) для дискретизації неперервних систем мають властивість **A**-стійкості, що унеможливлює нагромадження похибки дискретизації при тривалих перехідних процесах, які характерні для динамічних систем з високою добротністю. Цей результат підтверджено розрахунком кварцових генераторних пристроїв та високодобротних генераторних схем з тривалими перехідними процесами [6].

1. *Бондаренко В. М., Герасымив И. И., Мандзий Б. А., Маранов А. В.* Анализ точности и качественного соответствия дискретных моделей электрических цепей. – Киев, 1983. – 44 с. (Препринт НАН Украины. Ин-т электродинамики, № 307).
2. *Бутенин Н. В., Неймарк Ю. И., Фуфаев Н. А.* Введение в теорию нелинейных колебаний. – Москва: Наука, 1987. – 384 с.
3. *Васильев В. И., Шевченко А. И.* Комбинированный алгоритм оптимальной сложности // Праці Міжнар. конф. "Штучний інтелект". – Т. 1. – Крим, 2002. – С. 308–310.
4. *Заяць В. М.* Построение и анализ дискретной модели дискретной колебательной системы // Кибернетика и системный анализ. – 2000. – № 4. – С. 161–165.
5. *Заяць В. М.* Аналіз динаміки та умов стійкості дискретних моделей коливних систем // Вісн. НУ "Львівська політехніка". Інформаційні системи та мережі. – 2004. – № 519. – С. 132–142.
6. *Заяц В. М.* Ускоренный поиск установившихся режимов в высокочастотных автогенераторах с длительными переходными процессами // Изв. вузов. Радиоэлектроника. – 1993. – № 3. – С. 26–32.
7. *Заяць В. М.* Побудова комбінованих різницевих методів другого порядку // Зб. праць наук. техн. конф. "Обчислювальні методи і системи перетворення інформації". – Львів, 7–8 жовтня 2011. – ФМІ НАНУ. – 2011. – С. 34–36.

8. *Петренко А. І.* Числові методи в інформатиці. – Київ: В-во ВНУ, 1999. – 450 с.
9. *Самойленко А. М., Ронто Н. И.* Численно-аналитические методы исследования периодических решений. – Київ: Вища шк., 1976. – 180 с.
10. *Чуа Л. О., Лин П.-М.* Машинный анализ электронных схем (алгоритмы и вычислительные методы). – Москва: Энергия, 1980. – 640 с.

НУ "Львівська політехніка" *Надійшло до редакції 27.05.2013*

В. М. Заяць

Два подхода к построению оптимальных численных методов второго порядка и их применение к анализу нелинейных систем колебательной природы

Предложены итерационный и прямой подходы к минимизации погрешности дискретизации численных методов второго порядка. Итерационный подход основан на модификации метода трапеций и установлении момента времени, когда явный и неявный методы Эйлера имеют одинаковый вклад в поправки для следующей точки дискретизации динамической системы. При комбинировании полученной формулы с методом трапеции показана возможность построения оптимального по точности численного метода. Прямой подход основывается на установленнии момента времени, когда касательные, проведенные в соседние точки дискретизации непрерывной системы, пересекаются, что обеспечивает нулевую погрешность дискретизации. Подтверждена целесообразность их применения к анализу нелинейных динамических систем колеблющейся природы с малым коэффициентом затухания, длительными переходными процессами и высокой добротностью.

V. M. Zayats

Two approaches to the construction of optimal second-order numerical methods and their application to the analysis of oscillatory nonlinear systems

Iterative and direct approaches to the minimization of errors at a discretization of second-order numerical methods are proposed. The iterative approach is based on a modification of the method of trapezoids and setting the time when the explicit and implicit Euler methods give the same contribution to the amendment to the next discretization point of a dynamical system. Combining the derived formula with the method of trapezoids, the possibility of constructing the optimal precision numerical method is shown. The direct approach is based on determining a time when the tangents drawn to the nearby points of discretization of the continuous system intersect, which provides the zero error of a discretization. The expediency of their application to the analysis of nonlinear dynamical oscillatory systems with a low coefficient of attenuation, long transients, and high power is confirmed.