UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
WYDZIAŁ FIZYKI

R O Z P R A W A D O K T O R S K A

Dyskretyzacja niektórych modeli fizycznych: od podejścia standardowego do dyskretyzacji geometrycznej

mgr Bogusław Ratkiewicz

Praca wykonana pod kierunkiem
dra hab. Jana L. Cieślińskiego, prof. UwB

Poznań 2010

Pragnę podziękować dr hab. Janowi Cieslińskiemu
za cenne wskazówki, współpracę i pomoc w trakcie pisania pracy.
Dziękuję także prof. dr hab. Henrykowi Szydłowskiemu
i prof. dr hab. Wojciechowi Nawrocikowi za zachętę i życzliwość
okazaną mi w czasie trwania studiów doktoranckich.

Spis treści

1 Wstęp

Przedmiotem badań podjętych w tej pracy jest problem dyskretyzacji niektórych modeli fizycznych, które na ogół są jednowymiarowymi układami hamiltonowskimi. Konstruowanie schematów dyskretnych symulujących modele ciągłe tego typu jest, siłą rzeczy, blisko związane z numerycznym całkowaniem odpowiednich równań różniczkowych zwyczajnych. W niniejszej pracy koncentrujemy się na dyskretyzacji geometrycznej, czyli na takich równaniach różnicowych, których własności w możliwie dużym stopniu przypominają własności odpowiedniego równania różniczkowego.

Pierwotna motywacja podjęcia tej tematyki wiązała się z kwestiami dydaktycznymi. Pojęcia różniczkowe, takie jak prędkość czy przyspieszenie, definiuje się i wyjaśnia przy pomocy odpowiedników dyskretnych, takich jak iloraz różnicowy, które są bardziej naturalne i łatwiejsze do zrozumienia. Co więcej, w literaturze fizycznej jest stale obecny nurt zajmujący się numerycznym podejściem do całkowania równań różniczkowych, przy czym obok prac zaawansowanych, jak na przykład [42, 67, 79, 94], publikowane są prace stosunkowo elementarne, jak [32, 51, 59, 100]. Przykładem takiej elementarnej pracy jest także pierwsza publikacja związana z tematyką tej rozprawy [20], której wyniki zawarte są w rozdziale 3. Jednak kontynuacja naszych badań doprowadziła do uzyskania tak wielu nowych wyników, że kierunek prac uległ zmianie (co znalazło odbicie w zaproponowanym tytule i zakresie rozprawy), a ślad motywacji dydaktycznej jest ledwie widoczny i to tylko w niektórych rozdziałach (np. rozdziały 3 i 10).

Początkowa część pracy (a zwłaszcza rozdziały 3 i 4) poświęcona jest standardowym metodom numerycznym, z naciskiem na metody dobrze odtwarzające jakościowe i geometryczne cechy badanych równań. Metody te zostały porównane pod kątem jakościowego i ilościowego zachowania się dla długich i bardzo krótkich przebiegów czasowych. Centralną część pracy stanowi rozwinięcie i udoskonalenie metody dyskretnego gradientu (rozdziały 5, 6, 7), która jest jedną z ważniejszych metod geometrycznego całkowania numerycznego [41, 64, 66]. Trzecim wątkiem, obecnym w wielu miejscach niniejszego opracowania, są dyskretyzacje dokładne wybranych modeli fizycznych, w szczególności: klasycznego oscylatora harmonicznego i anharmonicznego (rozdziały 3 i 8) czy problemu Keplera (rozdział 9).

Całkowanie geometryczne polega na poszukiwaniu takich metod numerycznych, które ściśle zachowują pewne fizyczne lub matematyczne własności równań. Chodzi tu o takie ich cechy jak całki pierwsze, symetrie,

objętość przestrzeni fazowej czy struktura symplektyczna [64]. Zastosowania tych metod w fizyce są najrozmaitsze i rozciągają się od akceleratorów cząstek [30, 94], przez dynamikę molekularną [60, 78], mechanikę kwantową [111], mechanikę nieba [58, 105, 112, 114] aż do analizy układów złożonych [1] i układów z wieloma skalami czasowymi [50].

Jest wiele powodów szybkiego rozwoju całkowania geometrycznego, które pojawiło się praktycznie we wszystkich gałęziach analizy numerycznej. Poszukuje się metod szybszych, prostszych, bardziej stabilnych i dokładniejszych od schematów tradycyjnych. Szuka się metod takich, które dają lepsze zachowanie jakościowe, nawet jeśli nie udało się z ich pomocą zmniejszyć błędów numerycznych. Oczekuje się, że schematy zachowujące strukturę równań mogą pozwolić na przeprowadzenie obliczeń wcześniej uważanych za niemożliwe (np. całkowanie układów hamiltonowskich w bardzo długim okresie czasu) [64].

Tradycyjne podejście do numerycznego rozwiązywania równań różniczkowych polega na obliczaniu rozwiązania przy zadanych warunkach początkowych w określonej chwili czasu i z założonym błędem tak efektywnie, jak to tylko jest możliwe. Tak sformułowane zadanie warunkuje wybór rodzaju metody, jej rzędu i wielkości kroku czasowego. W przeciwieństwie do powyższego, używanie integratorów zachowujących strukturę równań często polega na wybraniu dość dużego kroku czasowego i obliczaniu orbit w bardzo długim okresie czasu. Typową sytuacją w przypadku metod geometrycznych jest dość duży błąd globalny otrzymywanych orbit, przy jednoczesnym zachowaniu prawidłowego zachowania jakościowego. Obraz trajektorii układu w przestrzeni fazowej może być bliski obrazowi właściwemu dla równania wyjściowego, co daje wiarygodną informację o ewolucji układu. Możemy się czasem spotkać z zaskakującym paradoksem: globalny błąd niektórych metod geometrycznych może... rosnąć wraz ze zmniejszaniem się kroku czasowego. Paradoks ten występuje przede wszystkim w przypadku dyskretyzacji dokładnych. Jedynym źródłem błędu są tam zaokrąglenia dokonywane w każdym kroku obliczeniowym. Im mniej tych kroków, tym mniejszy błąd.

Standardowe schematy numeryczne stały się punktem wyjścia do wielu ulepszeń i rozwinięcia nowych metod od czasu odkrycia ich właściwości geometrycznych. Wymienimy tu trzy popularne schematy geometryczne oraz niektóre ich cechy będące przedmiotem również naszego zainteresowania.

Metoda Störmera-Verleta, zwana też metodą żabiego skoku (*leap-frog*) [48], zastosowana do równań Newtona $\dot{x} = p, \quad \dot{p} = -\nabla V(x)$, ma postać

$$
\begin{aligned}
x_{k+1/2} &= x_k + \frac{1}{2} h p_k, \\
p_{k+1} &= p_k - h \nabla V(x_{k+1/2}), \\
x_{k+1} &= x_{k+1/2} + \frac{1}{2} h p_{k+1},
\end{aligned}
\tag{1.1}
$$

gdzie h oznacza krok czasowy, zaś x, p mogą być wektorami. Metoda ta jest tak naturalna, że była odkrywana co najmniej kilka razy, a doszukać się jej można już w pracach samego Newtona [41]. Jeden krok tej metody może być zinterpretowany jako ruch jednostajnie przyspieszony, przy czym siła wywołująca to przyspieszenie jest równa rzeczywistej sile obliczonej w środkowym momencie trwania tego ruchu

Leap-frog jest metodą otwartą (jawną) drugiego rzędu, wymagającą tylko jednego obliczenia wartości siły w każdym kroku czasowym. Metoda ta jest symplektyczna i odwracalna w czasie. Nie zachowuje wprawdzie energii, ale błąd energii nie rośnie w czasie. Zachowuje periodyczność orbit w przestrzeni fazowej (dla dostatecznie małych kroków czasowych). Metoda ta zachowuje również pęd i moment pędu. Niestety staje się niestabilna, jeśli krok czasowy jest zbyt duży. Metoda ta okazała się zaskakująco dobra, jak na swą prostotę, co w głównej mierze jest efektem jej symplektyczności i odwracalności w czasie.

Drugim przykładem jest metoda punktu środkowego (*implicit midpoint rule*), która dla układów pierwszego rzędu postaci $\dot{x} = f(x)$, $x \in \tilde{\mathrm{N}}^n$ jest dana przez

$$
x_{k+1} = x_k + hf\left(\frac{x_k + x_{k+1}}{2}\right). \tag{1.2}
$$

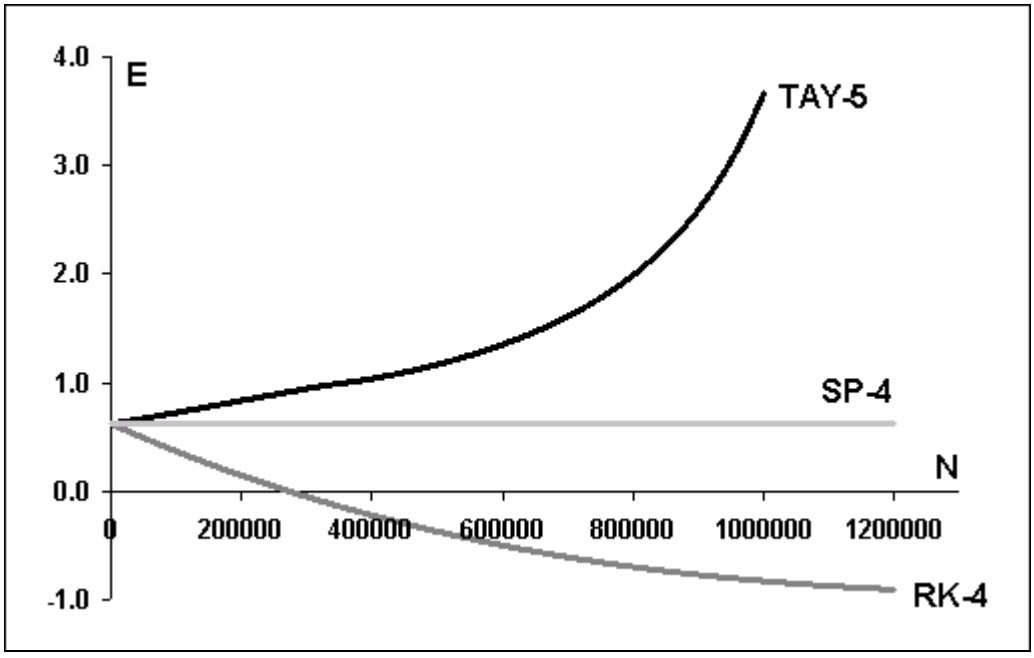
Metoda ta jest zamknięta (niejawna), czyli bardziej kosztowna numerycznie (w każdym kroku obliczeniowym trzeba rozwiązać algebraiczne równanie w postaci uwikłanej metodą iteracyjną), ale (podobnie jak *leap-frog*) jest symplektyczna i odwracalna w czasie. Nie zachowuje wprawdzie pędu i momentu pędu (o ile wielkości te mają zastosowanie do badanego układu), ale w zamian zachowuje dowolne kwadratowe całki pierwsze układu. Jest też liniowo stabilna dla wszystkich kroków czasowych. Należy jednak pamiętać o tym, że w praktyce żadna metoda, oprócz dyskretyzacji dokładnych, nie może zachowywać energii i być jednocześnie symplektyczna [34, 41].

Trzeci przykład, to metoda dyskretnego gradientu, zwana też zmodyfikowaną metodą punktu środkowego (*modified midpoint rule*) [49, 54, 98]. W przypadku jednowymiarowych równań Newtona ta niejawna metoda jest zadana równaniami

$$\frac{x_{n+1}-x_n}{h}=\frac{1}{2}(p_{n+1}+p_n),$$
$$\frac{p_{n+1}-p_n}{h}=-\frac{V(x_{n+1})-V(x_n)}{x_{n+1}-x_n} \qquad (1.3)$$

Bardzo ważną cecha tej metody jest dokładne zachowanie całki energii. Zasada zachowania energii może być łatwo wykazana poprzez pomnożenie stronami obu równań (tak, aby przyrost zmiennej *x* uległ skróceniu). Schemat ten nie jest symplektyczny.

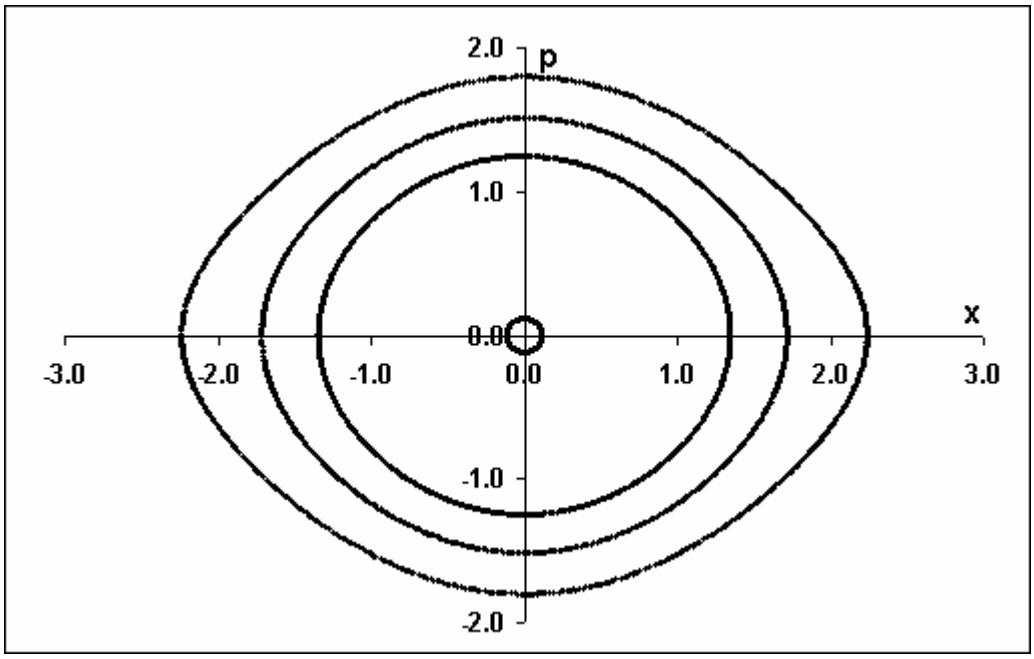
Przewagę geometrycznych metod numerycznych nad tradycyjnymi metodami, nawet wysokiego rzędu, zilustrujemy na przykładzie wahadła matematycznego, którego potencjał zadany jest wzorem $V(x)=-\cos(x)$. Wykres 1.1 pokazuje działanie trzech dyskretyzacji tego układu: schematu Taylora 5-ego rzędu (TAY-5), metody Rungego-Kutty rzędu 4-ego (RK-4) i metody symplektycznej 4-ego rzędu (SP-4) (szczegóły w rozdziale 7).



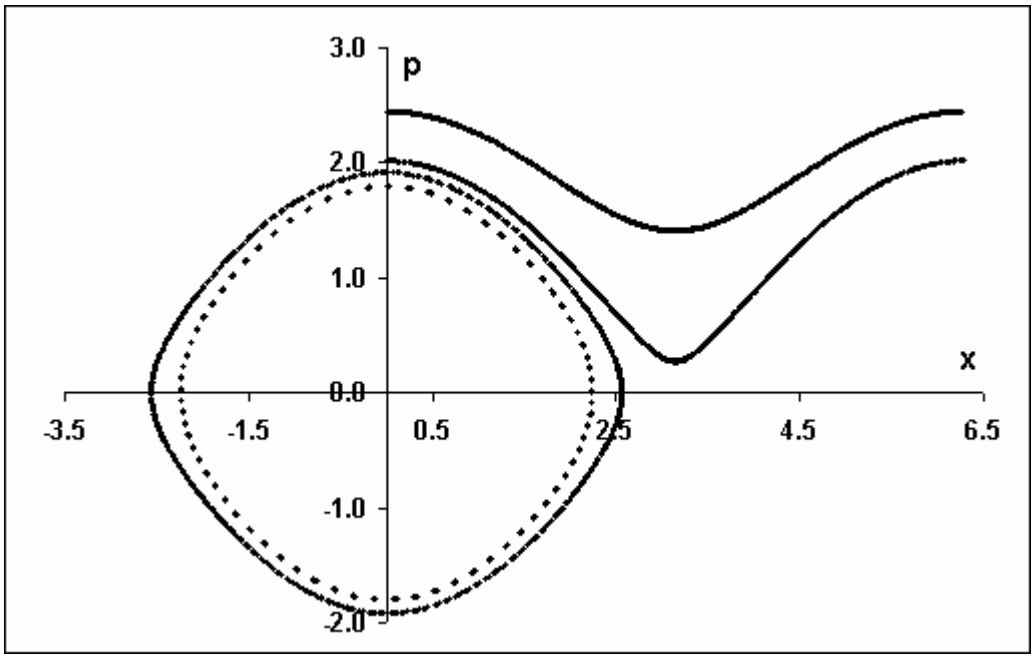
Wykres 1.1. Energia jako funkcja czasu (*t* = *Nh*) dla trzech dyskretyzacji wahadła matematycznego, p_0 = 1.8, *h* = 0.25, E_{ex} = 0.62.

Funkcjonowanie modeli RK-4 i TAY-5 wyjaśniają wykresy 1.2 i 1.3 przedstawiające ewolucję czasową generowanych przez nie krzywych fazowych (na każdym z tych rysunków mamy fragmenty jednej tylko dyskretnej krzywej fazowej). W przypadku wykresu 1.2 jest to krzywa spiralna, o malejącym promieniu, dążącym do zera. Widzimy więc, że metoda RK-4 modeluje raczej wahadło tłumione. Niewielkie „numeryczne" tłumienie zmniejsza

systematycznie energię drgań i ostatecznie układ asymptotycznie dąży do stanu spoczynku w punkcie równowagi trwałej.



Wykres 1.2. Ewolucja czasowa krzywej fazowej metody RK-4. Od zewnątrz: $t \in \{0; 5\cdot10^4; 10^5; 5\cdot10^5\}$, $p_0 = 1.8$, $h = 0.25$.



Wykres 1.3. Ewolucja czasowa krzywej fazowej metody TAY-5. Od wewnątrz: $t \in \{0; 5\cdot10^4; 10^5; 2\cdot10^5\}$, $p_0 = 1.8$, $h = 0.25$ (położenia w ruchu rotacyjnym modulo 2π).

Z kolei metoda TAY-5 systematycznie (i coraz szybciej) dodaje układowi energii, co kończy się przejściem od ruchu oscylacyjnego ($p_0 < 2$) do rotacyjnego, o rosnącej prędkości obiegu. Krzywa fazowa jest dyskretną spiralą

o wzrastającym promieniu. Obie metody produkują więc rozwiązania numeryczne o złych cechach jakościowych. W przeciwieństwie do powyższych wykresów obraz krzywych fazowych metody SP4 pozostaje niezmienny (i bardzo podobny do obrazu rzeczywistego) w bardzo długim okresie czasu. Podobną stabilność mają też inne metody symplektyczne, a także metody zachowujące energię, w tym wszystkie trzy metody wymienione wyżej, czyli (1.1), (1.2), (1.3).

Badania w dziedzinie całkowania geometrycznego koncentrują się na następujących kierunkach [64]:

1) poszukiwania nowych typów schematów numerycznych i schematów zachowujących nowe struktury równań,
2) poprawiania wydajności i dokładności schematów numerycznych poprzez znajdowanie metod wyższych rzędów, o mniejszych błędach lokalnych lub dopuszczających większy krok czasowy,
3) wyszukiwanie metod dostosowanych do wybranych szczególnych klas układów równań,
4) badanie zachowania różnych dyskretyzacji w długim okresie czasu oraz stopnia, w jakim zachowują obraz fazowy równań wyjściowych.

Praca niniejsza wpisuje się we wszystkie wymienione tu kierunki badań.

1) Zupełnie nowym, jak się wydaje, typem schematów numerycznych rozważanych w tej pracy są schematy „lokalnie dokładne”, zaproponowane przez promotora, a po raz pierwszy badane i testowane w niniejszej pracy (rozdz. 5, 6) i związanych z nią publikacjach.
2) Udało się znaleźć dość prosty sposób na dowolne zwiększanie rzędu schematu dyskretnego gradientu (rozdz. 7). Metody lokalnie dokładne także mają nieco wyższy rząd i dopuszczają duże kroki czasowe.
3) Znalezione zostały nowe dyskretyzacje dokładne, w tym dyskretyzacja dokładna dla oscylatora anharmonicznego (rozdz. 8), nowa dyskretyzacja problemu Keplera zachowująca trajektorie i całki ruchu (rozdz. 9) oraz nowa dyskretyzacja równania falowego z dokładną ewolucją czasową (rozdz. 10).
4) Szczegółowo przebadane zostało zachowanie się różnych dyskretyzacji w bardzo długich przedziałach czasu, oraz zachowanie różnych cech jakościowych (praktycznie wszystkie rozdziały, ale zwłaszcza rozdz. 4).

Struktura pracy, częściowo zasugerowana przez tytuł rozprawy, jest w dużej mierze chronologiczna. Rozdziały 3, 4, 5 oparte są na opublikowanych pracach:

1. J.L.Cieśliński, B.Ratkiewicz: "On simulations of the classical harmonic oscillator equation by difference equations", *Advances in Difference Equations* **2006** (2006) 40171 (17 pp).
2. J.L.Cieśliński, B.Ratkiewicz: "Long-time behaviour of discretizations of the simple pendulum equation", *Journal of Physics A: Mathematical and Theoretical* **42** (2009) 105204 (29 pp).
3. J.L.Cieśliński, B.Ratkiewicz: "Improving the accuracy of the discrete gradient method in the one-dimensional case", *Physical Review E* **81** (2010) 016704 (6pp).

Wyniki rozdziału 7 znalazły się w wysłanym do publikacji preprincie:

4. J.L.Cieśliński, B.Ratkiewicz: "Discrete gradient algorithms of high order for one-dimensional systems", *preprint arXiv:* 1008.3895 [physics.comp-ph].

Kolejny preprint jest przeglądem ważniejszych wyników i rozszerzeniem niektórych wątków z rozdziałów 5, 6 i 7:

5. J.L.Cieśliński, B.Ratkiewicz: "Energy-preserving numerical schemes of high accuracy for one-dimensional Hamiltonian systems", *preprint arXiv:* 1009.2738 [cs.NA].

Zawarty tam materiał był tematem mojego wystąpienia na konferencji „BIT 50. Trends in Numerical Mathematics" (Lund, 17-20.06.2010).

Razem z promotorem planujemy napisanie i publikację kolejnych prac, zawierających rezultaty rozdziałów 6, 8, 9 i 10.

Wszystkie omawiane w tej pracy schematy zostały przetestowane numerycznie. Wyniki tych eksperymentów numerycznych, będące ważnym elementem tej pracy, zostały omówione w końcowej części poszczególnych rozdziałów. Liczne szczegóły dotyczące praktycznych kwestii numerycznych umieszczone zostały w dodatkowym rozdziale 11.3.

Końcowy etap badań związanych z napisaniem tej rozprawy doktorskiej (15.09.2009-15.12.2010) dostał wsparcie finansowe z Ministerstwa Nauki i Szkolnictwa Wyższego w ramach grantu promotorskiego Nr N N202 238637.

2 Najważniejsze wyniki badań

W rozdziale tym przedstawiono w skrócie najważniejsze rezultaty uzyskane w trakcie prac badawczych. Pozwoli to czytelnikowi na szybkie ich ogarnięcie i ułatwi nawigację wśród pozostałych rozdziałów pracy zawierających szczegóły teoretyczne i eksperymentalne, które mogą czasem zaciemniać istotę rzeczy.

Wszelkie badania naukowe mają swoją chronologię, a różne koncepcje mają źródło we wcześniejszych wynikach. Często początkowe prace nie pozwalają od razu dostrzec przyszłych implikacji, ale zawarte w nich wyniki ważą na późniejszych rezultatach.

Dobrym przykładem jest dyskretyzacja dokładna równania oscylatora harmonicznego wyprowadzona w rozdziale 3. Rozpatrywano w nim równanie

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^{\,2} x = 0\,, \tag{2.1}$$

którego rozwiązanie jest dobrze znane i poszukiwano schematu numerycznego odtwarzającego to rozwiązanie dokładnie na dyskretnej siatce punktów. Schemat taki udało się znaleźć w postaci równania

$$x_{n+2} - 2e^{-\varepsilon\gamma}\cos(h\omega)x_{n+1} + e^{-2\varepsilon\gamma}x_n = 0 \tag{2.2}$$

gdzie h oznacza krok czasowy, a $\omega := \sqrt{\omega_0^{\,2} - \gamma^2}$. Przedstawiono w ten sposób konkretny przykład ogólniejszej relacji pomiędzy równaniami różniczkowymi i różnicowymi polegającej na tym, że każdemu liniowemu równaniu różniczkowemu ze stałymi współczynnikami odpowiada równanie różnicowe, które jest jego dyskretyzacją dokładną (oznacza to, że $x_n = x(t_n) = x(nh)$). Chociaż sam ten pomysł nie jest nowy [1, 83], jednak związane z nim wyniki zostały włączone do rozprawy (rozdział 3, napisany na bazie artykułu [20]), gdyż okazały się być bardzo użyteczne w dalszych badaniach.

Jednym z celów tej pracy było szczegółowe przetestowanie i porównanie ze sobą rozmaitych schematów numerycznych. Wyniki związane z tym tematem znajdują się w rozdziale 4, a modelem na którym zostały przeprowadzone eksperymenty numeryczne było tam wahadło matematyczne. Do testów (również w bardzo długich okresach czasu) wybrano szerokie spektrum metod od symplektycznych poczynając (*leap-frog*, *implicit midpoint*) poprzez metody rzutowane (*leap-frog* rzutowany na powierzchnię stałej energii), metodę Rungego-Kutty, kończąc na najprostszych wersjach metod gradientowych.

Wyniki testów numerycznych (rozdział 4, rozdział 11.3, a także artykuł [21]) są zbyt obszerne, aby je tu omawiać. Najważniejsze jest to, że przy okazji tych testów po raz pierwszy pojawił się pomysł linearyzacji równań standardowego

dyskretnego gradientu wokół punktu równowagi trwałej $\varphi = 0$, w którym potencjał wahadła, $V(\varphi) = -\cos\varphi$, posiada lokalne minimum. Można wówczas przybliżyć małe oscylacje wokół punktu równowagi przez równanie opisywanego wyżej klasycznego oscylatora harmonicznego. Pozwoliło to na zaproponowanie zmodyfikowanej metody dyskretnego gradientu w postaci

$$\begin{cases} \dfrac{\varphi_{n+1} - 2\varphi_n + \varphi_{n-1}}{\delta^2} = -\dfrac{1}{2}\left(\dfrac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n} + \dfrac{V(\varphi_n) - V(\varphi_{n-1})}{\varphi_n - \varphi_{n-1}} \right), \\ p_n = \dfrac{\varphi_{n+1} - \varphi_n}{\delta} + \dfrac{1}{2}\delta\left(\dfrac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n} \right), \end{cases} \tag{2.3}$$

gdzie $\delta = \dfrac{2}{\omega_0}\tan\left(\dfrac{\varepsilon\omega_0}{2}\right)$, natomiast $\omega_0 = \sqrt{V''(0)}$. Wzory powyższe można stosować do dowolnego potencjału mającego położenie równowagi trwałej w punkcie $\varphi = 0$. Dyskretyzacja ta, oznaczana skrótem MOD-GR, znakomicie spisuje się w przypadku małych drgań wokół położenia równowagi, dając względny błąd okresu o przynajmniej 4 rzędy wielkości mniejszy niż którakolwiek z pozostałych testowanych metod.

Koncepcja wykorzystana w dyskretyzacji (2.3) została rozwinięta w rozdziale 5 (opartym na artykule [22]). Zamiast ograniczać się do linearyzacji problemu nieliniowego wokół położenia równowagi trwałej, dokonujemy linearyzacji wokół innego wybranego punktu. W tym przypadku ewolucja czasowa wkrótce wyrzuca nas poza otoczenie tego punktu, zatem chcąc kontynuować musimy w kolejnych krokach zmieniać punkt, wokół którego linearyzujemy. Zatem parametr δ nie będzie już stały, ale zmienny (co czasem podkreślamy dodatkowym indeksem, pisząc δ_n). W rozdziale 5 rozpatrujemy więc układ równań postaci

$$\dot{p} = -V'(x), \quad \dot{x} = p, \tag{2.4}$$

dla którego zaproponowany został następujący schemat numeryczny zwany *lokalnie dokładną* modyfikacją metody dyskretnego gradientu, czyli GR-LEX:

$$\begin{aligned} &\frac{x_{n+1} - x_n}{\delta_n} = \frac{1}{2}(p_{n+1} + p_n), \\ &\frac{p_{n+1} - p_n}{\delta_n} = -\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n}, \end{aligned} \tag{2.5}$$

gdzie δ_n jest funkcją zdefiniowaną wzorami

$$\begin{aligned} &\delta_n = \frac{2}{\omega_n}\tan\frac{\varepsilon\omega_n}{2}, \text{ (jeżeli } V''(x_n) > 0), \\ &\delta_n = \varepsilon, \text{ (jeżeli } V''(x_n) = 0), \end{aligned} \tag{2.6}$$

$$\delta_n = \frac{2}{\omega_n}\tanh\frac{\varepsilon\omega_n}{2}, \text{ (jeżeli } V''(x_n) < 0\text{)},$$

przy czym ε oznacza krok czasowy, oraz $\omega_n = \sqrt{|V''(x_n)|}$. Postać funkcji δ_n wyznaczono żądając, aby zmodyfikowany schemat (2.5) był *lokalnie dokładny*. Żądanie to oznacza, że zlinearyzowana wokół punktu x_n dyskretyzacja (2.5) jest zgodna z *dokładną dyskretyzacją* zlinearyzowanego wokół punktu x_n układu równań (2.4). Wybierając zamiast punktu x_n środek odcinka między x_n a x_{n+1} otrzymaliśmy symetryczną modyfikację tej metody, zwaną w skrócie GR-SLEX. Testy numeryczne przeprowadzone na przykładzie wahadła matematycznego i potencjału Morse'a pokazały, że lokalnie dokładna metoda dyskretnego gradientu (GR-LEX) i jej symetryczna modyfikacja (GR-SLEX) przewyższają dokładnością standardową metodę dyskretnego gradientu o kilka rzędów wielkości nie tracąc przy tym stabilności i doskonałego jakościowego zachowania w długim okresie czasu. Warto dodać, że wszystkie omawiane metody gradientowe zachowują dokładnie trajektorie modelowanego układu w przestrzeni fazowej.

Omówione powyżej schematy lokalnie dokładne wykorzystano do skonstruowania klas dyskretyzacji zachowujących trajektorie 2-wymiarowego modelu Lotki-Volterry (poświęcono temu zagadnieniu osobny rozdział 6) zadanego równaniami

$$\begin{aligned} \dot{x} &= Ax + Bxy \\ \dot{y} &= Cy + Dxy \end{aligned} \qquad (2.7)$$

gdzie $x = x(t) \in \mathbf{R}$, $y = y(t) \in \mathbf{R}$ natomiast A, B, C, D są stałymi. Testy numeryczne pozwoliły wykreślić typowe (dokładne) trajektorie dla tego modelu i ustalić okres drgań z dokładnością 10^{-13}. Wynik ten jest interesujący, gdyż do tej pory model Lotki-Volterry był modelowany głównie przy pomocy schematów symplektycznych [69, 76, 92, 96], przy czym udawało się odtworzyć tylko ogólne cechy jakościowe trajektorii.

W rozdziale 7 skonstruowane zostały bardzo stabilne i doskonale sprawujące się w długich okresach czasu schematy numeryczne wysokiego rzędu (GR-*N*) przewyższające zdecydowanie dokładnością prezentowane wcześniej metody GR-LEX i GR-SLEX zwłaszcza dla dużych kroków czasowych. Schematy te nie są symplektyczne, nie zachowują objętości w przestrzeni fazowej, nie są też odwracalne w czasie. Swoje świetne cechy jakościowe i ilościowe potwierdzone w licznych symulacjach komputerowych zawdzięczają jedynie zachowywaniu energii i wysokiemu rzędowi. Wyniki te, na razie opublikowane niedawno w formie preprintów [23, 24], mogą się przyczynić do ożywienia

zainteresowania dyskretyzacjami całkowalnymi (zachowującymi całki ruchu). Obecnie bowiem doskonale są rozpracowane metody podwyższania rzędu dla schematów symplektycznych [9, 79, 106, 113], natomiast dość rzadkie są próby poprawiania dokładności schematów całkowalnych, a otrzymane wyniki są skomplikowane [78]. W przypadku metody dyskretnego gradientu, nasze wyniki są, jak się wydaje, pierwszymi rezultatami w tym kierunku.

W rozdziale 8 przedstawiliśmy zarówno ścisłe rozwiązanie oscylatora anharmonicznego, jak też sposób konstrukcji jego dyskretyzacji dokładnej. Podana została też duża rodzina schematów numerycznych zachowujących trajektorie układu w przestrzeni fazowej. W testach dokonano porównania dyskretyzacji dokładnej z dyskretyzacją podaną przez Hirotę [45, 72] oraz metodami wprowadzonymi we wcześniejszych rozdziałach pracy. Okazało się, że dyskretyzację dokładną cechuje dość szybki przyrost błędu globalnego spowodowany niedokładnością obliczeń. Testy pokazały, że dyskretyzacja zaproponowana przez Hirotę, choć nieco mniej dokładna od standardowej metody dyskretnego gradientu, jest bardzo stabilna i zdolna do symulowania układu w szerokim zakresie parametrów początkowych i kroków czasowych. Potwierdziła się też stabilność i bardzo wysoka dokładność metod gradientowych wyższych rzędów, skonstruowanych w rozdziale 7.

Rozdział 9 poświęcony jest ważnemu zagadnieniu mechaniki klasycznej, jakim jest problem Keplera. Pokazano w nim działanie kilku standardowych metod numerycznych oraz dwóch schematów zachowujących całki ruchu i trajektorie [16, 52]. Nowym wynikiem uzyskanym w tej pracy jest modyfikacja wyników pracy [16] w taki sposób, aby całki ruchu zachowywane były dokładnie (schemat skonstruowany w pracy [16] zachowywał wszystkie całki ruchu, ale ich wartości nieco różniły się od wartości dla przypadku ciągłego). Jeśli chodzi o drugą z tych metod, bardzo zaawansowaną matematycznie, to w pracy znalazło się szczegółowe omówienie jej teoretycznych aspektów, do których zaliczyć można zastosowanie 4-wymiarowego jednorodnego oscylatora harmonicznego i transformacji Kustaanheimo-Stiefela. Testy numeryczne pokazały, że obydwa zachowawcze schematy wykazują porównywalną kumulację błędów numerycznych z niewielką przewagą pierwszej z tych metod dla długich czasów.

W pierwszej części rozdziału 10 poddano analizie teoretycznej i eksperymentalnej prosty model ruchu falowego w postaci sprzężonych oscylatorów harmonicznych, który okazał się zupełnie poprawnie symulować wiele zjawisk falowych. W drugiej jego części pokazano, w jaki sposób

dokładna dyskretyzacja równania oscylatora harmonicznego pozwala na otrzymanie dokładnej dyskretyzacji jednowymiarowego równania falowego w postaci

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \tag{2.8}$$

Pomysł polega na rozpatrzeniu ewolucji czasowej modu odpowiadającego liczbie falowej k danego przez $u(x,t) = \hat{u}(k,t)e^{ikx}$ [17]. Po podstawieniu do równania (2.8) otrzymujemy wzór

$$\frac{d^2\hat{u}}{dt^2} = -k^2c^2\hat{u} = -\omega^2\hat{u}\,, \tag{2.9}$$

w którym rozpoznajemy równanie oscylatora harmonicznego. Na podstawie powyższego sformułowano następujący schemat numeryczny: dla zadanych warunków początkowych $u(x,0)$ i $\dot{u}(x,0)$ znajdujemy warunki początkowe dla ich transformat Fouriera

$$\hat{u}(k,0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(x,0)e^{-ikx}dx\,, \tag{2.10}$$

$$\dot{\hat{u}}(k,0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \dot{u}(x,0)e^{-ikx}dx\,. \tag{2.11}$$

Dokładną ewolucję czasową dla równania (2.9) obliczamy wykorzystując dyskretyzację dokładną oscylatora harmonicznego

$$\hat{u}^{n+1} - 2(\cos\omega\varepsilon)\hat{u}^n + \hat{u}^{n-1} = 0\,. \tag{2.12}$$

W ostatnim kroku wykonujemy transformatę odwrotną

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}(k,t)e^{ikx}dx\,, \tag{2.13}$$

aby odtworzyć przebieg $u(x,t)$. Algorytm ten został z powodzeniem przetestowany numerycznie. Wyniki umieszczono w drugiej części rozdziału 10. Zawartość przeglądowej pracy [14] sugeruje, że prezentowany tu algorytm, mimo swojej prostoty, nie jest znany specjalistom od geometrycznych metod numerycznych.

3 Symulacja równania klasycznego oscylatora harmonicznego przy pomocy równań różnicowych

3.1 Wprowadzenie

W rozdziale tym pokazane zostanie jak niewielkie i, zdawać by się mogło, niezbyt ważne zmiany wprowadzane w dyskretyzacjach równania różniczkowego prowadzą do równań różnicowych o całkowicie odmiennych właściwościach. Przez dyskretyzację rozumiana będzie symulacja równania różniczkowego poprzez równanie różnicowe [44]. Zaprezentowane tu zostaną procedury dyskretyzacji prostego modelu fizycznego, które same w sobie nie są zbyt ważne (bo model ten jest ściśle rozwiązywalny), jednak posłużą do przedstawienia pewnych dość ogólnych idei i wyników, które mogą i będą, jak się później okaże, wykorzystane do konstruowania lepszych dyskretyzacji innych równań różniczkowych.

Rozważmy tłumiony oscylator harmoniczny opisany równaniem

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^{\,2} x = 0\,, \tag{3.1}$$

gdzie $x = x(t)$, a kropka oznacza pochodną po czasie. Jest to równanie liniowe, którego ogólne rozwiązanie jest znane. Najbardziej naturalna dyskretyzacja, polega na zastąpieniu x przez x_n, $\dot{x}$ przez iloraz różnicowy $(x_{n+1} - x_n)/\varepsilon$, $\ddot{x}$ przez iloraz różnicowy ilorazów różnicowych, czyli

$$\ddot{x} \to \frac{1}{\varepsilon}\left(\frac{x_{n+2} - x_{n+1}}{\varepsilon} - \frac{x_{n+1} - x_n}{\varepsilon}\right) = \frac{x_{n+2} - 2x_{n+1} + x_n}{\varepsilon^2}\,, \tag{3.2}$$

itd. (jest to metoda Eulera). Równie dobrze możemy zastąpić x np. przez x_{n+1}, $\dot{x}$ przez iloraz różnicowy $(x_n - x_{n-1})/\varepsilon$, lub $\ddot{x}$ przez $(x_{n+1} - 2x_n + x_{n-1})/\varepsilon^2$. Ostatni wzór poprzez swoją symetrię wydaje się być zresztą bardziej naturalny niż (3.2) i rzeczywiście pracuje lepiej (zob. rozdz. 3.2).

W każdym przypadku żądamy, aby granica ciągła (polegająca na zastąpieniu x_n przez $x(t_n) = x(t)$ i wyznaczeniu sąsiednich wartości z rozwinięcia w szereg Taylora funkcji $x(t)$ w punkcie $t = t_n$), zastosowana do danej dyskretyzacji równania różniczkowego dawała w wyniku wyjściowe równanie różniczkowe. Czyli w rozważanym równaniu różnicowym podstawiamy

$$x_n = x(t_n), \quad t_n = n\varepsilon, \quad \varepsilon \to 0 \tag{3.3}$$

$$x_{n+k} = x(t_n + k\varepsilon) = x(t_n) + \dot{x}(t_n)k\varepsilon + \frac{1}{2}\ddot{x}(t_n)k^2\varepsilon^2 + \ldots. \tag{3.4}$$

Pozostawiając tylko wiodące wyrazy powinniśmy otrzymać rozważane równanie różniczkowe. Mówimy wtedy, że schemat numeryczny odpowiadający tej dyskretyzacji jest „zgodny” [83].

W dalszej części tego rozdziału porównane zostaną różne dyskretyzacje tłumionego i nietłumionego oscylatora harmonicznego włączając w to dyskretyzacje dokładne. Mówimy, że dyskretyzacja jest „dokładna” jeśli równość $x_n = x(t_n)$ ma miejsce dla dowolnej wartości ε, a nie tylko w granicy ciagłej (3.3).

3.2 Najprostsze dyskretyzacje oscylatora harmonicznego

Rozważmy następujące trzy równania dyskretne:

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + x_{n-1} = 0\,, \tag{3.5}$$

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + x_n = 0\,, \tag{3.6}$$

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + x_{n+1} = 0\,, \tag{3.7}$$

gdzie ε jest stałą. W granicy ciągłej (3.3) każde z nich prowadzi do równania oscylatora harmonicznego

$$\ddot{x} + x = 0\,. \tag{3.8}$$

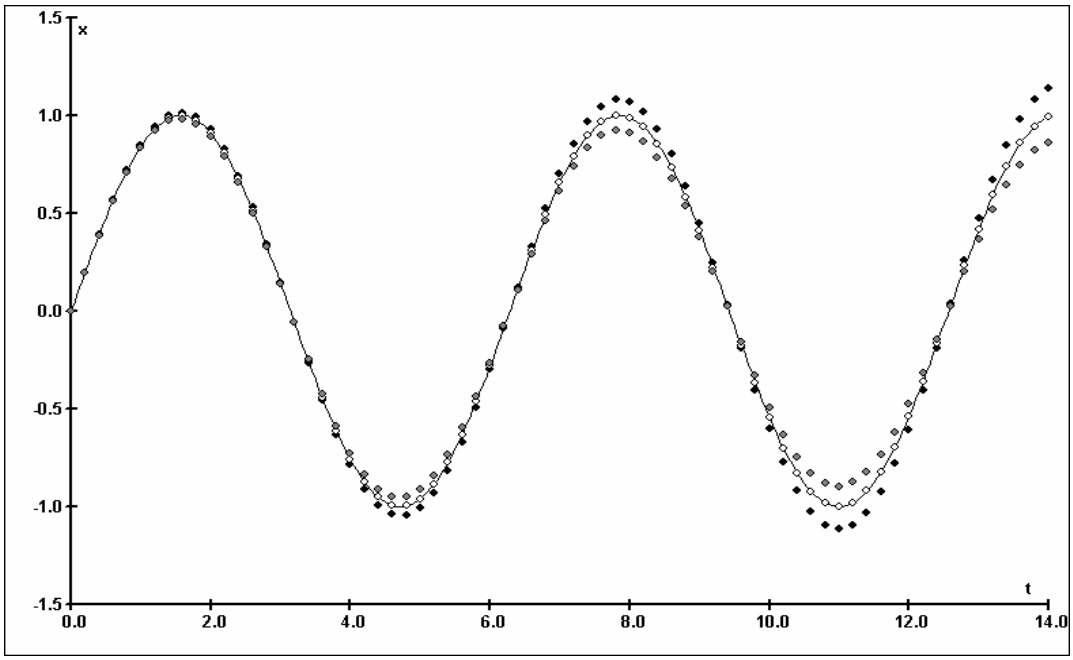
Dla ustalenia uwagi będziemy rozpatrywali tylko rozwiązania odpowiadające warunkom początkowym $x(0) = 0,\ \dot{x}(0) = 1$. Dane początkowe dla tych dyskretyzacji wybieramy tak, aby punkty x_0 i x_1 leżały na wykresie rozwiązania dokładnego.

Dla małych t_n i małych ε, każde z trzech rozwiązań dyskretnych aproksymuje odpowiednie rozwiązanie ciągłe zupełnie dobrze (wykres 3.1). Jednakże globalne zachowania tych rozwiązań (nawet dla bardzo małych ε) są zupełnie różne (wykres 3.2). Rozwiązanie (3.5) dla $t \to \infty$ zanika, podczas gdy (3.7) wykonuje oscylacje z gwałtownie rosnącą amplitudą (na wykresie 3.2 nie zmieścił się już żaden punkt odpowiadający temu rozwiązaniu). Jakościowo tylko rozwiązanie (3.6) przypomina przypadek ciągły (oscylacje o stałej amplitudzie), jednak i ono dla dużych czasów coraz bardziej różni się od rozwiązania dokładnego.

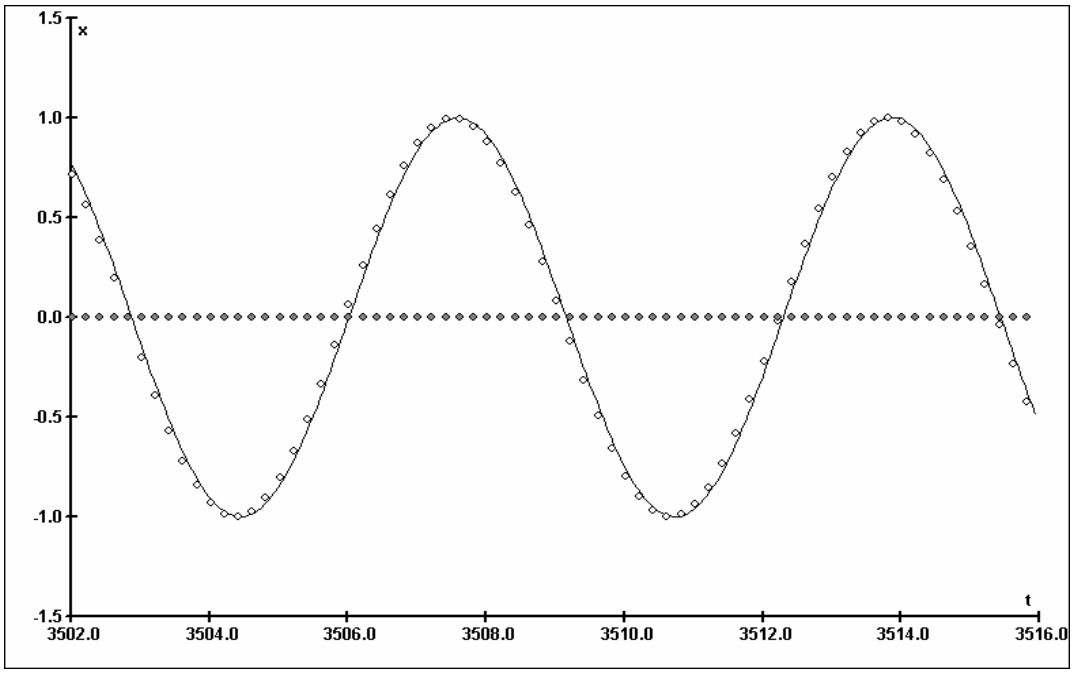
Powstaje pytanie: *jak znaleźć najlepszą dyskretyzację zachowującą globalne własności rozwiązania teoretycznego*?

W dalszej części tego rozdziału zostanie pokazane, jak znaleźć dokładną dyskretyzację tłumionego oscylatora harmonicznego. W szczególności

zaprezentowana zostanie też dyskretyzacja równania (3.1), która jest lepsza od (3.6) i wydaje się najlepszą z możliwych. Zaczniemy jednak od bardzo prostego przykładu ilustrującego podstawowe idee wykorzystywane w tym rozdziale.



Wykres 3.1. Najprostsze dyskretyzacje oscylatora harmonicznego dla małych czasów i $\varepsilon = 0.02$. Romby czarne – rozwiązanie (3.5), romby szare – rozwiązanie (3.7), białe kółka – rozwiązanie (3.6), linia ciągła – dokładne rozwiązanie teoretyczne.



Wykres 3.2. Najprostsze dyskretyzacje oscylatora harmonicznego dla dużych czasów i $\varepsilon = 0.02$. Romby czarne – rozwiązanie (3.5), romby szare – rozwiązanie (3.7), białe kółka – rozwiązanie (3.6), linia ciągła – dokładne rozwiązanie teoretyczne.

3.3 Dokładna dyskretyzacja równania wzrostu wykładniczego

Rozważmy dyskretyzację równania $\dot{x} = x$. Jego ogólne rozwiązanie ma postać

$$x(t) = x(0)e^t, \tag{3.9}$$

a najprostsza jego dyskretyzacja jest dana przez

$$\frac{x_{n+1} - x_n}{\varepsilon} = x_n. \tag{3.10}$$

To równanie dyskretne można rozwiązać bezpośrednio, gdyż sprowadza się ono do równania na ciąg geometryczny $x_{n+1} = (1 + \varepsilon)\, x_n$. Zatem:

$$x_n = (1+\varepsilon)^n x_0. \tag{3.11}$$

Aby porównać to z rozwiązaniem ciągłym, przepiszmy $(1 + \varepsilon)^n$ w formie

$$(1+\varepsilon)^n = \exp(n\ln(1+\varepsilon)) = \exp(\kappa t_n), \tag{3.12}$$

gdzie $t_n = \varepsilon n$ oraz $\kappa = \varepsilon^{-1}\ln(1 + \varepsilon)$. Wówczas rozwiązanie (3.11) można przepisać jako

$$x_n = x_0 e^{\kappa t_n}. \tag{3.13}$$

Widzimy zatem, że dla $\kappa \neq 1$ rozwiązanie ciągłe (3.9) wyznaczone w punkcie t_n

$$x(t) = x(0)e^{t_n}, \tag{3.14}$$

różni się od odpowiadającego mu rozwiązania dyskretnego (3.13). Łatwo można sprawdzić, że $0 < \kappa < 1$. Tylko w granicy $\varepsilon \to 0$, mamy $\kappa \to 1$.

Jakkolwiek jakościowe zachowanie „naiwnej" dyskretyzacji (3.10) dobrze zgadza się z rozwiązaniem ciągłym (wzrost eksponencjalny w obu przypadkach), to jednak (względne) różnice ilościowe przy $t \to \infty$ są bardzo duże z powodu różnych wykładników.

Dyskretyzacja (3.10) może być łatwo poprawiona. Wystarczy mianowicie zastąpić we wzorze (3.11) $1 + \varepsilon$ przez e^{ε}, aby uzyskać zgodność z rozwiązaniem dokładnym (3.14). Taka „dokładna dyskretyzacja" dana jest wzorem

$$\frac{x_{n+1} - x_n}{e^{\varepsilon} - 1} = x_n \tag{3.15}$$

lub po prostu $x_{n+1} = e^{\varepsilon} x_n$. Zauważmy, że $e^{\varepsilon} \approx 1 + \varepsilon$ (dla $\varepsilon \approx 0$) i właśnie to przybliżenie prowadzi do wzoru (3.10).

3.4 Dyskretyzacja oscylatora harmonicznego: rozwiązanie dokładne

Ogólne rozwiązanie równania oscylatora harmonicznego (3.8) ma postać

$$x(t) = x(0)\cos(t) + \dot{x}(0)\sin(t)\,. \tag{3.16}$$

W podrozdziale 3.2 zostało ono porównane z trzema najprostszymi symulacjami dyskretnymi (3.5), (3.6), (3.7). Poniżej zaprezentowane zostaną dokładne rozwiązania tych równań dyskretnych.

Ponieważ rozwiązania dyskretne są zazwyczaj mniej znane niż ciągłe, przypomnijmy, że najprostsze podejście polega szukaniu rozwiązań w formie $x_n = \Lambda^n$ (analogicznej do założenia postaci rozwiązania $x(t) = \exp(\lambda t)$ w przypadku ciągłym – szczegóły w rozdziale 11.1). W rezultacie otrzymujemy równanie charakterystyczne na Λ.

Podejście to zostanie zilustrowane na przykładzie równania (3.5) wynikającego z metody Eulera. Podstawiając $x_n = \Lambda^n$, otrzymujemy następujące równanie charakterystyczne:

$$\Lambda^2 - 2\lambda + (1+\varepsilon^2) = 0 \tag{3.17}$$

z rozwiązaniami $\Lambda_1 = 1 + i\varepsilon$, $\Lambda_2 = 1 - i\varepsilon$. Ogólne rozwiązanie równania (3.5) ma postać

$$x_n = c_1{\Lambda_1}^n + c_2{\Lambda_2}^n\,. \tag{3.18}$$

Wyrażając c_1 i c_2 przez warunki początkowe x_0 i x_1 otrzymujemy

$$x_n = x_1\frac{(1+i\varepsilon)^n - (1-i\varepsilon)^n}{2i\varepsilon} + x_0\frac{(1+i\varepsilon)(1-i\varepsilon)^n - (1-i\varepsilon)(1+i\varepsilon)^n}{2i\varepsilon}\,. \tag{3.19}$$

Oznaczając

$$1+i\varepsilon = \rho e^{i\alpha}\,, \tag{3.20}$$

gdzie $\rho = \sqrt{1+\varepsilon^2}$ oraz α = arctan(ε), po elementarnych przekształceniach otrzymujemy

$$x_n = \rho^n\left(x_0\cos(n\alpha) + \frac{x_1 - x_0}{\varepsilon}\sin(n\alpha)\right). \tag{3.21}$$

Wygodne będzie wprowadzenie oznaczeń

$$\rho = e^{\kappa\varepsilon},\quad t_n = n\varepsilon,\quad \kappa := \frac{1}{2\varepsilon}\ln(1+\varepsilon^2),\quad \omega := \frac{\arctan\varepsilon}{\varepsilon}\,, \tag{3.22}$$

wówczas

$$x_n = \rho^{\kappa t_n}\left(x_0\cos(\omega t_n) + \frac{x_1 - x_0}{\varepsilon}\sin(\omega t_n)\right). \tag{3.23}$$

Można sprawdzić, że $\kappa > 0$ i $\omega < 1$ dla dowolnej wartości $\varepsilon > 0$. Przy $\varepsilon \to 0$ mamy $\kappa \to 0$ i $\omega \to 1$. Dlatego rozwiązanie dyskretne (3.23) charakteryzuje się wykładniczym wzrostem amplitudy obwiedni i mniejszą częstotliwością drgań, niż odpowiadające mu rozwiązanie ciągłe (3.16).

Z podobną sytuacją mamy do czynienia w przypadku dyskretyzacji (3.7) z jedną (ale bardzo ważną) różnicą: zamiast wzrostu mamy wykładniczy zanik. Wzory (3.22) i (3.23) wymagają tylko jednej zmiany, aby być słusznymi również w tym przypadku. Wystarczy zmienić κ na $-\kappa$.

Trzecia z omawianych dyskretyzacji (3.6) charakteryzuje się wartością $\rho = 1$, dlatego w jej przypadku amplituda oscylacji jest stała (przypadek ten zostanie dokładniej omówiony dalej).

Powyższe wyniki są w doskonałej zgodności z zachowaniem rozwiązań dyskretnych prezentowanych na wykresach 3.1 i 3.2.

Rozważmy teraz następującą rodzinę równać dyskretnych (parametryzowanych rzeczywistymi p i q):

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + px_{n-1} + (1 - p - q)x_n + qx_{n+1} = 0 . \tag{3.24}$$

Granica ciągła (3.3) zastosowana do (3.24) prowadzi do równania oscylatora harmonicznego (3.8) dla dowolnych wartości p i q. Rodzina (3.24) obejmuje trzy dyskretyzacje z rozdziału 3.2, jak również (dla $p = q = ¼$) równanie wynikające z zastosowania metody Gaussa-Legendre'a-Rungego-Kutty (patrz rozdział 11.2):

$$x_{n+1} - 2\left(\frac{4 - \varepsilon^2}{4 + \varepsilon^2}\right)x_n + x_{n-1} = 0 . \tag{3.25}$$

Podstawiając $x_n = \Lambda^n$ do (3.24) otrzymujemy następujące równanie charakterystyczne:

$$(1 + q\varepsilon^2)\Lambda^2 - (2 + (p + q - 1)\varepsilon^2\Lambda + (1 + q\varepsilon^2) = 0 . \tag{3.26}$$

Można sformułować następujący problem: *znaleźć dyskretne równanie należące do rodziny* (3.24), *którego zachowanie globalne będzie możliwie jak najbardziej podobne do przypadku ciągłego*.

Aby uzyskać „dobrą" dyskretyzację, naturalne wydaje się żądanie spełnienia przynajmniej dwóch warunków: rozwiązanie ma oscylować i amplituda powinna być stała (tzn. $\rho = 1$, $\kappa = 0$). Warunki te mogą być łatwo wyrażone przez pierwiastki równania kwadratowego (3.26). Po pierwsze muszą one być urojone ($\Delta < 0$), po drugie ich moduł powinien być równy 1, a zatem: $\Lambda_1 = e^{i\alpha}$, $\Lambda_2 = e^{-i\alpha}$. Stąd wynika, że $1 + p\varepsilon^2 = 1 + q\varepsilon^2$, czyli $p = q$. W takim przypadku wyróżnik Δ równania kwadratowego (3.26) dany jest przez:

$$\Delta = -4\varepsilon^2 + \varepsilon^4(1 - 4p) . \tag{3.27}$$

Mamy tu dwie możliwości, jeśli $p \geq ¼$, wówczas $\Delta < 0$ dla dowolnego $\varepsilon \neq 0$, jeśli natomiast $p < ¼$, wówczas $\Delta < 0$ dla dostatecznie małego ε, mianowicie $\varepsilon^2 < 4(1 - 4p)^{-1}$. W każdym z tych przypadków wymagania nie są zbyt restrykcyjne i pozwalają na uzyskanie rodziny dobrych dyskretyzacji oscylatora harmonicznego, parametryzowanej rzeczywistym p. Jeśli $\Lambda_1 = e^{i\alpha}$ i $\Lambda_2 = e^{-i\alpha}$, to rozwiązanie (3.24) ma postać:

$$x_n = x_0 \cos(t_n \omega) + \frac{x_1 - x_0 \cos\alpha}{\sin\alpha} \sin(t_n \omega), \tag{3.28}$$

gdzie $\alpha = \omega \varepsilon$, natomiast

$$\omega = \frac{1}{\varepsilon} \arctan\left(\varepsilon \frac{\sqrt{1 + \varepsilon^2 (p - 1/4)}}{1 + (p - 1/2)\varepsilon^2} \right). \tag{3.29}$$

Zauważmy, że wzór (3.28) jest niezmienniczy względem transformacji $\alpha \to -\alpha$, co oznacza, że jako Λ_1 można wybrać dowolny z dwóch pierwiastków równania (3.26).

Równanie (3.6) jest szczególnym przypadkiem (3.24) dla $p = q = 0$. W podrozdziale (3.2) przekonaliśmy się, że dla małych ε dyskretyzacja ta symuluje zachowanie oscylatora harmonicznego dużo lepiej niż (3.5) i (3.7). Jednak przy dostatecznie dużych ε (mianowicie $\varepsilon > 2$), jej właściwości radykalnie się zmieniają i otrzymujemy wzrost wykładniczy bez oscylacji.

Rozwijając (3.29) w szereg MacLaurina względem ε, otrzymujemy

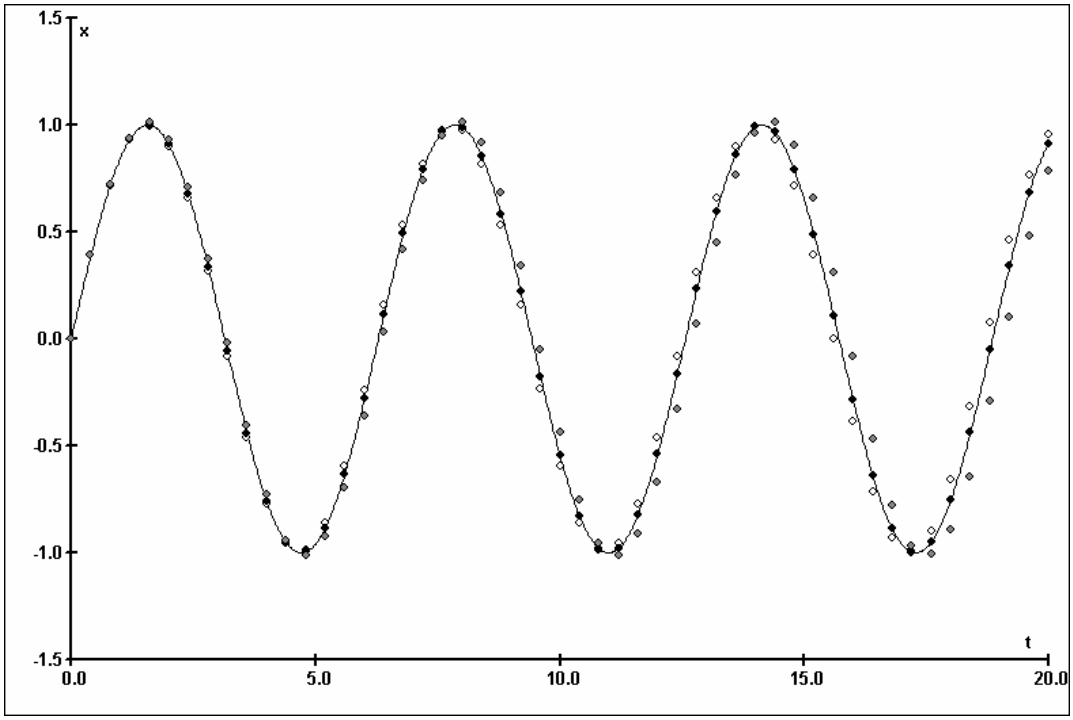
$$\omega \approx 1 + \frac{1 - 12p}{24} \varepsilon^2 + \frac{3 - 40p + 240p^2}{640} \varepsilon^4 + \ldots. \tag{3.30}$$

Dlatego najlepsza aproksymacja równania (3.8) należąca do rodziny (3.24) charakteryzuje się wartością $p = 1/12$:

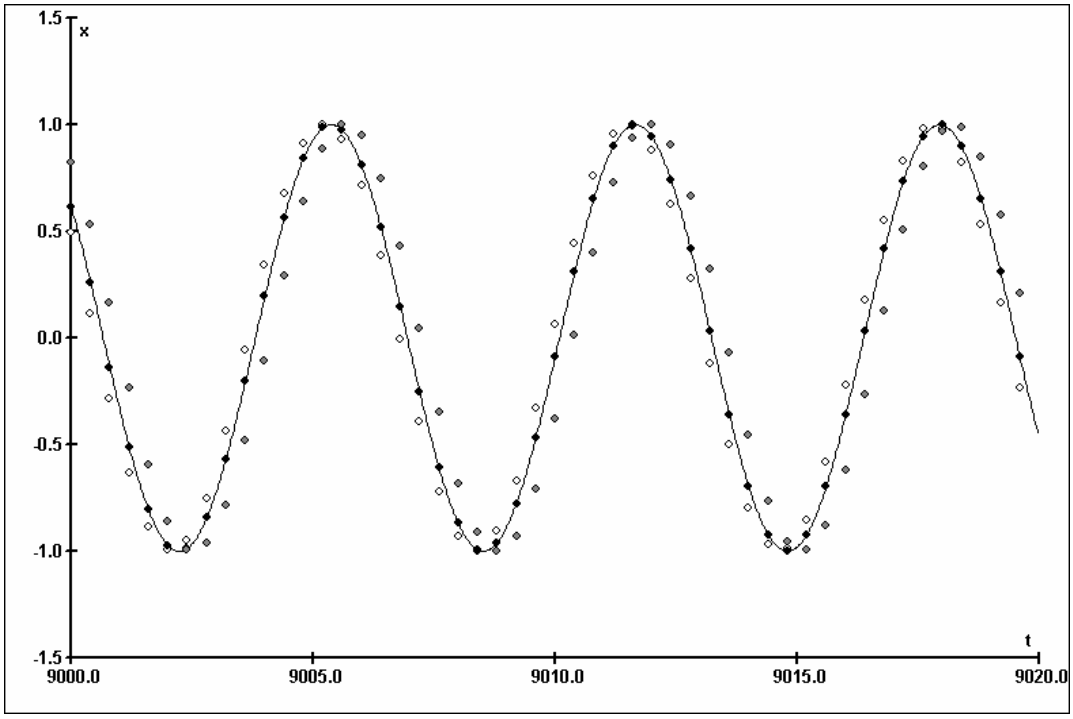
$$x_{n+1} - 2\left(\frac{12 - 5\varepsilon^2}{12 + \varepsilon^2} \right) x_n + x_{n-1} = 0 \tag{3.31}$$

i wtedy $\omega \approx 1 + \varepsilon^4/480$, co jest najbliższe wartości dokładnej $\omega = 1$.

Standardowe metody numeryczne dają wartości podobne (patrz rozdział 11.2). Dyskretyzacja drugiej pochodnej zazwyczaj jest symetryczna, taka jak we wzorze (3.6). Równania dyskretne odpowiadające tym schematom numerycznym modelują równanie oscylatora (3.8) podobnie, lub niewiele lepiej, jak dyskretyzacje opisane w rozdziale 3.2.



Wykres 3.3. Dobre dyskretyzacje oscylatora harmonicznego ($\omega_0 = 1$) dla małych czasów i $\varepsilon = 0.4$. Punkty czarne – dyskretyzacja dokładna (3.49), koła białe – równanie (3.36), koła szare – metoda Rungego-Kutty, linia ciągła – rozwiązanie dokładne.



Wykres 3.4. Dobre dyskretyzacje oscylatora harmonicznego ($\omega_0 = 1$) dla dużych czasów i $\varepsilon = 0.02$. Punkty czarne – dyskretyzacja dokładna (3.49), koła białe – równanie (3.36), koła szare – metoda Rungego-Kutty, linia ciągła – rozwiązanie dokładne.

3.5 Tłumiony oscylator harmoniczny i jego dyskretyzacje

Przechodzimy teraz do równania tłumionego oscylatora harmonicznego (3.1). Jego rozwiązanie ogólne może być wyrażone przez pierwiastki λ_1, λ_2 równania charakterystycznego $\lambda^2 + 2\gamma\lambda + \omega_0^2 = 0$ i dane początkowe $x(0)$, $\dot{x}(0)$:

$$x(t) = \left(\frac{\dot{x}(0) - \lambda_2 x(0)}{\lambda_1 - \lambda_2} \right) e^{\lambda_1 t} + \left(\frac{\dot{x}(0) - \lambda_1 x(0)}{\lambda_2 - \lambda_1} \right) e^{\lambda_2 t} . \tag{3.32}$$

Jeśli tłumienie jest słabe ($\omega > \gamma > 0$), to $\lambda_1 = -\gamma + i\omega$ oraz $\lambda_2 = -\gamma - i\omega$, gdzie $\omega = \sqrt{\omega_0^2 - \gamma^2}$. Wówczas

$$x(t) = x(0)e^{-\gamma t} \cos \omega t + \omega^{-1}(\dot{x}(0) + \gamma x(0))e^{-\gamma t} \sin \omega t . \tag{3.33}$$

Aby otrzymać proste dyskretyzacje równania (3.1), możemy zastąpić pierwszą i drugą pochodną przez ich dyskretne odpowiedniki. Wyniki podrozdziału drugiego sugerują, że najlepszym sposobem dyskretyzacji drugiej pochodnej jest wybór wersji symetrycznej, podobnie jak to miało miejsce we wzorze (3.6). Z drugiej strony mamy przynajmniej 3 możliwości dyskretyzacji pierwszej pochodnej, co prowadzi do następujących dyskretnych symulacji tłumionego oscylatora harmonicznego:

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + 2\gamma \frac{x_n - x_{n-1}}{\varepsilon} + \omega_0^2 x_n = 0 , \tag{3.34}$$

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + 2\gamma \frac{x_{n+1} - x_{n-1}}{2\varepsilon} + \omega_0^2 x_n = 0 , \tag{3.35}$$

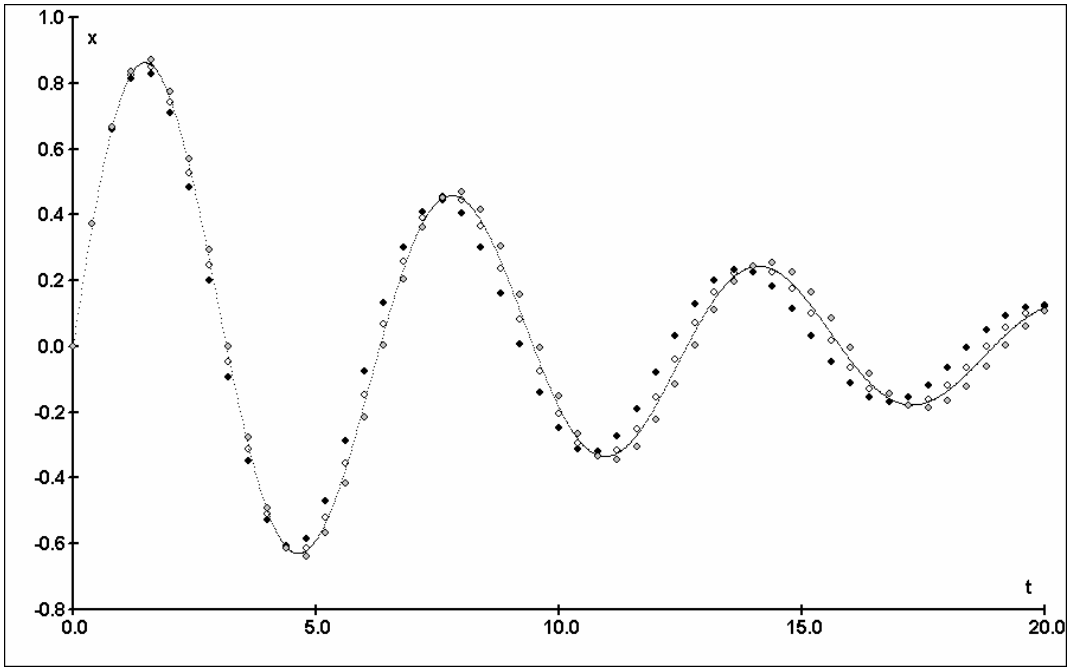
$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + 2\gamma \frac{x_{n+1} - x_n}{\varepsilon} + \omega_0^2 x_n = 0 . \tag{3.36}$$

Można oczekiwać, że najlepsze wyniki da najbardziej symetryczne równanie, czyli (3.35), co faktycznie ma miejsce (wykres 3.3).

3.6 Dokładna dyskretyzacja równania tłumionego oscylatora harmonicznego

W celu znalezienia dokładnej dyskretyzacji (3.1) rozważmy liniowe równanie dyskretne drugiego rzędu:

$$x_{n+2} = 2Ax_{n+1} + Bx_n . \tag{3.37}$$



Wykres 3.5. Najprostsze dyskretyzacje tłumionego oscylatora harmonicznego ($\omega_0 = 1$, $\gamma = 0.1$) dla małych czasów i $\varepsilon = 0.4$. Punkty czarne – równanie (3.34), białe – równanie (3.35), szare – równanie (3.36), linia kropkowana – dokładne rozwiązanie ciągłe.

Ogólne rozwiązanie (3.37) ma następującą postać (szczegóły w rozdziale 11.1):

$$x_n = \frac{x_0(\Lambda_1\Lambda_2{}^n - \Lambda_2\Lambda_1{}^n) + x_1(\Lambda_1{}^n - \Lambda_2{}^n)}{\Lambda_1 - \Lambda_2}, \tag{3.38}$$

gdzie Λ_1, Λ_2 są pierwiastkami równania charakterystycznego $\Lambda^2 - 2A\Lambda - B = 0$, to jest

$$\Lambda_1 = A + \sqrt{A^2 + B}, \qquad \Lambda_2 = A - \sqrt{A^2 + B}\,. \tag{3.39}$$

Wzór (3.38) jest słuszny dla $\Lambda_1 \neq \Lambda_2$, co jest równoważne warunkowi $A^2 + B \neq 0$. Jeśli wartości własne są równe ($\Lambda_1 = \Lambda_2$, $B = -A^2$), mamy $\Lambda_1 = A$ oraz

$$x_n = (1-n)\Lambda_1{}^n x_0 + n\Lambda_1{}^{n-1} x_1\,. \tag{3.40}$$

Czy możliwe jest utożsamienie x_n danego (3.38) *z $x(t_n)$, gdzie $x(t)$ dane jest wzorem* (3.32)*?* Okazuje się, że tak. Wystarczy wyrazić we właściwy sposób λ_1 i λ_2 przez Λ_1 i Λ_2 oraz warunki początkowe $x(0)$, $\dot{x}(0)$ przez x_0, x_1. Jest przy tym całkiem zaskakujące, że powyższe utożsamienie może być wykonane dla dowolnej wartości ε.

Kluczowym będzie następujące powiązanie odpowiadających sobie wartości własnych równań charakterystycznych

$$\Lambda_k{}^n = \exp(n \ln \Lambda_k) = \exp(t_n \lambda_k)\,, \tag{3.41}$$

gdzie, jak zwykle $t_n := n\varepsilon$. Oznacza to, że

$$\lambda_k := \varepsilon^{-1} \ln \Lambda_k \tag{3.42}$$

(dla zespolonych Λ_k, czyli $\Lambda_k = \rho_k e^{i\alpha_k}$, przyjmujemy $\ln\Lambda_k = \ln\rho_k + i\alpha_k$). Wówczas (3.38) przyjmuje postać

$$x_n = \left(\frac{x_1 - x_0 e^{\varepsilon\lambda_2}}{e^{\varepsilon\lambda_1} - e^{\varepsilon\lambda_2}} \right) e^{\lambda_1 t_n} + \left(\frac{x_1 - x_0 e^{\varepsilon\lambda_1}}{e^{\varepsilon\lambda_2} - e^{\varepsilon\lambda_1}} \right) e^{\lambda_2 t_n} . \tag{3.43}$$

Porównując (3.38) z (3.32) otrzymujemy $x_n = x(t_n)$ pod warunkiem, że

$$x(0) = x_0, \quad \dot{x}(0) = \frac{(\lambda_1 - \lambda_2)x_1 - (\lambda_1 e^{\varepsilon\lambda_2} - \lambda_2 e^{\varepsilon\lambda_1})x_0}{e^{\varepsilon\lambda_1} - e^{\varepsilon\lambda_2}} . \tag{3.44}$$

Przypadek zdegenerowany, $\Lambda_1 = \Lambda_2$ (który odpowiada przypadkowi $\lambda_1 = \lambda_2$) może byś rozpatrywany analogicznie (porównaj rozdział 11.1). Wzór (3.40) został otrzymany z (3.38) w granicy $\Lambda_2 \to \Lambda_1$. Dlatego wszystkie wyrażenia dla przypadku zdegenerowanego mogą być wyprowadzone poprzez zastosowanie przejścia granicznego $\lambda_2 \to \lambda_1$.

Zatem mamy bijekcję pomiędzy równaniami różniczkowymi drugiego rzędu ze stałymi współczynnikami oraz równaniami dyskretnymi drugiego rzędu ze stałymi współczynnikami. Ta relacja, odpowiadająca dokładnej dyskretyzacji, wynika ze związku (3.42) pomiędzy wartościami własnymi odpowiednich równań charakterystycznych.

Tłumiony oscylator harmoniczny (3.1) odpowiada równaniu dyskretnemu (3.37) w taki sposób, że

$$2A = e^{-\varepsilon\gamma}\left(e^{\varepsilon\sqrt{\gamma^2 - \omega_0^2}} + e^{-\varepsilon\sqrt{\gamma^2 - \omega_0^2}} \right), \quad B = -e^{-2\varepsilon\gamma} . \tag{3.45}$$

W przypadku słabo tłumionego oscylatora harmonicznego ($\omega_0 > \gamma > 0$), dyskretyzacja dokładna dana jest przez

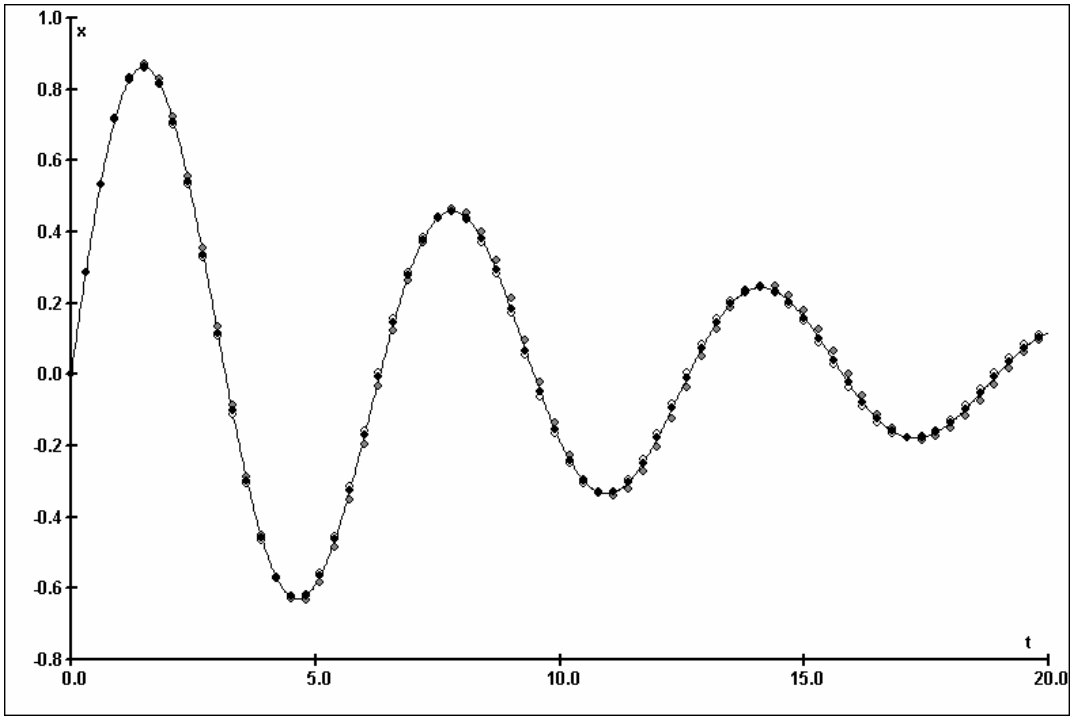
$$A = e^{-\varepsilon\gamma}\cos(\varepsilon\omega), \quad B = -e^{-2\varepsilon\gamma}, \tag{3.46}$$

gdzie $\omega := \sqrt{\omega_0^2 - \gamma^2}$. Innymi słowy, dokładna dyskretyzacja (3.1) ma postać

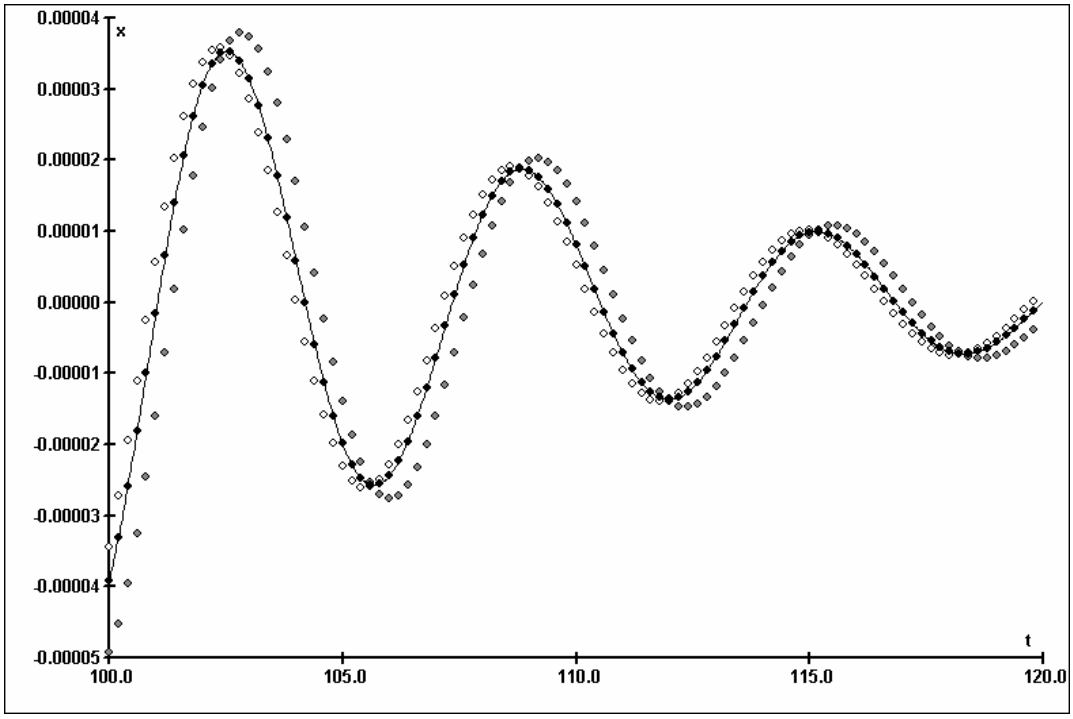
$$x_{n+2} - 2e^{-\varepsilon\gamma}\cos(\varepsilon\omega)x_{n+1} + e^{-2\varepsilon\gamma}x_n = 0 . \tag{3.47}$$

Warunki początkowe powiązane są jak następuje (patrz (3.44)):

$$\begin{aligned} x(0) = x_0, \quad \dot{x}(0) &= \frac{x_1\omega e^{\gamma\varepsilon} - (\gamma\sin(\omega\varepsilon) + \omega\cos(\omega\varepsilon))x_0}{\sin(\omega\varepsilon)}, \\ x_1 &= \left(\dot{x}(0)\frac{\sin(\omega\varepsilon)}{\omega} + \left(\gamma\frac{\sin(\omega\varepsilon)}{\omega} + \cos(\omega\varepsilon) \right) x(0) \right) e^{-\varepsilon\gamma}. \end{aligned} \tag{3.48}$$



Wykres 3.6. Dobre dyskretyzacje tłumionego oscylatora harmonicznego (ω_0 = 1, γ = 0.1) dla małych czasów i ε = 0.2. Punkty czarne – dyskretyzacja dokładna, białe – schemat Rungego-Kutty, szare – równanie (3.35), linia ciągła – rozwiązanie dokładne.



Wykres 3.7. Dobre dyskretyzacje tłumionego oscylatora harmonicznego (ω_0 = 1, γ = 0.1) dla dużych czasów i ε = 0.2. Punkty czarne – dyskretyzacja dokładna, białe – schemat Rungego-Kutty, szare – równanie (3.35), linia ciągła – rozwiązanie dokładne.

Wykresy 3.6 i 3.7 porównują dyskretyzację dokładną z dwiema innymi dobrymi dyskretyzacjami słabo tłumionego oscylatora harmonicznego. Dyskretyzacja dokładna jest rzeczywiście dokładna, czyli punkty trajektorii dyskretnej należą do wykresu dokładnego rozwiązania ciągłego (dla dowolnego ε oraz n). Podobnie jak w przypadku nietłumionym, w pełni symetryczna dyskretyzacja (3.35) jest lepsza niż dyskretyzacja wynikająca z metody GLRK.

Dokładna dyskretyzacja równania oscylatora harmonicznego $\ddot{x} + x = 0$ jest przypadkiem szczególnym wzoru (3.47) i dana jest przez

$$x_{n+2} - 2(\cos\varepsilon)x_{n+1} + x_n = 0\,. \tag{3.49}$$

Można łatwo sprawdzić, że wzór (3.49) można przepisać w postaci

$$\frac{x_{n+2} - 2x_{n+1} + x_n}{(2\sin(\varepsilon/2))^2} + x_{n+1} = 0\,, \tag{3.50}$$

przypominającej „symetryczną" wersję dyskretyzacji wynikającej z metody Eulera (patrz (3.2) i (3.6)), w której ε (pojawiający się w dyskretyzacji drugiej pochodnej) zamieniony został przez $2\sin(\varepsilon/2)$. Dla małych ε mamy $2\sin(\varepsilon/2) \approx \varepsilon$.

Rozwiązanie zagadnienia wartości początkowej dla (3.49) dane jest przez

$$x_n = x_0\cos(n\varepsilon) + \frac{x_1 - x_0\cos\varepsilon}{\sin\varepsilon}\sin(n\varepsilon)\,, \tag{3.51}$$

(porównaj z (3.44)). W ten sposób dyskretnym analogiem $x(0)$ jest po prostu x_0, podczas gdy analogiem prędkości początkowej $\dot{x}(0)$ jest $v_0 = (x_1 - x_0\cos\varepsilon)/\sin\varepsilon$.

Porównanie dyskretyzacji dokładnej (3.49) z trzema innymi równaniami dyskretnymi symulującymi oscylator harmoniczny znajdziemy na wykresach 3.3 i 3.4. Zwróćmy uwagę, że rozważane dyskretyzacje są bardzo dobre nawet dla dużych czasów, jednak nie mogą być lepsze od dyskretyzacji dokładnej. Dyskretyzacja (3.31) wypada również świetnie. Współczynnik przy wyrazie $-2x_n$ we wzorze (3.31)

$$\frac{12 - 5\varepsilon^2}{12 + \varepsilon^2} \approx 1 - \frac{1}{2!}\varepsilon^2 + \frac{1}{4!}\varepsilon^4 + \ldots\,, \tag{3.52}$$

aproksymuje $\cos\varepsilon$ z dokładnością do wyrazów 4-go rzędu. Jeśli zastosowalibyśmy parametry z wykresu 3.4 na wykresie 3.3, to dyskretyzacja (3.31) byłaby nie do odróżnienia od dokładnej.

3.7 Podsumowanie

W rozdziale tym, opartym na pracy [20], pokazane zostało, że dla liniowych równań różniczkowych drugiego rzędu ze stałymi współczynnikami istnieją równania dyskretne, które prawidłowo modelują wszystkie cechy równania różniczkowego. Rozwiązania tych równań dyskretnych są zgodne z rozwiązaniami równań różniczkowych w węzłach siatki dyskretnej. Tego typu dokładne dyskretyzacje mogą być znalezione dla dowolnej stałej sieci ε.

W omawianym przypadku istnieje relacja jeden-do-jednego pomiędzy równaniami różniczkowymi i różnicowymi: każdemu liniowemu równaniu różniczkowemu ze stałymi współczynnikami odpowiada równanie różnicowe, które nazywamy dyskretyzacją dokładną (nazywaną również „najlepszą" [1]). Analogiczne rozważania mogą być przeprowadzone dla równań różniczkowych zwyczajnych ze stałymi współczynnikami dowolnego rzędu [1, 71, 86, 92]. Wielowymiarowe uogólnienia oscylatora harmonicznego zostały ostatnio przedyskutowane w pracy [17].

Należy zaakcentować fakt, że uzyskanie dyskretyzacji prezentowanych w tym rozdziale wymagało założenia istotnych ich zależności od rozważanych równań, co zdecydowanie kontrastuje ze standardowym podejściem numerycznym do równań różniczkowych zwyczajnych, w którym nie przyjmuje się praktycznie żadnych założeń i konstruuje się uniwersalne metody pasujące do dowolnego równania.

W rozdziale tym mamy do czynienia z sytuacją ekstremalną: zastosowana metoda pasuje do bardzo wąskiej klasy równań, ale w wyniku otrzymano dyskretyzację zaskakująco dobrą, wręcz dokładną. Tendencja w tym kierunku jest zresztą coraz bardziej zauważalna wśród specjalistów od metod numerycznych. Tradycyjnie koncentrowano się na stabilności i dokładności schematów różnicowych dla krótkich przedziałów czasowych. Współczesne badania przesuwają akcent w kierunku zachowania niezmienników i prawidłowego odtwarzania cech jakościowych [48, 80]. Wymaga to jednak starannego dopasowania schematu numerycznego do rozważanego równania różniczkowego.

4 Całkowalne dyskretyzacje równania wahadła matematycznego

4.1 Wprowadzenie

W rozdziale tym przedstawione zostaną wyniki eksperymentów porównujących dokładność i wydajność szeregu standardowych i geometrycznych dyskretyzacji na przykładzie równania wahadła matematycznego. Większość prezentowanego materiału została opublikowana w pracy [21].

Równanie wahadła matematycznego jest bardzo dobrze znane, jednak jego odpowiedniki dyskretne pokazują wiele ciekawych i zaskakujących cech, jak choćby pojawianie się zachowań chaotycznych dla dużych kroków czasowych [31, 114]. Zachowanie takie nie występuje w przypadku dyskretyzacji całkowalnych, natomiast nie są wolne od niego dyskretyzacje symplektyczne. Główna uwaga poświęcona zostanie stabilności dyskretyzacji oraz zależności ich okresu i amplitudy od wartości kroku czasowego. Zostaną też opisane i wyjaśnione niewielkie oscylacje wymienionych parametrów ruchu wokół ich wartości średnich.

Spośród mnogości możliwych schematów numerycznych do badań zostały wybrane tylko dyskretyzacje symplektyczne lub całkowalne (czyli zachowujące całkę energii). Stabilność metod symplektycznych w kontekście zachowania energii układów fizycznych jest dobrze znana i od początku lat dziewięćdziesiątych dwudziestego wieku była z powodzeniem wykorzystywana do badania ewolucji układu słonecznego w długiej perspektywie czasowej [105, 111, 114]. Główną przyczyną obserwowanej stabilności jest to, że schemat symplektyczny n-tego rzędu wprawdzie nie zachowuje dokładnie całki energii [34], ale daje dla dużych czasów błąd hamiltonianiu rzędu $O(\varepsilon^n)$, gdzie ε jest stałym krokiem czasowym całkowania [6, 41, 62]. Algorytmy symplektyczne mają zatem od samego początku wielką przewagę we wszelkich analizach porównawczych zachowania się układów fizycznych w długiej perspektywie czasowej. Do tej kategorii metod numerycznych należą dobrze znane schematy takie jak klasyczny *leap-frog* oraz metoda punktu środkowego (*implicit midpoint rule*), bardzo skuteczne mimo swej prostoty, które zostaną porównane z kilkoma (też znanymi) metodami geometrycznymi zachowującymi całkę energii.

Zaproponowana zostanie też jedna nowa dyskretyzacja będąca modyfikacją metody dyskretnego gradientu, która jest bardzo dokładna w przypadku małych drgań (nawet dla bardzo dużych kroków czasowych) i zachowuje przy tym najlepsze cechy swojej metody źródłowej, jak choćby precyzję opisu ruchu

w pobliżu separatrysy. Wszystkie inne rozważane w tym rozdziale dyskretyzacje są znane. W kolejnych rozdziałach wprowadzimy całą serię jeszcze lepszych, zupełnie nowych algorytmów, które są konsekwencją rozwinięcia i uogólnienia pomysłu zaprezentowanego w rozdziale 4.6.

4.2 Symplektyczne dyskretyzacje równań Newtona

Wahadło matematyczne jest szczególnym przypadkiem jednowymiarowego równania Newtona z siłą niezależną od czasu:

$$\ddot{\varphi} = f(\varphi), \tag{4.1}$$

(masę przyjęliśmy równą jedności), które może być zapisane jako układ równań pierwszego rzędu

$$\dot{\varphi} = p, \quad \dot{p} = f(\varphi). \tag{4.2}$$

Dobrze wiadomo, że równania te mają całkę ruchu (energię całkowitą) dla dowolnej funkcji $f(\varphi)$. Zasada zachowania energii dana jest wzorem:

$$\frac{1}{2}\dot{\varphi}^2 + V(\varphi) = E, \quad f(\varphi) = -\frac{dV(\varphi)}{d\varphi}, \tag{4.3}$$

gdzie E = const. Innymi słowy, układ równań (4.2) posiada hamiltonian

$$H(p,q) = \frac{p^2}{2} + V(q). \tag{4.4}$$

Jako przykład do testowania ilościowego różnych metod numerycznych użyte zostanie równanie wahadła matematycznego

$$\ddot{\varphi} = -k\sin(\varphi). \tag{4.5}$$

W tym przypadku zasada zachowania energii ma postać

$$\frac{1}{2}p^2 - k\cos(\varphi) = E. \tag{4.6}$$

Stała k nie jest istotna i może zostać wyeliminowana poprzez zmianę zmiennej t, dlatego we wszystkich obliczeniach numerycznych będziemy zakładali, że $k = 1$.

Dyskretyzacją równania (4.1) nazywać będziemy rodzinę równań różnicowych drugiego rzędu, parametryzowanych krokiem czasowym ε, które w granicy $\varepsilon \to 0$ prowadzą do równania (4.1). Warunek początkowy powinien mieć również postać dyskretną, co oznacza istnienie odwzorowania $\varphi(0) \mapsto \varphi_0$ oraz $\dot{\varphi}(0) \mapsto p_0$.

Wygodnie jest dyskretyzować układ równań pierwszego rzędu (4.2), co automatycznie daje również dyskretyzację p. Mamy wtedy zależne od ε odwzorowanie $(\varphi_n, p_n) \mapsto (\varphi_{n+1}, p_{n+1})$.

Odwzorowanie $(\varphi_n, p_n) \mapsto (\varphi_{n+1}, p_{n+1})$. nazywamy *symplektycznym*, jeśli dla dowolnego n zachodzi

$$d\varphi_{n+1} \wedge d p_{n+1} = d\varphi_n \wedge dp_n \,. \tag{4.7}$$

Poniższe dwa lematy umożliwiają szybkie sprawdzenie symplektyczności dość dużych klas odwzorowań i są dogodne do praktycznych zastosowań.

Lemat 1. *Odwzorowanie* $(\varphi_n, p_n) \mapsto (\varphi_{n+1}, p_{n+1})$ *zdefiniowane w sposób uwikłany poprzez układ równań*

$$\varphi_{n+1} - \varphi_n = P(p_n, p_{n+1}, \varepsilon), \quad p_{n+1} - p_n = R(\varphi_n, \varphi_{n+1}, \varepsilon)\,, \tag{4.8}$$

gdzie P i R są funkcjami różniczkowalnymi, jest symplektyczne wtedy i tylko wtedy, gdy

$$\frac{\partial P}{\partial p_n}\frac{\partial R}{\partial \varphi_n} = \frac{\partial P}{\partial p_{n+1}}\frac{\partial R}{\partial \varphi_{n+1}} \neq 1\,. \tag{4.9}$$

Przeprowadzimy dowód wprost. Różniczkując mianowicie równania (4.8) otrzymujemy:

$$\begin{aligned} d\varphi_{n+1} - d\varphi_n &= P_{,1}\, dp_n + P_{,2}\, dp_{n+1} \\ dp_{n+1} - dp_n &= R_{,1}\, d\varphi_n + R_{,2}\, d\varphi_{n+1} \end{aligned}, \tag{4.10}$$

(przecinek oznacza różniczkowanie cząstkowe). Wówczas

$$\begin{aligned} d\varphi_{n+1} &= \frac{1+P_{,2}\,R_{,1}}{1-P_{,2}\,R_{,2}}\, d\varphi_n + \frac{P_{,1}+P_{,2}}{1-P_{,2}\,R_{,2}}\, dp_n \\ dp_{n+1} &= \frac{R_{,1}+R_{,2}}{1-P_{,2}\,R_{,2}}\, d\varphi_n + \frac{1+P_{,1}\,R_{,2}}{1-P_{,2}\,R_{,2}}\, dp_n \end{aligned}, \tag{4.11}$$

co wymaga aby $P_{,2}R_{,2} \neq 1$ (warunek ten oznacza, że odwzorowanie zdefiniowane przez P, R jest niezdegenerowane). Stąd otrzymujemy

$$d\varphi_{n+1} \wedge dp_{n+1} = \frac{1-P_{,1}\,R_{,1}}{1-P_{,2}\,R_{,2}}\, d\varphi_n \wedge dp_n\,. \tag{4.12}$$

Widzimy zatem, że odwzorowanie jest symplektyczne, jeśli $P_{,1}R_{,1} = P_{,2}R_{,2} \neq 1$, co kończy dowód.

Lemat 2. *Odwzorowanie* $(\varphi_n, p_n) \mapsto (\varphi_{n+1}, p_{n+1})$ *zdefiniowane poprzez układ równań*

$$\varphi_{n+1} - A(\varphi_n, \varepsilon) + \varphi_{n-1} = 0 \quad p_n = \mu_0(\varepsilon)\varphi_{n+1} + B(\varphi_n, \varepsilon), \tag{4.13}$$

jest symplektyczne dla dowolnych różniczkowalnych funkcji A, B.

Aby udowodnić lemat 2, obliczamy

$$dp_{n+1} = \mu_0 d\varphi_{n+2} + TB'd\varphi_{n+1} = \mu_0 TA'd\varphi_{n+1} - \mu_0 d\varphi_n + TB'd\varphi_{n+1}, \tag{4.14}$$

gdzie prim oznacza różniczkowanie (względem pierwszej zmiennej), a T operator przesunięcia, czyli $(TB)(\varphi_n, \varepsilon) = B(\varphi_{n+1}, \varepsilon)$. Stąd otrzymujemy:

$$d\varphi_{n+1} \wedge dp_{n+1} = -\mu_0 d\varphi_{n+1} \wedge d\varphi_n. \tag{4.15}$$

Z drugiej strony jednak $d\varphi_n \wedge dp_n = \mu_0 d\varphi_n \wedge d\varphi_{n+1}$, co kończy dowód.

4.3 Niecałkowalne dyskretyzacje symplektyczne

W podrozdziale tym zaprezentowanych zostanie kilka znanych dyskretyzacji, zachowujących strukturę symplektyczną równań Newtona (porównaj [41], s. 189-190), ale nie zachowujących żadnych całek ruchu. Zatem będą to niecałkowalne dyskretyzacje symplektyczne. Trzeba zaznaczyć, że używając w tej pracy terminu „całkowalność" mamy na myśli istnienie całki ruchu, podobnie jak czyni to autor pracy [104], i nie ma to bezpośredniego związku z teorią układów całkowalnych (solitonowych).

4.3.1 Dyskretyzacja standardowa

Dyskretyzacją standardową równania (4.1) nazywamy schemat, w którym drugą pochodną dyskretyzujemy w sposób symetryczny, czyli $(\varphi_{n+1} - 2\varphi_n + \varphi_{n-1})/\varepsilon^2$, natomiast prawą stronę obliczamy w punkcie φ_n [104]. Dyskretyzacja standardowa równania wahadła matematycznego

$$\frac{\varphi_{n+1} - 2\varphi_n + \varphi_{n-1}}{\varepsilon^2} = -k \sin \varphi_n \tag{4.16}$$

jest niecałkowalna [104]. Otrzymujemy ją poprzez zastosowanie schematu Störmera-Verleta lub symplektycznej metody Eulera (patrz podrozdział 4.3.3). W każdym tych przypadków otrzymujemy to samo równanie dyskretne (4.16), lecz coraz to inną zależność p_n od φ_n i φ_{n+1} (porównaj [54], [85]):

$$p_n = \frac{\varphi_{n+1} - \varphi_n}{\varepsilon} + ck\varepsilon \sin \varphi_n \tag{4.17}$$

gdzie c = 0, ½, 1. Na mocy lematu 2, dyskretyzacje standardowe są symplektyczne dla dowolnej wartości c.

4.3.2 Schemat Störmera-Verleta (*leap-frog*)

Wspomniana już we wstępie metoda Störmera-Verleta (*leap-frog*) (LF) zadana jest następującymi wzorami (porównaj np. [41]):

$$\begin{cases} p_{n+\frac{1}{2}} = p_n + \frac{1}{2}\varepsilon f(\varphi_n), \\ \varphi_{n+1} = \varphi_n + \varepsilon p_{n+\frac{1}{2}}, \\ p_{n+1} = p_{n+\frac{1}{2}} + \frac{1}{2}\varepsilon f(\varphi_{n+1}), \end{cases} \tag{4.18}$$

Eliminując $p_{n+\frac{1}{2}}$, możemy łatwo przekształcić ten schemat do zwykłej postaci jednokrokowej:

$$\begin{aligned} \varphi_{n+1} &= \varphi_n + \varepsilon p_n + \frac{1}{2}\varepsilon^2 f(\varphi_n), \\ p_{n+1} &= p_n + \frac{1}{2}\varepsilon(f(\varphi_n + f(\varepsilon p_n + \frac{1}{2}\varepsilon^2 f(\varphi_n))). \end{aligned} \tag{4.19}$$

Można go również przedstawić w formie układu równań

$$\frac{\varphi_{n+1} - 2\varphi_n + \varphi_{n-1}}{\varepsilon^2} = f(\varphi_n), \tag{4.20}$$

$$p_n = \frac{\varphi_{n+1} - \varphi_n}{\varepsilon} - \frac{\varepsilon}{2} f(\varphi_n). \tag{4.21}$$

W przypadku wahadła matematycznego, $f(\varphi) = -k\sin(\varphi)$, rozpoznajemy w nich dyskretyzację naturalną, czyli wzory: (4.16) oraz (4.17) z c = ½..

4.3.3 Symplektyczne schematy Eulera

Układ równań (4.2) jest przykładem ogólniejszej klasy układów równań postaci

$$\dot{\varphi} = g(\varphi, p), \quad \dot{p} = h(\varphi, p). \tag{4.22}$$

gdzie g, h są zadanymi funkcjami dwóch zmiennych. Układ (4.22) można poddać dyskretyzacji na jeden z dwóch sposobów:

$$\varphi_{n+1} = \varphi_n + \varepsilon g(\varphi_n, p_{n+1}), \quad p_{n+1} = p_n + \varepsilon h(\varphi_n, p_{n+1}), \tag{4.23}$$

$$\varphi_{n+1} = \varphi_n + \varepsilon g(\varphi_{n+1}, p_n), \quad p_{n+1} = p_n + \varepsilon h(\varphi_{n+1}, p_n). \tag{4.24}$$

Obie te dyskretyzacje nazywane są albo symplektycznymi metodami Eulera [41] lub metodami podziału symplektycznego [66]. Dyskretyzując w ten sposób równanie (4.2), otrzymujemy odpowiednio

$$\varphi_{n+1} = \varphi_n + \varepsilon p_{n+1}, \quad p_{n+1} = p_n + \varepsilon f(\varphi_n), \tag{4.25}$$

$$\varphi_{n+1} = \varphi_n + \varepsilon p_n, \quad p_{n+1} = p_n + \varepsilon f(\varphi_{n+1}). \tag{4.26}$$

Obydwa powyższe układy równań prowadzą do równania (4.20), lecz zamiast (4.21) otrzymujemy odpowiednio

$$p_n = \frac{\varphi_{n+1} - \varphi_n}{\varepsilon} - \varepsilon f(\varphi_n) \quad \text{lub} \quad p_n = \frac{\varphi_{n+1} - \varphi_n}{\varepsilon}, \tag{4.27}$$

co w przypadku wahadła matematycznego daje dyskretyzację standardową z $c = 1$ lub $c = 0$.

4.3.4 Metoda punktu środkowego (*implicit midpoint rule*)

Dowolne równanie różniczkowe pierwszego rzędu $\dot{x} = F(x)$ może zostać zdyskretyzowane przy pomocy (niejawnej) metody punktu środkowego (*implicit midpoint*) (MID), która w tym przypadku sprowadza się do metody pierwszego rzędu Gaussa-Legendre'a-Runge-Kutty (porównaj [41]). Pierwszą pochodną zamieniamy w niej na iloraz różnicowy, a prawa strona jest obliczana w punkcie środkowym $\frac{1}{2}(x_n + x_{n+1})$. Metoda ta w przypadku najprostszego układu równań Hamiltona (4.2) daje:

$$\begin{aligned} \varphi_{k+1} &= \varphi_k + \frac{1}{2}\varepsilon(p_k + p_{k+1}), \\ p_{k+1} &= p_k + \varepsilon f\left(\frac{\varphi_k + \varphi_{k+1}}{2}\right). \end{aligned} \tag{4.28}$$

W szczególnym przypadku wahadła matematycznego dostajemy:

$$\begin{aligned} \frac{\varphi_{k+1} - 2\varphi_k + \varphi_{k-1}}{\varepsilon^2} &= -\frac{1}{2}k\left(\sin\left(\frac{\varphi_{k+1} + \varphi_k}{2}\right) + \sin\left(\frac{\varphi_k + \varphi_{k-1}}{2}\right)\right), \\ p_k &= \frac{\varphi_{k+1} - \varphi_k}{\varepsilon} + \frac{1}{2}\varepsilon k \sin\left(\frac{\varphi_{k+1} + \varphi_k}{2}\right). \end{aligned} \tag{4.29}$$

Metoda punktu środkowego ma zupełnie dobre właściwości: jest symplektyczną, symetryczną (odwracalną w czasie) metodą rzędu drugiego. Jej symplektyczność wynika bezpośrednio z Lematu 1, gdyż wzory (4.28) dają $P_{,1} = P_{,2}$ oraz $R_{,1} = R_{,2}$. Wzory te są też jawnie „odwracalne w czasie" (symetryczne). Mianowicie zamiana miejscami indeksów k i $k+1$ wraz z jednoczesną zmianą znaku zmiennych p_k i p_{k+1} nie zmienia wzorów (4.28).

4.4 Metody rzutowania na powierzchnię stałej energii

Dyskretyzacje niecałkowalne mogą być modyfikowane tak, aby zachowywały całkę energii „na siłę", czyli poprzez rzutowanie wyniku każdego kroku na powierzchnię stałej energii. Zatem dla dowolnej metody jednokrokowej można skonstruować odpowiadającą jej metodę rzutowania. Działanie rzutowania zademonstrowane zostanie na przykładzie dyskretyzacji naturalnej, czyli dla schematu Störmera-Verleta.

4.4.1 Metoda rzutowania standardowego

Niech $\dot{x} = F(x), x \in R^2$ będzie równaniem różniczkowym pierwszego rzędu, $x_{n+1} = \Phi_\varepsilon(x_n)$ - dowolną jednokrokową dyskretyzacją tego równania, zaś $g(x) = 0$ równaniem definiującym więzy (w tym przypadku: powierzchnię zanurzoną w R^3), które chcemy zachować. Prezentowana metoda (PROJ) polega na obliczeniu punktu $\tilde{x}_{n+1} := \Phi_\varepsilon(x_n)$, a następnie wyznaczeniu jego rzutu prostopadłego na powierzchnię $g(x) = 0$ (patrz [41]), który oznaczymy poprzez x_{n+1}. Definiuje to procedurę uzyskania następnego kroku: $x_n \to x_{n+1}$. Innymi słowy

$$x_{n+1} = \tilde{x}_n + \lambda \nabla g(\tilde{x}_{n+1}) \tag{4.30}$$

przy czym λ wybieramy tak, aby $g(x_{n+1}) = 0$.

Stosując to podejście do wahadła matematycznego wygodnie będzie zdefiniować x jako

$$x = \left(\varphi, \frac{\dot{\varphi}}{\omega}\right) \equiv (\varphi, p), \tag{4.31}$$

gdzie $\omega = \sqrt{k}$. Dzięki powyższej definicji p, składowe x są bezwymiarowe. Jeśli $k = 1$, (co zakładamy w tym rozdziale), wówczas definicja ta pokrywa się podanym wcześniej wzorem (4.2). Z (4.6) wynika, że funkcja występująca w równaniu więzów $g(x) = 0$ ma postać

$$g(x) = \frac{1}{2} p^2 - k\cos(\varphi) - h, \tag{4.32}$$

gdzie $h = E/\omega^2$. Z (4.30) dostajemy wówczas

$$\varphi_{n+1} = \tilde{\varphi}_{n+1} + \lambda \sin\tilde{\varphi}_{n+1}, \quad p_{n+1} = (1+\lambda)\tilde{p}_{n+1} \tag{4.33}$$

przy czym parametr λ obliczany jest z równania

$$\frac{1}{2}(1+\lambda^2)\tilde{p}^2{}_{n+1} - \cos(\tilde{\varphi}_n + \lambda \sin\tilde{\varphi}_{n+1}) = h. \tag{4.34}$$

Możemy posłużyć się tu iteracją Newtona $\lambda_{i+1} = \lambda_j - \Delta\lambda_j$, gdzie

$$\Delta\lambda_j = -\frac{\frac{1}{2}(1+\lambda^2)\tilde{p}^2{}_{n+1} - \cos(\tilde{\varphi}_n + \lambda\sin\tilde{\varphi}_{n+1}) - h}{\tilde{p}^2{}_{n+1} + \sin^2\tilde{\varphi}^2{}_{n+1}} \tag{4.35}$$

przy czym wystarczający i zarazem wygodny jest wybór $\lambda_0 = 0$. Granica $\lim_{j\to\infty}\lambda_j$ daje nam przybliżone rozwiązanie (4.35).

4.4.2 Metoda rzutowania symetrycznego

Algorytm jednokrokowy $x_{n+1} = \Phi_\varepsilon(x_n)$ nazywany jest symetrycznym (lub odwracalnym w czasie), jeśli $\Phi_{-\varepsilon} = \Phi_\varepsilon^{-1}$. Odwracalne w czasie są równania mechaniki klasycznej, dlatego zachowanie tej własności jest wygodne i z zasady poprawia wyniki numeryczne. Spośród wymienionych wcześniej metod nieodwracalne w czasie są obie symplektyczne metody Eulera oraz metoda rzutowania standardowego, natomiast odwracalne są schematy Störmera-Verleta oraz punktu środkowego.

Metoda rzutowania symetrycznego (SYM-PROJ) jest modyfikacją rzutowania standardowego mającą na celu zachowanie odwracalności w czasie. Stosujemy ją przy podobnych założeniach jak w rzutowaniu standardowym, ale żądamy dodatkowo symetrii funkcji Φ_ε. Metoda ta sprowadza się do wykonania następujących kroków [4, 38]:

$$\begin{aligned}
&\hat{x}_n = x_n + \lambda\nabla g(x_n),\\
&\tilde{x}_{n+1} = \Phi_\varepsilon(\hat{x}_n),\\
&x_{n+1} = \tilde{x}_{n+1} + \lambda\nabla g(\tilde{x}_{n+1}),
\end{aligned} \tag{4.36}$$

przy czym zakładamy, że $g(x_n) = 0$ i obliczamy parametr λ z warunku $g(x_{n+1}) = 0$.

4.5 Dyskretyzacje całkowalne

Przypomnijmy, że w pracy tej całkowalność schematu numerycznego oznacza, że istnieją w nim całki ruchu. Równanie Newtona (4.1) posiada całkę energii, więc jego dyskretyzację nazywamy całkowalną, o ile posiada całkę, którą możemy traktować jako dyskretną analogię energii. Wymagamy, aby w granicy ciągłej ta dyskretna całka przechodziła w zwykłą całką energii równania Newtona.

4.5.1 Symplektyczne dyskretyzacje Surisa

Dyskretyzacje podobne do standardowej (*standard-like*) są definiowane poprzez równania [104]:

$$\begin{aligned} \varphi_{n+1} &= \varphi_n + \varepsilon p_{n+1}, \\ p_{n+1} &= p_n + \varepsilon F(\varphi_n, \varepsilon) \end{aligned} \tag{4.37}$$

przy czym F musi spełniać zależność $F(\varphi_n, 0) = f(\varphi_n)$. Dla danej funkcji f istnieje nieskończenie wiele funkcji F spełniających ten warunek. Wszystkie dyskretyzacje tej postaci są symplektyczne, co można sprawdzić stosując Lemat 1 i zauważając, że $P_{,1} = R_{,2} = 0$. Podobnie jak w przypadku niecałkowalnych dyskretyzacji symplektycznych, z równań (4.37) otrzymujemy:

$$\begin{aligned} &\varphi_{n+1} - 2\varphi_n + \varphi_{n-1} = \varepsilon^2 F(\varphi_n, \varepsilon), \\ &p_n = \frac{\varphi_{n+1} - \varphi_n}{\varepsilon} - \varepsilon F(\varphi_n, \varepsilon) = \frac{\varphi_n - \varphi_{n-1}}{\varepsilon}. \end{aligned} \tag{4.5-1}$$

Interesują nas przypadki dyskretyzacji całkowalnych, to jest takich, które zachowują całkę energii. Dwie dyskretyzacje wahadła matematycznego odpowiadające temu warunkowi znalazł Suris (Sur1, Sur2) [103, 104]:

$$\varphi_{n+1} - 2\varphi_n + \varphi_{n-1} = -2\arctan\left(\frac{k\varepsilon^2 \sin(\varphi_n)}{2 + k\varepsilon^2 \cos(\varphi_n)}\right), \tag{4.38}$$

$$\varphi_{n+1} - 2\varphi_n + \varphi_{n-1} = -4\arctan\left(\frac{k\varepsilon^2 \sin(\varphi_n)}{4 + k\varepsilon^2 \cos(\varphi_n)}\right). \tag{4.39}$$

Wykonując bezpośrednie obliczenia można sprawdzić, że równanie (4.38) ma całkę ruchu daną wzorem

$$E_1 = \frac{1}{2}\left(\frac{2\sin\frac{\varphi_{n+1} - \varphi_n}{2}}{\varepsilon}\right)^2 - \frac{1}{2}k(\cos\varphi_n + \cos\varphi_{n+1}), \tag{4.40}$$

którą można wyrazić przy pomocy φ_n oraz p_n:

$$E_1 = \frac{1 - \cos\varepsilon p_n}{\varepsilon^2} - \frac{1}{2}k(\cos\varphi_n + \cos(\varphi_n - \varepsilon p_n)). \tag{4.41}$$

Z kolei dyskretyzacja (4.39) posiada całkę ruchu

$$E_2 = \frac{1}{2}\left(\frac{4\sin\frac{\varphi_{n+1} - \varphi_n}{4}}{\varepsilon}\right)^2 - k\cos\frac{\varphi_n + \varphi_{n+1}}{2}, \tag{4.42}$$

która, wyrażona poprzez φ_n i p_n, przyjmuje postać:

$$E_2 = \frac{4}{\varepsilon^2}\left(1-\cos\frac{\varepsilon p_n}{2}\right) - k\cos\left(\varphi_n - \frac{\varepsilon p_n}{2}\right). \tag{4.43}$$

Warto zauważyć, że hamiltonian dany wzorem (4.4) nie jest zachowywany przez te dyskretyzacje, czyli $H(p_n,\varphi_n)$ nie jest całką ruchu. Jedynie dla $\varepsilon \approx 0$ całki ruchu dane wzorami (4.41) i (4.43) są w przybliżeniu równe $H(p_n,\varphi_n)$.

4.5.2 Metoda dyskretnego gradientu

Metoda dyskretnego gradientu (GR) [66, 85, 87] jest ogólną i bardzo silną metodą budowania schematów numerycznych zachowujących dowolną liczbę całek ruchu oraz szereg innych własności równań Hamiltona opisujących układ fizyczny [65]. Jednak metoda ta nie jest symplektyczna. W przypadku jednowymiarowych układów hamiltonowskich (z hamiltonianem niezależnym od czasu), czyli równań

$$\dot{\varphi} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial \varphi} \tag{4.44}$$

metoda dyskretnego gradientu redukuje się do następującego prostego schematu. Lewe strony formuł (4.44) poddajemy dyskretyzacji w najprostszy sposób (stosując ilorazy różnicowe), podczas gdy prawe strony zastępujemy poprzez tak zwane dyskretne (lub średnie) gradienty:

$$\frac{\varphi_{n+1}-\varphi_n}{\varepsilon} = \frac{\Delta H}{\Delta p}, \quad \frac{p_{n+1}-p_n}{\varepsilon} = -\frac{\Delta H}{\Delta \varphi}. \tag{4.45}$$

Dyskretny gradient $\overline{\nabla} H \equiv \left(\frac{\Delta H}{\Delta \varphi}, \frac{\Delta H}{\Delta p}\right)$ różniczkowanej funkcji $H(\varphi, p)$ z definicji (patrz [66]) spełnia warunek

$$H(\varphi_{n+1}, p_{n+1}) - H(\varphi_n, p_n) = \frac{\Delta H}{\Delta \varphi}(\varphi_{n+1}-\varphi_n) + \frac{\Delta H}{\Delta p}(p_{n+1}-p_n). \tag{4.46}$$

Określenie to jest niejednoznaczne. Istnieją różne definicje dyskretnego gradientu, spełniające powyższy warunek. Najprostszą z tych definicji jest [103]:

$$\begin{aligned} \frac{\Delta H}{\Delta \varphi} &= \frac{H(\varphi_{n+1}, p_n) - H(\varphi_n, p_n)}{\varphi_{n+1}-\varphi_n} \\ \frac{\Delta H}{\Delta p} &= \frac{H(\varphi_{n+1}, p_{n+1}) - H(\varphi_{n+1}, p_n)}{p_{n+1}-p_n} \end{aligned}. \tag{4.47}$$

czyli w każdym z dwóch wzorów mamy przyrost hamiltonianu względem innej zmiennej. Inne możliwe definicje są przedstawione np. w pracach [66] i [35].

Wszystkie te definicje dają ten sam wynik gdy zastosujemy je do przypadku rozpatrywanego w naszej pracy, czyli $H(\varphi, p) = T(p) + V(\varphi)$. Wówczas $\overline{\nabla} H = \overline{\nabla} T + \overline{\nabla} V$, gdzie

$$\overline{\nabla} T = \frac{T(p_{n+1}) - T(p_n)}{p_{n+1} - p_n}, \qquad \overline{\nabla} V = \frac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n}. \tag{4.48}$$

Zatem

$$\begin{cases} \dfrac{p_{n+1} + p_n}{2} = \dfrac{\varphi_{n+1} - \varphi_n}{\varepsilon}, \\ \dfrac{p_{n+1} - p_n}{\varepsilon} = -\dfrac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n}. \end{cases} \tag{4.49}$$

Powyższy schemat numeryczny może być również otrzymany jako przypadek szczególny zmodyfikowanej metody punktu środkowego (*modified midpoint rule*) [54]. Układ równań (4.49) może być przepisany w postaci równania drugiego stopnia na φ_n oraz równania definiującego p_n:

$$\begin{cases} \dfrac{\varphi_{n+1} - 2\varphi_n + \varphi_{n-1}}{\varepsilon^2} = -\dfrac{1}{2}\left(\dfrac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n} + \dfrac{V(\varphi_n) - V(\varphi_{n-1})}{\varphi_n - \varphi_{n-1}} \right), \\ p_n = \dfrac{\varphi_{n+1} - \varphi_n}{\varepsilon} + \dfrac{1}{2}\varepsilon \left(\dfrac{V(\varphi_{n+1}) - V(\varphi_n)}{\varphi_{n+1} - \varphi_n} \right). \end{cases} \tag{4.50}$$

Podstawiając $V(\varphi) = -k\cos\varphi$ otrzymujemy dyskretyzację wahadła matematycznego. Mnożąc równania (4.49) stronami, można łatwo udowodnić, że układ równań (4.50) posiada całkę pierwszą ruchu

$$E = \frac{1}{2} p_n^{\,2} + V(\varphi_n), \tag{4.51}$$

która jest dokładnie równa wartości hamiltonianu (4.4) w punkcie φ_n, p_n. Całki ruchu (4.41) i (4.43) osiągają tę zgodność tylko w granicy $\varepsilon \to 0$.

4.6 Poprawka zachowująca okres małych drgań

Równanie klasycznego oscylatora harmonicznego $\ddot{\varphi} + \omega^2 \varphi = 0$ posiada dyskretyzację dokładną ([20], porównaj także [1, 89]), to jest taką, która w n-tym kroku φ_n przyjmuje taką samą wartość, jak rozwiązanie tego równania $\varphi(t)$ w punkcie $n\varepsilon$ (dla dowolnych ε oraz n):

$$\begin{aligned} &\varphi_{n+1} - 2\varphi_n \cos\omega\varepsilon + \varphi_{n-1} = 0, \\ &p_n = \frac{\omega}{\sin\omega\varepsilon}(\varphi_{n+1} - \varphi_n \cos\omega\varepsilon). \end{aligned} \tag{4.52}$$

Bezpośrednim rachunkiem można łatwo sprawdzić, że energia jest również dokładnie zachowana, co oznacza, że wartość wyrażenia

$$E = \frac{1}{2} p_n^{\,2} + \frac{1}{2} \omega^2 \varphi_n^{\,2} \tag{4.53}$$

nie zależy od n. Istnienie dokładnej dyskretyzacji równania oscylatora harmonicznego zostało ostatnio wykorzystane do otrzymania dyskretyzacji zagadnienia Keplera zachowującej wszystkie całki ruchu i trajektorię [16].

Rozważmy równania Newtona postaci (4.2). Ograniczymy się do równań, które opisują układy będące w stanie równowagi trwałej w punkcie $\varphi = 0$ ($f'(0) < 0$). Wówczas potencjał $V = V(\varphi)$ posiada lokalne minimum w punkcie $\varphi = 0$, co oznacza, że $V'(0) = f(0) = 0$. Oznaczmy

$$\omega_0 = \sqrt{V''(0)}\,, \tag{4.54}$$

wtedy

$$V(\varphi) = V_0 + \frac{1}{2} \omega_0^{\,2} \varphi^2 + ..., \tag{4.55}$$

i małe oscylacje względem punktu równowagi mogą być przybliżane poprzez równanie klasycznego oscylatora harmonicznego z $\omega = \omega_0$.

Czy istnieje dyskretyzacja, które w granicy $\varphi_n \approx 0$ (ε jest ustalone) staje się dokładna? Znane dyskretyzacje, włączając przedstawione wcześniej w tym rozdziale, nie posiadają takiej własności. Można ją jednak otrzymać poprzez modyfikację metody dyskretnego gradientu. Okazuje się, że wystarczy zastąpić ε przez pewną funkcję $\delta = \delta(\varepsilon)$ we wzorach (4.50). Postać tej funkcji otrzymana zostanie poprzez porównanie z równaniem oscylatora harmonicznego (w granicy $\varphi \approx 0$).

Linearyzujemy równanie (4.50) (z ε zastąpionym przez δ) wokół $\varphi_n = 0$ biorąc pod uwagę (4.55). Dostajemy wówczas

$$\begin{aligned} &\frac{\varphi_{n+1} - 2\varphi_n + \varphi_{n-1}}{\delta^2} = -\frac{\omega_0^{\,2}}{4}(\varphi_{n+1} + 2\varphi_n + \varphi_{n-1}), \\ &p_n = \frac{\varphi_{n+1} - \varphi_n}{\delta} + \frac{1}{4}\omega_0^{\,2}\delta(\varphi_{n+1} + \varphi_n), \end{aligned} \tag{4.56}$$

co jest równoważne

$$\begin{aligned} &\varphi_{n+1} - 2\left(\frac{4 - \omega_0^{\,2}\delta^2}{4 + \omega_0^{\,2}\delta^2}\right)\varphi_n + \varphi_{n-1} = 0, \\ &p_n = \frac{4 + \omega_0^{\,2}\delta^2}{4\delta}\left(\varphi_{n+1} - \left(\frac{4 - \omega_0^{\,2}\delta^2}{4 + \omega_0^{\,2}\delta^2}\right)\varphi_n\right). \end{aligned} \tag{4.57}$$

Porównując układy równań (4.52) oraz (4.56) stwierdzamy, że są one zgodne wtedy i tylko wtedy, gdy zachodzą równości

$$\left(\frac{4-\omega_0{}^2\delta^2}{4+\omega_0{}^2\delta^2}\right)=\cos\omega\varepsilon,\quad \frac{4+\omega_0{}^2\delta^2}{4\delta}=\frac{\omega}{\sin\omega\varepsilon}. \tag{4.58}$$

Rozwiązując ten układ równań otrzymujemy

$$\omega=\omega_0,\quad \delta=\frac{2}{\omega_0}\tan\left(\frac{\varepsilon\omega_0}{2}\right). \tag{4.59}$$

Można zatem zaproponować następującą nową dyskretyzację równania Newtona (4.1), (4.2) (zmodyfikowaną metodę dyskretnego gradientu – MOD-GR):

$$\begin{cases}\dfrac{\varphi_{n+1}-2\varphi_n+\varphi_{n-1}}{\delta^2}=-\dfrac{1}{2}\left(\dfrac{V(\varphi_{n+1})-V(\varphi_n)}{\varphi_{n+1}-\varphi_n}+\dfrac{V(\varphi_n)-V(\varphi_{n-1})}{\varphi_n-\varphi_{n-1}}\right),\\ p_n=\dfrac{\varphi_{n+1}-\varphi_n}{\delta}+\dfrac{1}{2}\delta\left(\dfrac{V(\varphi_{n+1})-V(\varphi_n)}{\varphi_{n+1}-\varphi_n}\right),\end{cases} \tag{4.60}$$

gdzie δ jest zdefiniowana poprzez (4.59) a ω_0 jest dana wzorem (4.55). Dyskretyzacja ta staje się dokładna dla małych drgań przy dowolnej stałej wartości ε. Oznacza to, że dla $\varphi_n \approx 0$ okres i amplituda rozwiązania przybliżonego powinny być bardzo bliskie wartościom dokładnym nawet dla dużych ε.

4.7 Eksperymenty numeryczne

Przeprowadzono szereg eksperymentów numerycznych z zastosowaniem schematów zaprezentowanych w poprzednich podrozdziałach. Parametrem wszystkich doświadczeń była prędkość początkowa p_0 (położenie początkowe było zawsze to samo: $j_0 = 0$. W przypadku ciągłym (4.5) mamy 3 możliwości: ruch drgający ($|p_0| < 2$), ruch obrotowy ($|p_0| > 2$) oraz ruch wzdłuż separatrysy ($p_0 = \pm 2$) od j = 0 do (asymptotycznie) j = ± p. Teoretyczna amplituda oscylacji może być łatwo wyznaczona z zasady zachowania energii (4.6) (gdzie $k = 1$, tj. $\frac{1}{2}p_0{}^2 - 1 = -\cos A_{th}$):

$$2\sin\frac{A_{th}}{2}=p_0. \tag{4.61}$$

W szczególności wykonano wiele obliczeń numerycznych dla następujących danych początkowych:

- $p_0 = 0.1$, wówczas $A_{th} \approx 0.0318443\text{p} \approx 0.1000417$ (mała amplituda)
- $p_0 = 1.8$, wówczas $A_{th} \approx 0.712867\text{p} \approx 2.239539$ (bardzo duża amplituda).

Do estymacji amplitudy danej dyskretyzacji wykorzystano następującą procedurę: jeżeli j_m jest lokalnym maksimum trajektorii dyskretnej (tj. $\text{j}_m > \text{j}_{m-1}$

i $j_m > j_{m+1}$), wówczas utożsamialiśmy maksimum aproksymowanej funkcji z maksimum paraboli dopasowanej do następujących 5 punktów: j_{m-2}, j_{m-1}, j_m, j_{m+1}, j_{m+2}. Analogiczną procedurę stosowano w przypadku lokalnych minimów (wykorzystywany był wówczas moduł otrzymanego minimum). W ten sposób otrzymywano ciągi A_N amplitud. Indeks N jest wspólny dla wszystkich ekstremów (minimów i maksimów) i na niektórych wykresach oznaczany jest przez $N_{1/2}$ (liczba połówek okresu) dla odróżnienia od N (liczby okresów).

Każdy badany schemat numeryczny generował trajektorię dyskretną z praktycznie stałą amplitudą. W rzeczywistości amplitudy oscylowały w bardzo regularny sposób wokół pewnej wartości średniej:

$$A_N = A(1+\alpha_N), \tag{4.62}$$

gdzie zarówno średnia amplituda A, jak i bezwymiarowy współczynnik względnych oscylacji mogą zależeć od wartości kroku czasowego ε oraz prędkości początkowej p_0, tzn. $A = A(p_0, \varepsilon)$ i $\alpha_N = \alpha_N(p_0, \varepsilon)$. Oczywiście funkcje te są różne dla różnych schematów numerycznych.

W podobny sposób określano okres ruchu dyskretnego. Dokładna okresowość ($j_{k+n} = j_k$ dla pewnych k, n) jest zjawiskiem rzadkim, które oczywiście nie zostało zaobserwowane. Aby zdefiniować przybliżony okres, możemy dopasować krzywą ciągłą do wykresu dyskretnego, wyznaczyć miejsca zerowe tej funkcji i obliczyć odległości pomiędzy sąsiednimi punktami.

Przypuśćmy, że $j_m j_{m+1} < 0$ dla pewnego m. Oznacza to, że jedno z miejsc zerowych, powiedzmy z_N, leży pomiędzy j_m i j_{m+1}. Jego położenie przybliżamy za pomocą miejsca zerowego wielomianu interpolacyjnego trzeciego stopnia zbudowanego w oparciu o punkty j_{m-1}, j_m, j_{m+1}, j_{m+2} (inna naturalna możliwość to połączenie odcinkiem j_m i j_{m+1}). Jeśli oznaczymy kolejne wyznaczone w powyższy sposób miejsca zerowe przez z_N ($N = 1, 2, 3,\ldots$) ustalając przy tym, że $z_0 = j_0 = 0$, możemy zdefiniować wielkość

$$T_N = z_{2N} - z_{2N-2}, \tag{4.63}$$

którą przyjmiemy za przybliżoną wartość okresu (lokalnego, w punkcie N).

Eksperymenty numeryczne pokazały, że T_N nie jest dokładnie stały, lecz oscyluje ze stosunkowo małą amplitudą. Średnia wartość T_N jest stała z dużą dokładnością, dlatego, podobnie jak w przypadku amplitudy, możemy napisać

$$T_N = T(1+\tau_N), \tag{4.64}$$

gdzie zarówno średni okres T, jak i bezwymiarowy współczynnik jego względnych oscylacji τ_N mogą zależeć od wartości kroku czasowego ε oraz prędkości początkowej p_0, tzn. $T = T(p_0, \varepsilon)$ i $\tau_N = \tau_N(p_0, \varepsilon)$. Tutaj również funkcje te w sposób istotny zależą od rozpatrywanej dyskretyzacji.

Amplituda tych małych oscylacji definiowana jest w sposób naturalny

$$\tau(\varepsilon, p_0) := \max_N |\tau_N(\varepsilon, p_0)|, \quad \alpha(\varepsilon, p_0) := \max_N |\alpha_N(\varepsilon, p_0)|. \tag{4.65}$$

Ponieważ $|\tau_N|$ i $|\alpha_N|$ oscylują (jako funkcje N) z niewielką amplitudą w sposób bardzo regularny, możemy wyznaczać $\tau(p_0, \varepsilon)$ i $\alpha(p_0, \varepsilon)$ rozważając serie, powiedzmy 40 lokalnych ekstremów τ_N i α_N i biorąc wartość średnią.

4.8 Okresowość i stabilność modeli dyskretnych

Trajektorie dyskretne generowane przez schematy symplektyczne lub całkowalne rozważane w tym rozdziale są bardzo stabilne, jeśli krok czasowy ε przyjmuje rozsądne wartości (dla bardzo dużych wartości ε można zaobserwować zachowania chaotyczne [31, 114]). W eksperymentach numerycznych dotyczących amplitud krok czasowy był ograniczany do $\varepsilon \leq 0.5$, natomiast w przypadku okresów stosowano nawet $\varepsilon \approx 1.0$ (dla $p_0 < 1.5$). W obszarze tym ruch dyskretny jest bardzo stabilny, a średni okres T i amplituda A są dobrze zdefiniowane. Średnia amplituda była definiowana po prostu jako

$$A_{avg}(N, M) = \frac{1}{M} \sum_{j=0}^{M-1} \left| A_{N+j} \right|, \tag{4.66}$$

przy czym przyjmowano zwykle $M = 50$. Definicja średniego okresu jest podobna. W wielu przypadkach używano wzoru

$$T_{avg}(N, M) = \frac{1}{M}(z_{N+2M} - z_N), \tag{4.67}$$

w którym pominięto, gwoli zwięzłości, zależność (bardzo istotną) od ε i p_0. Zauważmy, że $T_N \equiv T_{avg}(2N - 2, 1)$. Obliczając T_{avg} należy przyjąć arbitralnie wartość M, zazwyczaj stosowano $M = 20$. Czasami stosowano oznaczenie $N \equiv N_0$, aby zaznaczyć, że uśrednianie przebiega po indeksach większych od N_0.

W sytuacjach, gdy przedmiotem zainteresowania był tylko średni okres (początkowy lub po upływie wielu tysięcy okresów) i należało go uwolnić od wpływu opisywanych wcześniej małych, regularnych oscylacji, stosowano inną jego definicję. Uśredniano mianowicie $T_{avg}(N, M)$ po pewnym zakresie parametru M ($K < M \leq L$):

$$\overline{T}_{avg}(N, K, L) = \frac{1}{L - K} \sum_{M=K+1}^{L} T_{avg}(N, M). \tag{4.68}$$

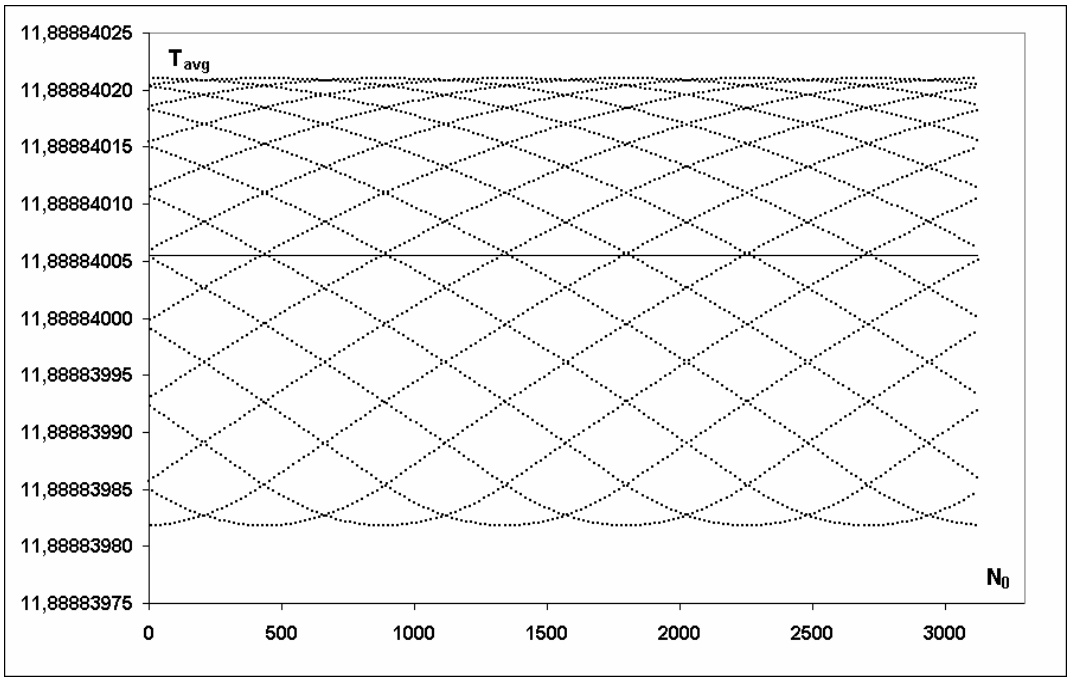
Zazwyczaj wybierano $K = 100$, $L = 200$, co w wystarczającym stopniu eliminowało oscylacje $T_{avg}(N, M)$.

Wszystkie dyskretyzacje rozważane w tym rozdziale charakteryzują się bardzo wysoką stabilnością okresu i amplitudy niezależną od sposobu

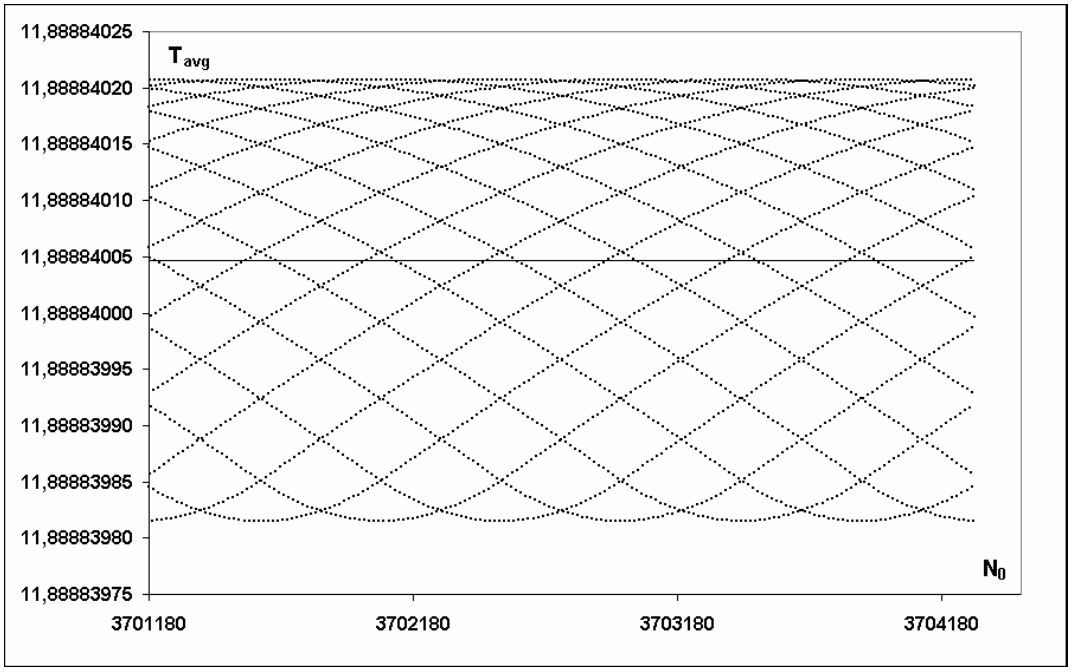
uśredniania. Trudno dostrzec jakąkolwiek zależność A_{avg} i T_{avg} (i tym bardziej $\overline{T}_{avg}$) od N nawet przy bardzo dużych N (takich jak 10^3, 10^5 czy 10^6).

Jako typowy przykład prezentujemy zachowanie schematu Sur1 dla bardzo długich czasów (wykresy 4.1 i 4.2), gdzie użyto definicji (4.67) z $M = 20$. Zależność czasowa tak zdefiniowanego średniego okresu daje interesujący, okresowy wzór geometryczny (można z niej wyodrębnić różne „krzywe dyskretne"). Pochodzenie tego typu wzorów zostanie wyjaśnione w następnym podrozdziale. Ciekawy jest efekt związany ze zmianą M. Otóż dla różnych wartości M otrzymujemy bardzo podobne wzory na wykresach z malejącą amplitudą towarzyszącą wzrostowi M (porównajmy wykres 4.3, gdzie $M = 1$ z 4.2, gdzie $M = 20$).

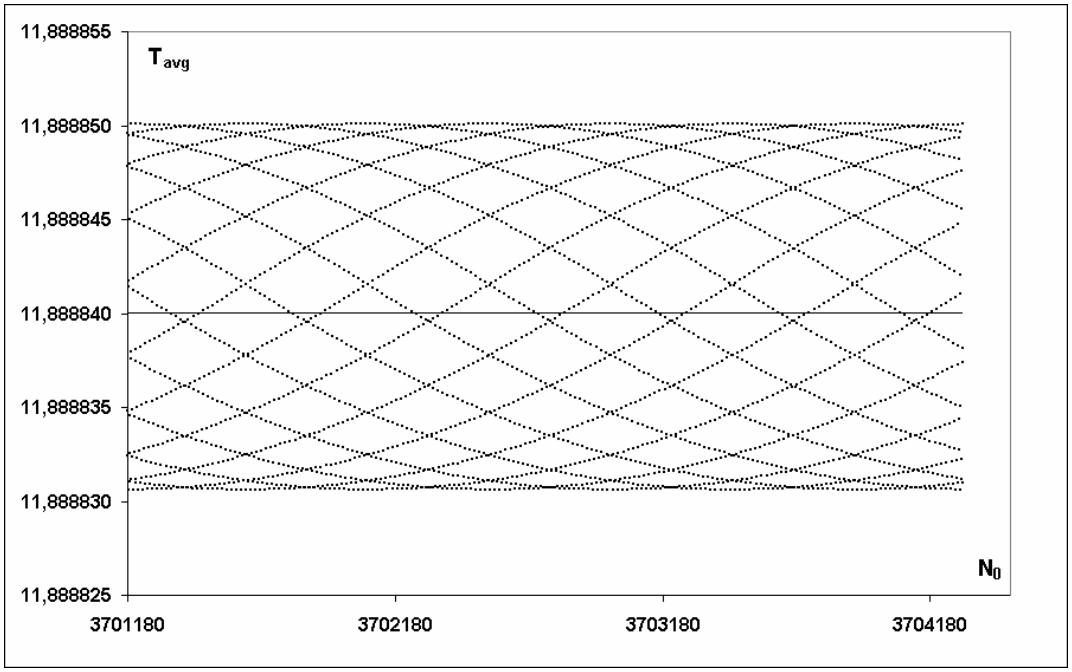
Tabela 4.1 pokazuje jak stabilne są okresy oscylacji. Maksymalny T_N jest zdefiniowany jako $\max_{J+100<N\leq J+200} T_N$ dla $J = 0$ lub $J \approx 1.8 \times 10^6$. Minimalne wartości i średnia jest wyznaczana w tym samym przedziale. Ponieważ odchylenie standardowe średniej jest rzędu 5.7×10^{-8} (przy błędzie maksymalnym około 10^{-7}), zatem średni okres jest praktycznie stały dla wszystkich badanych dyskretyzacji. Szczególnie wielką stabilnością wyróżnia się schemat Sur1. W jego przypadku wszelkie zmiany wartości okresu mieszczą się całkowicie w granicach błędu i nie zaobserwowano jakiejkolwiek zależności $T_{avg}(N, M)$ od N. Biorąc pod uwagę obserwowaną stabilność okresów przyjmiemy w dalszej części tego rozdziału oznaczenie $T \equiv T_{avg}(0, 20)$.



Wykres 4.1. $T_{avg}(N_0, 20)$ dla dyskretyzacji Sur1 ($N_0 < 3100$), $\varepsilon = 0.2$, $p_0 = 1.95$, $T_{th} = 11.65758528$, $T = 11.88884005$.



Wykres 4.2. $T_{avg}(N_0, 20)$ **dla dyskretyzacji Sur1 (dla bardzo dużych** N_0**),** $\varepsilon = 0.2$**,** $p_0 = 1.95$**,** $T_{th} = 11.65758528$**,** $T = 11.88884005$**.**

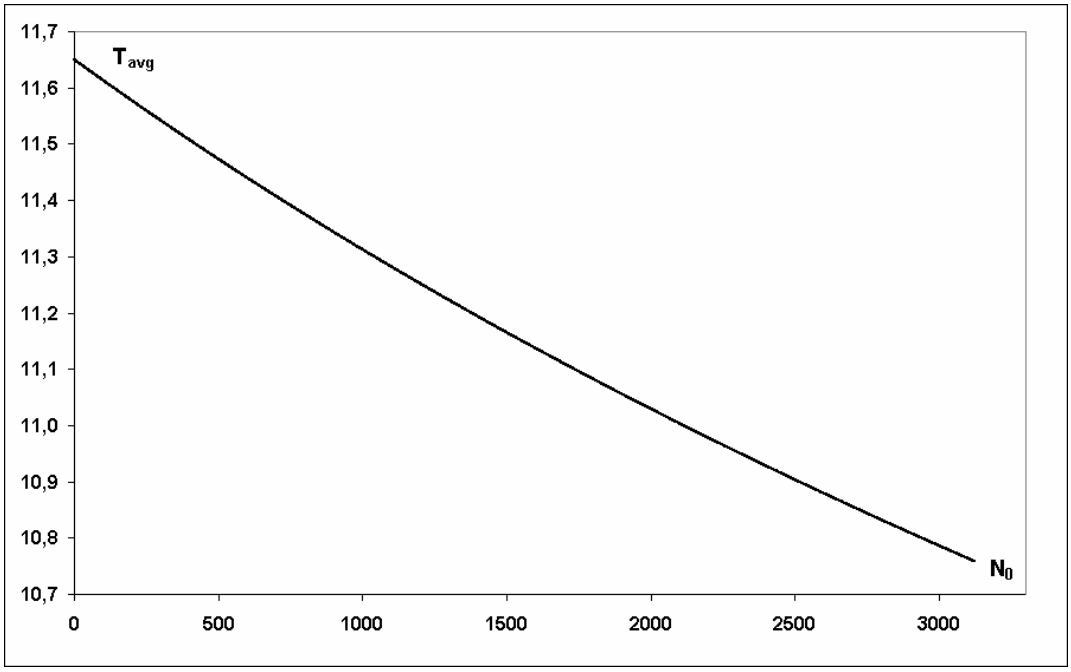


Wykres 4.3. T_N **dla dyskretyzacji Sur1 (dla bardzo dużych** N_0**),** $\varepsilon = 0.2$**,** $p_0 = 1.95$**,** $T_{th} = 11.65758528$**,** $T = 11.88884005$**.**

Tabela 4.1. Stabilność okresu. Minimalna, maksymalna i średnia ($\overline{T}_{avg}(N,100,200)$) wartość T_N. (p_0 = 1.95, ε = 0.2, T_{th} = 11.65758528).

Okres		LF	Sur1	Sur2	GR
Początkowy $N < 100$	Maksymalny	11,93166041	11,88885008	11,91074473	11,64698500
	Średni	11,93165174	11,88884005	11,91073493	11,64697732
	Minimalny	11,93164145	11,88883061	11,91072454	11,64697157
Końcowy $N \approx 1.8\times10^6$	Maksymalny	11,93166040	11,88885008	11,91074470	11,64698540
	Średni	11,93165162	11,88884001	11,91073482	11,64697764
	Minimalny	11,93164140	11,88883061	11,91072450	11,64697190

Obserwowana stabilność okresu (dyskretyzacji symplektycznych i całkowalnych) ostro kontrastuje z wynikami otrzymywanymi za pomocą metod standardowych (niesymplektycznych i niecałkowalnych). Przykładowo, bardzo popularna (jawna) metoda 4-go rzędu Rungego-Kutty daje okres wyraźnie malejący w czasie (wykres 4.4). Przy małych wartościach N_0 dostajemy dobre przybliżenie okresu (interpolacja krzywej dyskretnej daje T = 11.64602 dla N_0 = 0, co jest bliskie teoretycznej wartości T_{th} = 11.65758528). Wśród innych badanych dyskretyzacji tylko dwa schematy gradientowe mogły z tą dokładnością konkurować (a nawet być nieco lepsze) – metoda dyskretnego gradientu daje T = 11.64698. Niestety, przy rosnących N_0 schemat Rungego-Kutty coraz gorzej przybliża okres drgań (w rzeczywistości otrzymujemy wykładnicze, choć bardzo powolne jego zmniejszanie), podczas gdy obydwie metody gradientowe pracują ze stałym okresem przez bardzo długi okres czasu (tabela 4.1). W przypadku parametrów zastosowanych na powyższych wykresach (p_0 = 1.95; ε = 0.2), błąd metody Rungego-Kutty staje się większy od błędów wszystkich prezentowanych metod poczynając od $N_0 \approx 2000$.

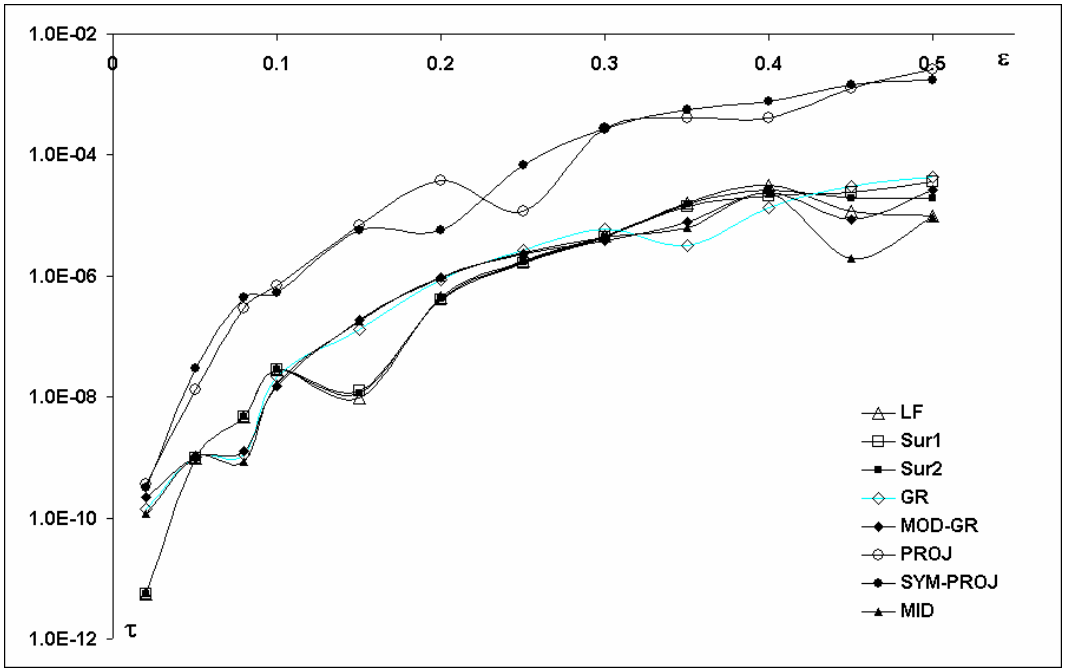


Wykres 4.4. $T_{avg}(N_0, 20)$ dla metody 4-go rzędu Rungego-Kutty, $\varepsilon = 0.2$, $p_0 = 1.95$, $T_{th} = 11.65758528$.

Eksperymenty numeryczne pokazują, że oscylacje okresu i amplitudy są bardzo małe. W granicy $\varepsilon \to 0$ otrzymujemy $\tau(\varepsilon, p_0) \to 0$ z dokładnością do błędu zaokrąglenia. Największe wartości $\tau(\varepsilon, p_0)$ otrzymano dla dwóch metod projekcyjnych (przy dużych ε i małych p_0) – były one rzędu 0.2. Wszystkie inne dyskretyzacje dają oscylacje mniejsze o jeden lub dwa rzędy wielkości (nawet dla dużych ε). Typowy obraz przebiegu funkcji $\tau(\varepsilon, p_0)$ prezentowany jest na wykresie 4.5 dla $p_0 = 1.8$.

4.9 Dlaczego okres i amplituda oscylują w bardzo regularny sposób?

W szerokim zakresie parametrów oscylacje τ_N są bardzo regularne i ich amplituda jest większa niż błędy numeryczne o kilka rzędów wielkości. Zjawisko to okazało się systematycznym, ubocznym efektem numerycznym.



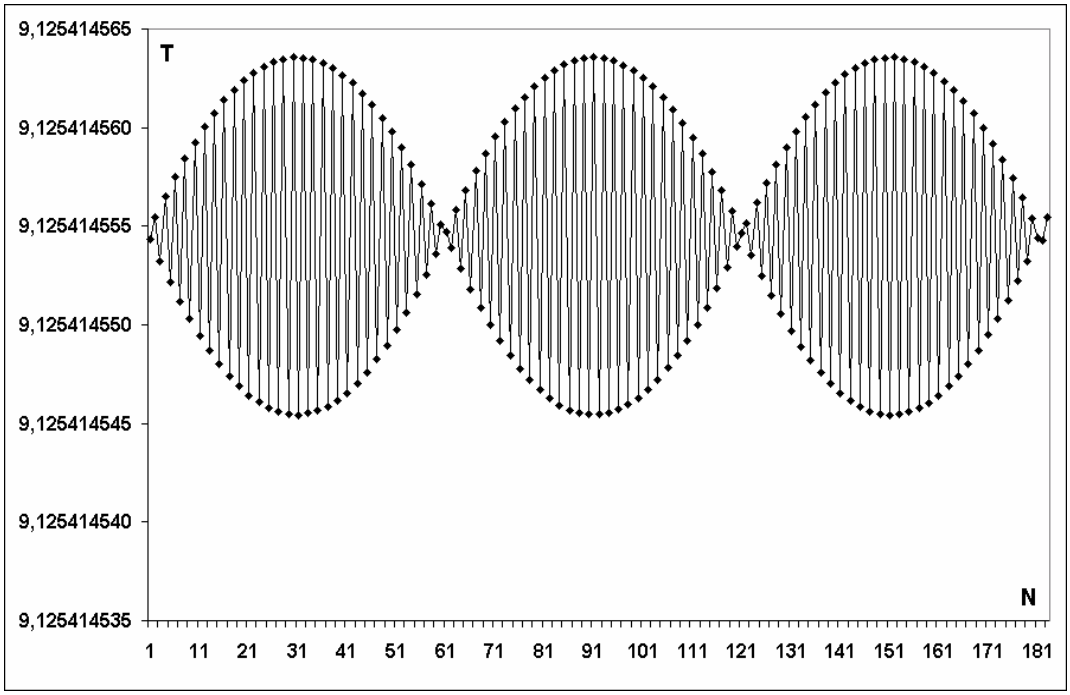
Wykres 4.5. Względne amplitudy oscylacji okresów (τ) dla p_0 = 1.8. (Sur1 i LF przybierają prawie takie same wartości, podobnie jest z dyskretyzacjami MID, GR i MOD-GR, szczególnie dla 0.1 < ε < 0.3).

Wyjaśnienie jest związane z przedstawioną wcześniej procedurą estymowania miejsc zerowych dyskretyzacji. W ogólności okres $T \equiv T_{avg}$ i ε są niewspółmierne, dlatego też względna pozycja z_N pomiędzy j_m i j_{m+1} zależy od N. Okresowe oscylacje, które można obserwować na wykresach od 4.6 do 4.11 są związane z własnościami liczby rzeczywistej T/ε, a konkretnie z aproksymacją liczb T/ε i $T/(2\varepsilon)$ przy pomocy liczb wymiernych.

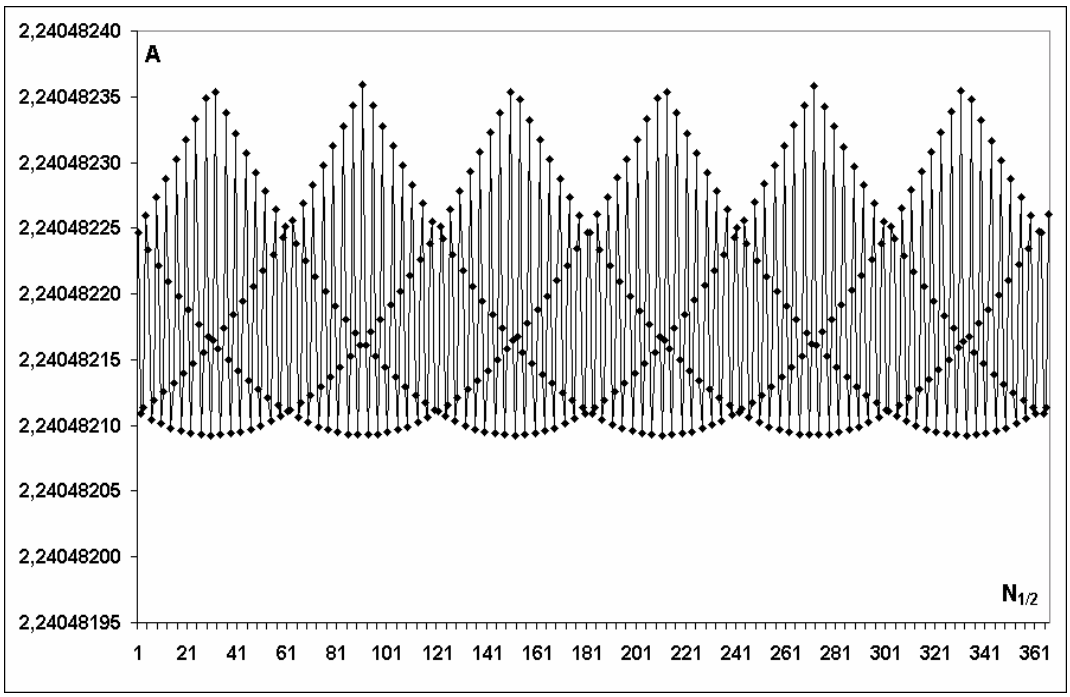
Rozpoczniemy od kilku prostych definicji. Dla danych T, $\varepsilon \in \boldsymbol{R}$ ($T > \varepsilon > 0$) i $K \in \boldsymbol{N}$ zdefiniujmy

$$\mu_K := \frac{KT}{\varepsilon} - M_K, \quad \nu_K := \frac{KT}{2\varepsilon} - L_K, \tag{4.69}$$

takie, że $-0.5 < \mu_K \le 0.5$, $-0.5 < \nu_K \le 0.5$ oraz M_K, $K_K \in \boldsymbol{N}$. Innymi słowy, dla danego K przyjmujemy M_K takie, że M_K/K jest najlepszym wymiernym przybliżeniem (przy danym mianowniku K) liczby rzeczywistej T/ε, i L_K/K jest najlepszym przybliżeniem wymiernym (przy mianowniku K) $T/(2\varepsilon)$. Dla danych T, ε, K formuły (4.69) definiują jednoznacznie μ_K, ν_K, M_K, L_K. Z powyższych definicji wynika bezpośrednio następujący lemat.



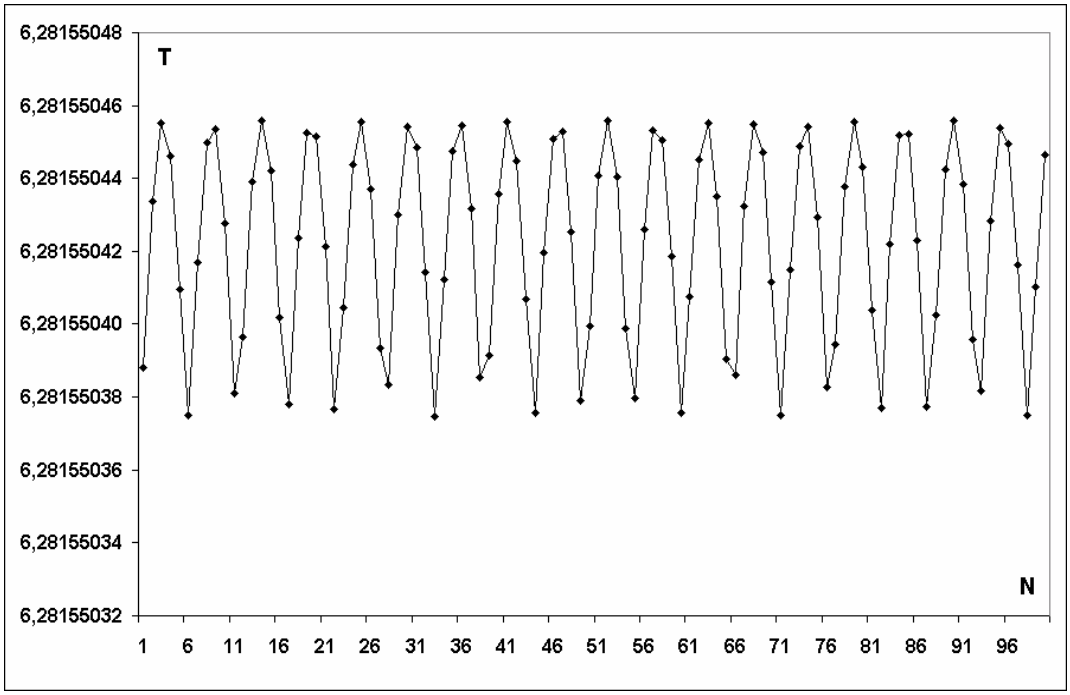
Wykres 4.6. T_N dla metody LF, $\varepsilon = 0.05$, $p_0 = 1.8$, $T = 9.1254145545$.



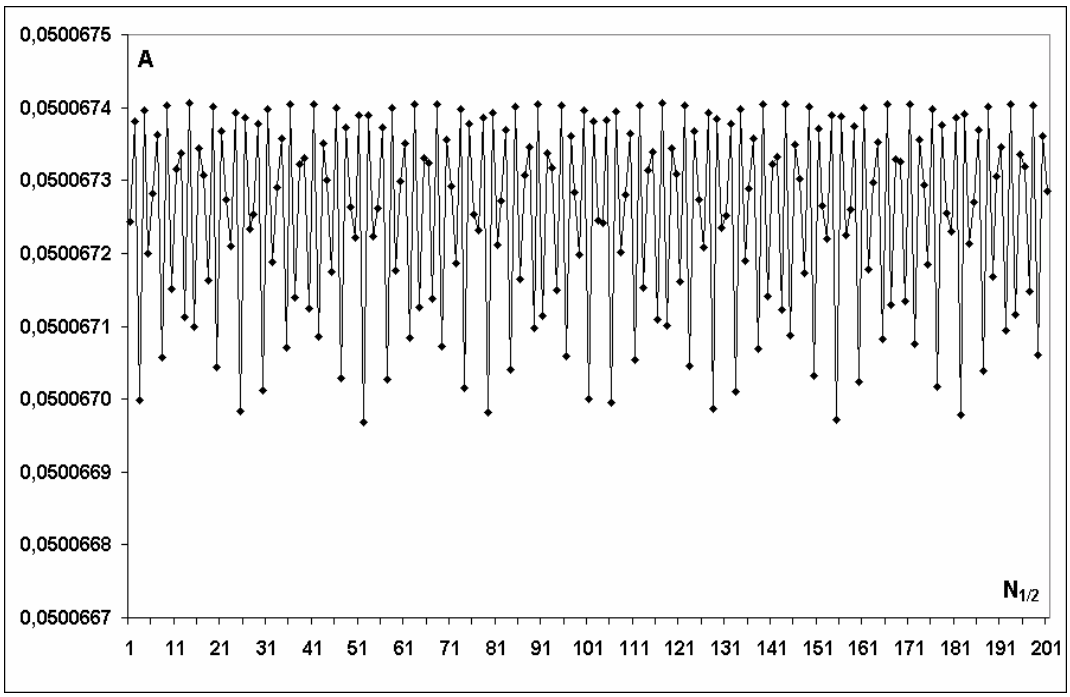
Wykres 4.7. A_N dla metody LF, $\varepsilon = 0.05$, $p_0 = 1.8$, $T = 9.1254145545$.

Lemat 3. Przypuśćmy, że $T > \varepsilon > 0$, wówczas:

1. Jeżeli $|\mu_K + \mu_J| < 0.5$, wówczas $M_{K+J} = M_K + M_J$ i $\mu_{K+J} = \mu_K + \mu_J$.
2. Jeżeli $|\nu_K + \nu_J| < 0.5$, wówczas $L_{K+J} = L_K + L_J$ i $\nu_{K+J} = \nu_K + \nu_J$.
3. Jeżeli $|\nu_K| < 0.25$, wówczas $M_K = 2L$ i $\nu_K = 2\nu_K$.
4. Jeżeli K jest parzyste, to $M_{K/2} = L_K$ i $\mu_{K/2} = \nu_K$.



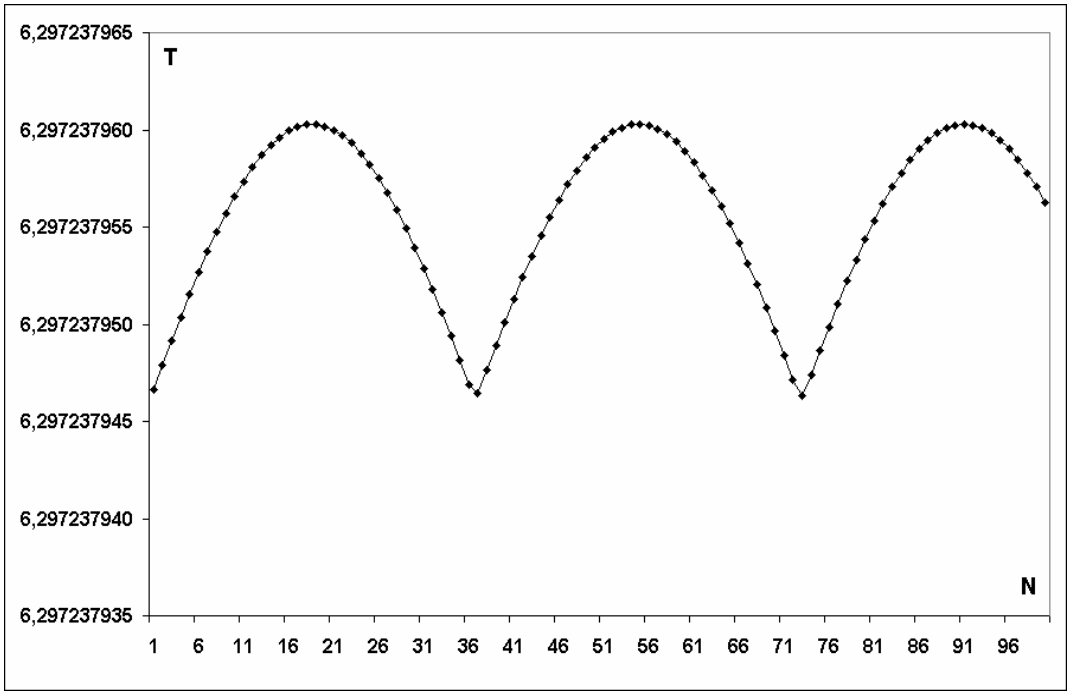
Wykres 4.8. T_N dla metody LF, $\varepsilon = 0.1$, $p_0 = 0.05$, $T = 6.2815504224$.



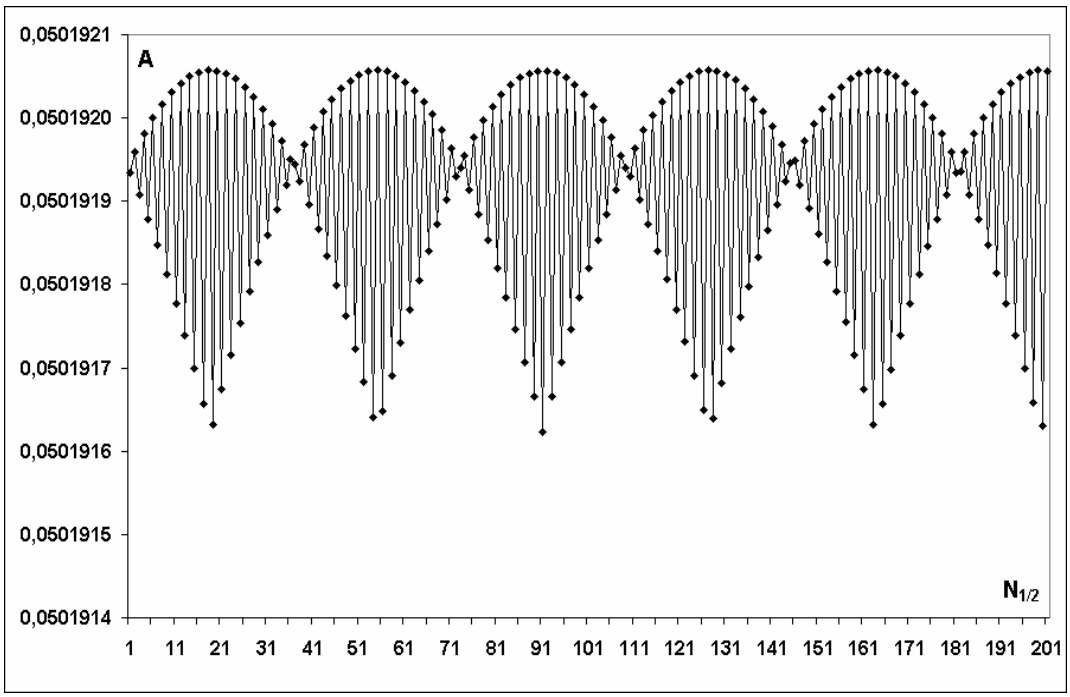
Wykres 4.9. A_N dla metody LF, $\varepsilon = 0.1$, $p_0 = 0.05$, $T = 6.2815504224$.

Wniosek 1. Jeżeli $\nu_K \approx 0$, to $\mu_K \approx 0$ i, dla parzystych K, również $\mu_{K/2} \approx 0$.

Jeżeli $\mu_K \approx 0$, wówczas konfiguracja $z_N,\ j_m,\ j_{m+1}$ w praktyce powtarza się po każdych K okresach. Dlatego naturalnym jest oczekiwanie pewnych cyklicznych zdarzeń z okresem KT. W szczególności $\tau_{N+K} \approx \tau_N$ dla dowolnego N.



Wykres 4.10. T_N dla metody Sur1, $\varepsilon = 0.1$, $p_0 = 0.05$, $T = 6.297327955$.



Wykres 4.11. A_N dla metody Sur1, $\varepsilon = 0.1$, $p_0 = 0.05$, $T = 6.297327955$.

Aby otrzymać „dobrą” aproksymację zazwyczaj żądamy przynajmniej $\mu_K < 0.01$. Czasami, szczególnie dla małych K (tj. $K \leq 5$), ciekawe efekty mogą wystąpić również dla większych μ_K (ale w każdym razie μ_K <0.1): wykres funkcji $N \to T_N$ najwyraźniej rozszczepia się na K „krzywych dyskretnych” (T_N i T_M należą do tej samej krzywej jeżeli $N = M \pmod K$).

Podobne rozważania można przeprowadzić dla oscylacji α_N amplitudy. W tym przypadku okres wynosi $T/2$ i „dobra” aproksymacja odpowiada $\nu_K \approx 0$.

Przykład 1 (schemat LF, $\varepsilon = 0.05$, $p_0 = 1.8$, $T \approx 9.1254146$). Obliczamy $T/\varepsilon \approx 182.508291$ i sprawdzamy, że $\mu_2 \approx 0.017$, $\mu_{59} \approx -0.011$, $\mu_{61} \approx 0.0058$, $\mu_{120} \approx -0.0051$, $\mu_{181} \approx 0.00067$. Wykres 4.6 potwierdza istnienie charakterystycznych wzorów oscylacji o okresach 2, 120 i 181.

Okres 2 odpowiada przejściom między dwiema krzywymi podobnymi do sinusoidy. Mianowicie T_N należy do pierwszej „sinusoidy” dla nieparzystego N i do drugiej „sinusoidy” dla parzystego N. Obie krzywe dyskretne zmieniają się cyklicznie z okresem 120. Właściwe cały wykres wydaje się mieć symetrię względem przesunięć o 60. Różnica pomiędzy T_{N+60} a T_N jest jednak duża (w tym sensie 60 nie jest okresem), jednak T_N leży pomiędzy T_{N+59} a T_{N+61}.

Następny okres, 181, jest trudniejszy do zaobserwowania i odpowiada subtelniejszym efektom, takim jak konfiguracje punktów w pobliżu przecięcia dwóch „sinusoid”, które powtarzają się, w przybliżeniu, co trzy półokresy „sinusoidy”.

Podobnie obliczamy $\nu_4 \approx 0.017$, $\nu_{59} \approx -0.0054$, $\nu_{181} \approx 0.00034$, $\nu_{240} \approx -0.0051$. Na wykresie 4.7 widzimy cztery krzywe dyskretne powtarzające się z okresem 240. Cały wykres ma okres 60, lecz przypatrując się bliżej pewnym detalom (np. obszarom w pobliżu przecięć) możemy zauważyć konfiguracje powtarzające się z okresem 181.

Na koniec podkreślmy, że spełnione są wszystkie równości sugerowane w lemacie 3 (tj. $\mu_{61} = \mu_2 + \mu_{59}, \nu_{240} = \nu_{59} + \nu_{181}, \mu_4 = \nu_2$ itd.)

Przykład 2 (schemat LF, $\varepsilon = 0.1$, $p_0 = 0.05$, $T \approx 6.28155042$). $T/\varepsilon \approx 62.815504$ i sprawdzamy, że $\mu_5 \approx 0.078$, $\mu_{11} \approx -0.029$, $\mu_{27} \approx 0.019$, $\mu_{38} \approx -0.011$, $\mu_{65} \approx 0.0078$ $\mu_{103} \approx -0.0031$. Wykres 4.8 nie sprawia wrażenia równie regularnego jak 4.6. Zwróćmy uwagę, że μ_K są stosunkowo duże – pierwszy μ_K mniejszy niż 0.01 ma indeks $K = 65$, a następny $K = 103$. Bliższa analiza ujawnia podobne cechy na obu wykresach. Znajdujemy podobne do sinusoid krzywe (powtarzające się z okresem 65). Odstęp między nimi wynosi 13, jednak różnica pomiędzy T_{N+13} a T_N jest duża. Zwróćmy uwagę, że okres $103 \approx 8 \times 13$, więc w praktyce punkty tylko co ósmej „sinusoidy” się powtarzają.

Inne okresy ($K = 11, 27, 38$) mogą być wyprowadzone ze 103 i 65. Mianowicie $38 = 103 - 65$, $27 = 65 - 38$, $11 = 38 - 27$. Można je dostrzec na wykresie 4.8. Przykładowo najniższe punkty (T_N pomiędzy 6.28155037 i 6.28155038) mają $N = 6, 17, 22, 33, 44, 49, 60, 71, 82, 87, 98$; odstępy między nimi są dane przez $\Delta N = 11, 5, 11, 11, 5, 11, 11, 11, 5, 11$ (zauważmy, że $11 + 11 + 5 = 27$).

Aby wyjaśnić regularności wykresu 9 obliczamy $\nu_5 = 0.039$, $\nu_{22} = -0.029$, $\nu_{27} = 0.0093$, $\nu_{49} = -0.020$, $\nu_{76} = -0.011$, $\nu_{103} = -0.0015$, $\nu_{130} = 0.0078$ oraz $\nu_{645} = 0.00010$. W tym przypadku struktura wzoru jest dość skomplikowana, ponieważ mamy kilku kandydatów na okres. Kilku z nich pozwala na jasną interpretację. Łącząc co piąty punkt otrzymamy pięć sinusoidalnych krzywych z okresem 130. Dystans pomiędzy sąsiadującymi „sinusoidami" wynosi 26, co jest bardzo bliskie okresowi 27. Późniejsze minima wypadają w punktach $N = 3, 25, 52, 79, 106, 128, 155, 182, 209$, dlatego $\Delta N = 22, 27, 27, 27, 22, 27, 27, 27$ (zauważmy, że $|\nu_{22}|$ jest również stosunkowo małe). Przyglądając się punktom w pobliżu każdego minimum możemy zauważyć cykliczność z okresem 103.

Przykład 3 (schemat Sur1, $\varepsilon = 0.1$, $p_0 = 0.05$, $T \approx 6.29723795$). W tym przypadku przebieg oscylacji okresu prezentowany na wykresie 4.10 jest wyjątkowo prosty (pojedyncza krzywa dyskretna). Może być on wyjaśniony przez brak jakichkolwiek „krótkich" okresów. Najkrótszy z nich, wyraźnie widoczny na wykresie 4.10 wynosi 36. Porównajmy to z wartościami $\mu_{36} \approx 0.0057$, $\mu_{145} \approx 0.0057$, $\mu_{181} \approx 0.00069$. Okres 181 jest nawet bardziej dokładny niż 36 (μ_{181} jest znacznie mniejsze od μ_{36}). Dlatego co pięć bazowych okresów ($181 \approx 5 \times 36$) cykliczność krzywej się poprawia.

Wykres 4.11 składa się z dwóch przecinających się krzywych dyskretnych (okresowych z okresem 72) gdyż $\nu_2 = -0.028$ jest stosunkowo małe i $\nu_{72} = 0.0057$. Istotne jest, że $\nu_3 = 0.46$ jest dużo większe niż $|\nu_2|$. Zauważmy, że $\mu_2 = -0.055$ również nie jest zbyt duże, ale $\mu_3 = -0.083$ jest tego samego rzędu. Cała struktura ma okres 36, ale (podobnie jak w przykładzie 1) różnica pomiędzy A_{N+36} a A_N jest całkiem duża; A_N jest bliskie A_{N+35} i A_{N+36} ($\nu_{35} = 0.017$, $\nu_{37} = -0.011$). Co więcej, mamy całkiem dokładny okres 181 ($\nu_{181} = 0.00034$). Tę okresowość można zauważyć obserwując minima lub punkty, gdzie krzywe dyskretne się przecinają.

Podobne uwagi dotyczą wykresu 4.3, gdzie $T \approx 11.88884005$ oraz $\mu_9 = -0.0044$, $\mu_{448} = 0.0035$, $\mu_{457} = -0.000093$. Wzór przedstawiony na tym wykresie składa się z dziewięciu krzywych dyskretnych i powtarza się cyklicznie z okresem bliskim 457.

Zachowanie opisane na powyższych przykładach jest typowe. Podobne efekty okresowe można obserwować dla innych dyskretyzacji i innych wyborów parametrów z wyjątkiem bardzo małych wartości ε (np. $\varepsilon \leq 0.01$), gdy systematyczne oscylacje stają się porównywalne lub mniejsze od błędów zaokrąglenia, co sprawia, że stają się chaotyczne.

4.10 Numeryczne szacowanie amplitudy i okresu

Wszystkie dyskretyzacje analizowane w tym rozdziale charakteryzują się bardzo wysoką stabilnością trajektorii. Dzięki temu wielkości takie jak średni okres i średnia amplituda (w przypadku ruchu oscylacyjnego) są dobrze zdefiniowane dla każdej dyskretyzacji (zakładając, że rozważamy trajektorie wystarczająco oddalone od separatrysy i ε nie jest zbyt duże (wystarczy przyjąć, że $\varepsilon \leq 0.5$).

4.10.1 Średnia amplituda

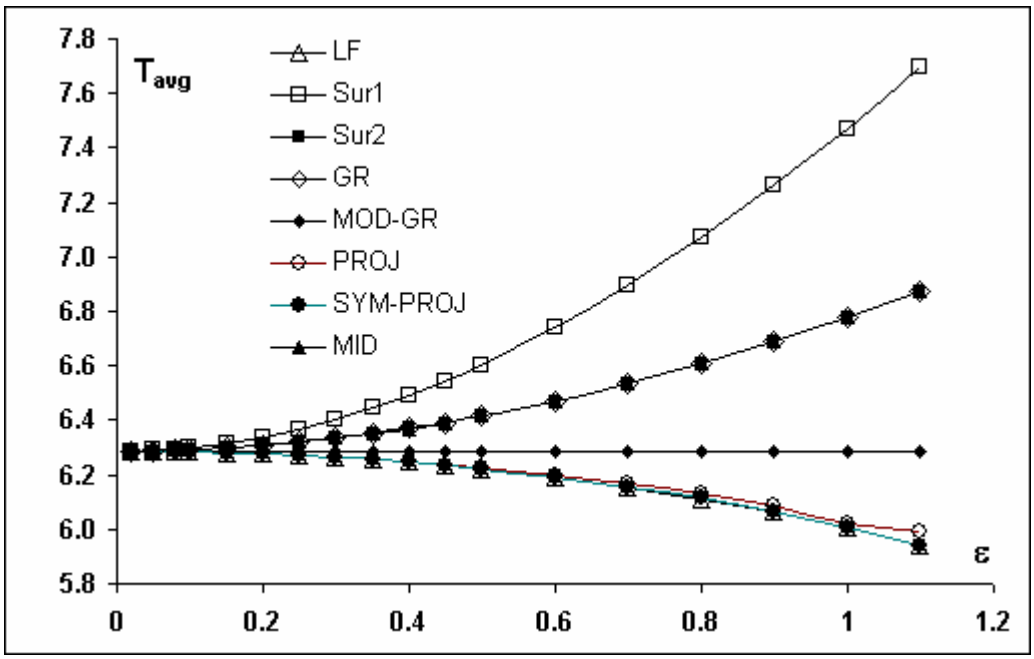
Względne odchylenia średnich amplitud od teorii znajdują się w tabeli 4.2 (dla $\varepsilon = 0.02$ i $\varepsilon = 0.5$). Zostały one obliczone jako różnice pomiędzy wartościami numerycznymi i amplitudami teoretycznymi wynikającymi z warunków początkowych ruchu.

Tabela 4.2. Względne odchylenie średniej amplitudy od teorii ($A = A_{avg}(0, 50)$).

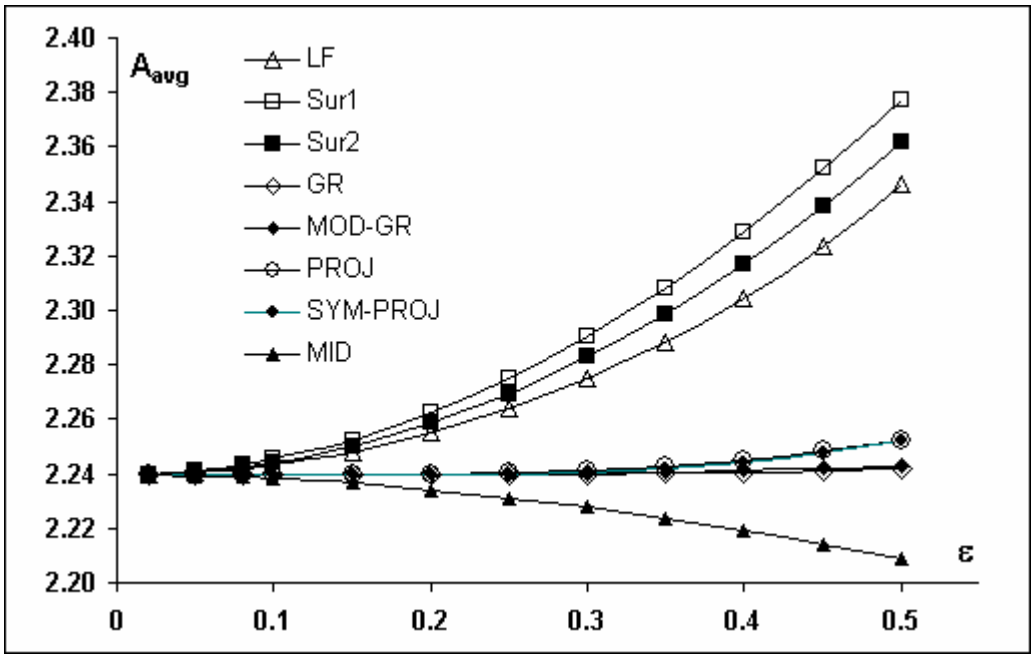
p_0	LF	Sur1	Sur2	GR	MOD-GR	PROJ	SYM-PROJ	MID
ε = 0,02								
0,05	5,00E-05	1,50E-04	1,00E-04	-1,86E-08	-1,87E-08	-1,68E-08	-1,71E-08	-2,89E-08
0,1	5,00E-05	1,50E-04	9,99E-05	-1,85E-08	-1,84E-08	-1,10E-08	-1,21E-08	-6,01E-08
0,3	5,00E-05	1,48E-04	9,92E-05	-1,74E-08	-1,68E-08	4,58E-08	5,09E-08	-3,95E-07
0,5	5,00E-05	1,46E-04	9,79E-05	-1,55E-08	-1,56E-08	1,69E-08	-8,59E-09	-1,08E-06
0,8	5,02E-05	1,39E-04	9,47E-05	-9,03E-09	-8,37E-09	-1,46E-07	-2,05E-07	-2,84E-06
1,2	5,13E-05	1,26E-04	8,86E-05	-3,85E-09	-3,88E-09	-4,55E-09	2,18E-09	-7,00E-06
1,6	5,66E-05	1,08E-04	8,24E-05	2,71E-09	2,68E-09	5,42E-10	1,33E-08	-1,53E-05
1,8	6,73E-05	1,02E-04	8,48E-05	4,07E-09	3,96E-09	4,56E-09	4,10E-09	-2,49E-05
ε = 0,5								
0,05	2,54E-02	8,52E-02	5,58E-02	-6,34E-03	-6,87E-03	-3,15E-02	-3,16E-02	-6,35E-03
0,1	2,55E-02	8,52E-02	5,57E-02	-6,32E-03	-6,80E-03	-3,05E-02	-3,14E-02	-6,34E-03
0,3	2,58E-02	8,46E-02	5,56E-02	-6,07E-03	-6,59E-03	-2,55E-02	-2,76E-02	-6,27E-03
0,5	2,65E-02	8,34E-02	5,54E-02	-5,72E-03	-6,13E-03	-1,44E-02	-2,13E-02	-6,30E-03
0,8	2,81E-02	8,05E-02	5,46E-02	-4,58E-03	-5,01E-03	8,86E-03	-5,54E-03	-6,12E-03
1,2	3,07E-02	7,42E-02	5,32E-02	-2,44E-03	-2,69E-03	3,40E-02	1,95E-02	-6,54E-03
1,6	3,84E-02	6,52E-02	5,19E-02	1,80E-04	1,46E-04	9,31E-03	1,76E-02	-9,00E-03
1,8	4,76E-02	6,14E-02	5,46E-02	1,22E-03	1,31E-03	5,56E-03	5,70E-03	-1,36E-02

Widzimy od razu, że w obu przypadkach najlepsze wyniki dają obydwa schematy gradientowe (a najgorsze Sur1 i Sur2). Względny błąd dyskretyzacji LF i metod Surisa praktycznie nie zależy od p_0. Dokładność metod gradientowych wzrasta dla dużych p_0 zarówno dla $\varepsilon = 0.02$, jak i $\varepsilon = 0.5$. Przy małych wartościach ε (np. $\varepsilon = 0.02$) również schematy rzutowane dają bardzo małe odchylenia rzędu 10^{-8} lub 10^{-9} (podobne do metod gradientowych) przy czym ich dokładność zależy od p_0. Przy niektórych wartościach p_0 (np. p_0 =1.6)

ich dokładność jest bardzo duża, podczas gdy przy innych (np. p_0 = 0.8) – wyraźnie mniejsza.



Wykres 4.12. $T_{avg} \equiv \overline{T}_{avg}(0, 100, 200)$ **jako funkcja** ε **dla** p_0 **= 0.1, (**T_{th} **= 6.28711782). (Sur2 i GR dają praktycznie te same rezultaty, MOD-GR jest bardzo bliski teorii).**



Wykres 4.13. $A_{avg} \equiv A_{avg}(0, 50)$ **jako funkcja** ε **dla** p_0 **= 1.8, (**A_{th} **= 2.239539). GR daje praktycznie te same wartości co MOD-GR, podobnie jest w przypadku dwóch metod rzutowanych.**

Schemat *implicit midpoint* jest porównywalny z metodami gradientowymi, jednak tylko przy małych wartościach p_0 ($p_0 < 0.1$). Przy dużych wartościach p_0 ($p_0 > 1.6$) dyskretyzacje LF, obie Surisa i *implicit midpoint* dają znacznie większe odchylenia (nawet cztery rzędy wielkości przy małych wartościach ε).

Dla większych ε (np. $\varepsilon = 0.5$) różnice pomiędzy badanymi metodami są znacznie mniejsze (różnią się najwyżej o dwa rzędy wielkości). Regułą pozostaje najwyższa dokładność metod gradientowych. *Implicit midpoint* dorównuje im w dokładności dla $p_0 < 1.2$, podczas gdy metody rzutowane są prawie zawsze znacznie gorsze z wyjątkiem wybranych wartości p_0 (np. $p_0 = 0.8$). LF i obie metody Surisa mają większe względne odchylenie od teorii w całym zakresie zmienności p_0. Warto zaznaczyć jednakże, że nawet te „duże" odchylenia nie przekraczają kilku procent z wyjątkiem sytuacji, gdy p_0 zbliża się do 2 (wówczas dyskretyzacje te nie są w stanie odtworzyć teorii nawet jakościowo).

Wykres 4.13 ilustruje zależność średniej amplitudy od ε dla $p_0 = 1.8$. Metody gradientowe i (zwłaszcza dla $\varepsilon < 0.3$) metody projekcyjne są najdokładniejsze.

4.10.2 Średni okres

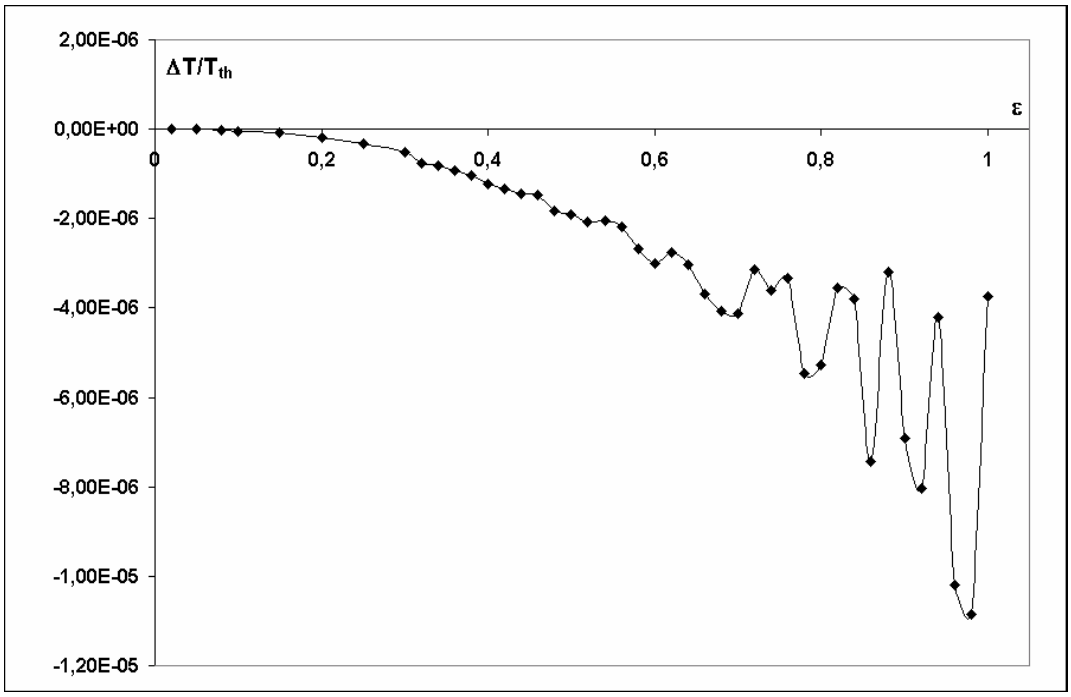
Względne odchylenia od teorii średniego okresu są przedstawione w tabeli 4.3 (dla $\varepsilon = 0.02$ i $\varepsilon = 0.5$). Przy $p_0 < 0.5$ wszystkie dyskretyzacje z wyjątkiem modyfikowanego dyskretnego gradientu mają podobne błędy względne (Sur1 jest najgorszy wśród nich). Modyfikowany gradient jest znacznie lepszy (dla $p_0 \approx 0$ daje odchylenia mniejsze o co najmniej cztery rzędy wielkości) – porównajmy wykresy 4.12 ($p_0 = 0.1$) i 4.14 ($p_0 = 0.02$).

Przy większych wartościach p_0 wszystkie dyskretyzacje mają zbliżoną dokładność z dwoma interesującymi wyjątkami: schematy LF oraz *implicit midpoint* wydają się posiadać „rezonansowe wartości" p_0, przy których ich dokładność jest znacznie lepsza od dokładności wszystkich innych metod. Wykres 4.15 pokazuje, jak dokładny jest schemat LF przy $p_0 = 1.21$ dla praktycznie wszystkich wartości ε. Pokazane tam są również dwie inne dyskretyzacje: *implicit midpoint* i zmodyfikowany dyskretny gradient (kolejne najlepsze przy tej wartości p_0) – znacznie gorsze niż LF. Odchylenia pozostałych dyskretyzacji są jeszcze większe. *Implicit midpoint* posiada analogiczną „rezonansową wartość" $p_0 \approx 1.6$. Warto tu podkreślić dość zaskakujący fakt, że rzutowanie zastosowane do metody LF wpływa negatywnie na dokładność odwzorowania średniego okresu w przedziale $0.8 < p_0 < 1.8$ zwłaszcza dla większych wartości ε (np. $\varepsilon = 0.5$).

Tabela 4.3. Względne odchylenie okresu od teorii ($T_{avg} \equiv \overline{T}_{avg}(0, 100, 200)$).

p_0	LF	Sur1	Sur2	GR	MOD-GR	PROJ	SYM-PROJ	MID
ε = 0,02								
0,02	-1,67E-05	8,33E-05	3,33E-05	3,33E-05	-3,34E-09	-1,66E-05	-1,66E-05	3,33E-05
0,05	-1,66E-05	8,33E-05	3,33E-05	3,33E-05	-2,08E-08	-1,64E-05	-1,65E-05	3,33E-05
0,1	-1,66E-05	8,32E-05	3,33E-05	3,32E-05	-8,34E-08	-1,56E-05	-1,59E-05	3,32E-05
0,3	-1,59E-05	8,18E-05	3,30E-05	3,26E-05	-7,52E-07	-6,74E-06	-1,01E-05	3,24E-05
0,5	-1,45E-05	7,92E-05	3,23E-05	3,12E-05	-2,10E-06	1,11E-05	1,70E-06	3,07E-05
0,8	-1,08E-05	7,28E-05	3,10E-05	2,79E-05	-5,45E-06	5,58E-05	3,16E-05	2,63E-05
1,0	-6,99E-06	6,71E-05	3,01E-05	2,47E-05	-8,63E-06	9,86E-05	6,05E-05	2,20E-05
1,2	-1,48E-06	6,05E-05	2,95E-05	2,07E-05	-1,27E-05	1,53E-04	9,80E-05	1,62E-05
1,4	6,86E-06	5,37E-05	3,03E-05	1,56E-05	-1,77E-05	2,21E-04	1,46E-04	8,30E-06
1,6	2,12E-05	4,92E-05	3,52E-05	9,31E-06	-2,40E-05	3,05E-04	2,07E-04	-3,63E-06
1,8	5,64E-05	5,91E-05	5,77E-05	9,19E-07	-3,24E-05	4,08E-04	2,87E-04	-2,75E-05
1,95	2,17E-04	1,90E-04	2,03E-04	-9,09E-06	-4,24E-05	4,99E-04	3,72E-04	-1,15E-04
1,99	8,96E-04	8,50E-04	8,73E-04	-1,50E-05	-4,83E-05	5,19E-04	4,08E-04	-4,57E-04
$2,0 - 10^{-3}$	7,12E-03	7,06E-03	7,09E-03	-1,95E-05	-5,29E-05	5,14E-04	4,26E-04	-3,40E-03
$2,0 - 10^{-4}$	9,17E-02	9,16E-02	9,16E-02	-2,22E-05	-5,56E-05	5,05E-04	4,34E-04	-2,40E-02
$2,0 - 10^{-5}$				-2,43E-05	-5,58E-05	4,99E-04	4,40E-04	-1,03E-01
$2,0 - 10^{-6}$				-2,80E-05	-5,69E-05	4,94E-04	4,43E-04	-2,13E-01
$2,0 - 10^{-7}$				-7,33E-05	-2,09E-05	4,91E-04	4,46E-04	-3,08E-01
$2,0 - 10^{-8}$				1,38E-04	1,15E-04	4,88E-04	4,48E-04	-3,83E-01
$2,0 - 10^{-9}$				-1,61E-03	1,18E-03	4,86E-04	4,50E-04	-4,43E-01
$2,0 + 10^{-8}$	-4,15E-01	-4,15E-01	-4,15E-01	-5,16E-05	-4,23E-06	2,82E-04	3,32E-04	
$2,0 + 10^{-7}$	-3,44E-01	-3,44E-01	-3,44E-01	-1,59E-05	-6,26E-05	2,60E-04	3,15E-04	
$2,0 + 10^{-6}$	-2,54E-01	-2,54E-01	-2,54E-01	-2,90E-05	-6,44E-05	2,31E-04	2,94E-04	
$2,0 + 10^{-5}$	-1,43E-01	-1,43E-01	-1,43E-01	-2,45E-05	-5,74E-05	1,94E-04	2,67E-04	
2,0001	-4,26E-02	-4,27E-02	-4,27E-02	-2,22E-05	-5,55E-05	8,93E-05	1,76E-04	3,38E-02
2,001	-6,68E-03	-6,73E-03	-6,70E-03	-1,96E-05	-5,29E-05	1,62E-04	2,69E-04	3,49E-03
2,05	-2,44E-04	-2,78E-04	-2,61E-04	-1,14E-05	-4,47E-05	-3,47E-06	1,57E-04	1,14E-04
2,1	-1,46E-04	-1,74E-04	-1,60E-04	-9,25E-06	-4,26E-05	-3,91E-05	1,31E-04	6,62E-05
2,2	-9,25E-05	-1,13E-04	-1,03E-04	-7,02E-06	-4,04E-05	-7,22E-05	1,04E-04	4,10E-05
2,5	-5,71E-05	-6,97E-05	-6,34E-05	-4,20E-06	-3,75E-05	-1,08E-04	6,77E-05	2,54E-05
3	-4,45E-05	-5,18E-05	-4,81E-05	-2,44E-06	-3,58E-05	-1,28E-04	4,18E-05	2,04E-05
5	-3,61E-05	-3,83E-05	-3,72E-05	-7,26E-07	-3,41E-05	-1,43E-04	1,46E-05	1,75E-05
10	-3,40E-05	-3,45E-05	-3,42E-05	-1,70E-07	-3,35E-05	-1,42E-04	9,53E-06	1,69E-05
ε = 0,5								
0,1	-1,06E-02	5,06E-02	2,05E-02	2,04E-02	-5,02E-05	-9,86E-03	-1,03E-02	2,04E-02
0,3	-1,01E-02	4,97E-02	2,03E-02	2,01E-02	-4,53E-04	-3,29E-03	-7,54E-03	1,99E-02
0,5	-9,17E-03	4,80E-02	1,98E-02	1,93E-02	-1,27E-03	1,01E-02	-1,69E-03	1,89E-02
0,8	-6,71E-03	4,40E-02	1,89E-02	1,73E-02	-3,30E-03	4,43E-02	1,42E-02	1,64E-02
1	-4,13E-03	4,02E-02	1,83E-02	1,53E-02	-5,25E-03	7,85E-02	3,12E-02	1,38E-02
1,2	-4,05E-04	3,58E-02	1,79E-02	1,29E-02	-7,74E-03	1,24E-01	5,55E-02	1,03E-02
1,4	5,31E-03	3,11E-02	1,83E-02	9,82E-03	-1,09E-02	1,84E-01	9,04E-02	5,56E-03
1,6	2,40E-02	3,74E-02	3,08E-02	8,57E-03	-2,13E-02	4,11E-01	2,14E-01	-1,91E-03
1,8	4,28E-02	3,27E-02	3,80E-02	6,42E-04	-2,03E-02	3,15E-01	2,19E-01	-1,56E-02
1,95	3,41E-01	1,86E-01	2,56E-01	-5,72E-03	-2,68E-02	2,86E-01	3,06E-01	-5,94E-02
1,99				-9,51E-03	-3,06E-02	2,23E-01	3,31E-01	-1,54E-01
$2,0 - 10^{-3}$				-1,24E-02	-3,36E-02	1,59E-01	3,31E-01	-3,21E-01
$2,0 - 10^{-4}$				-1,41E-02	-3,53E-02	1,19E-01	3,26E-01	-4,48E-01
$2,0 - 10^{-5}$				-1,52E-02	-3,65E-02	9,09E-02	3,22E-01	-5,36E-01

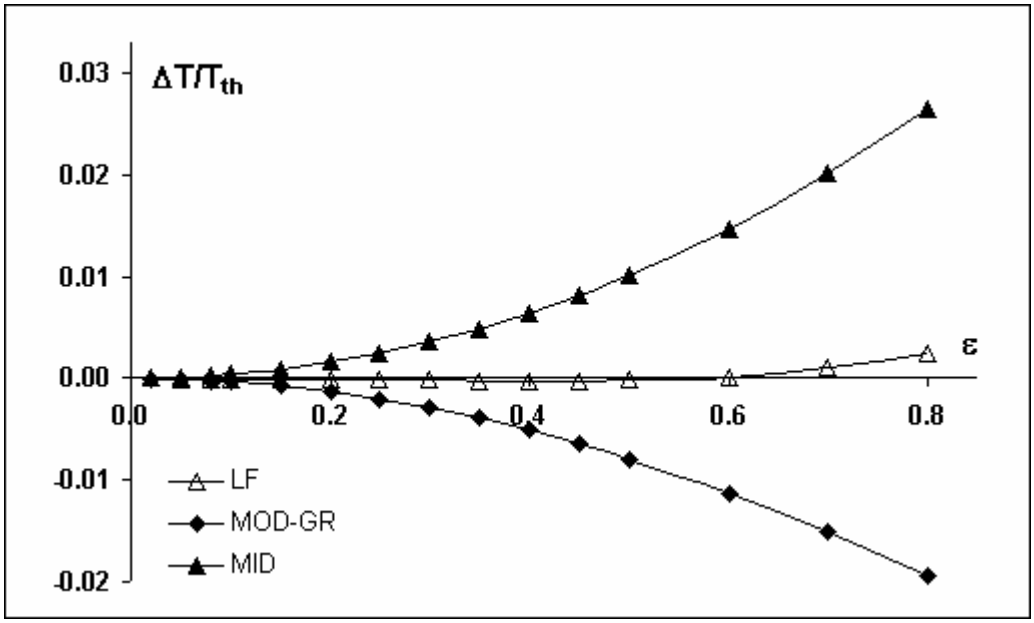
p_0	LF	Sur1	Sur2	GR	MOD-GR	PROJ	SYM-PROJ	MID
$2,0 - 10^{-6}$				-1,61E-02	-3,73E-02	7,16E-02	3,19E-01	-6,01E-01
$2,0 - 10^{-7}$				-1,67E-02	-3,80E-02	5,87E-02	3,17E-01	-6,49E-01
$2,0 - 10^{-8}$				-1,73E-02	-3,86E-02	4,46E-02	3,16E-01	-6,88E-01
$2,0 - 10^{-9}$				-1,73E-02	-3,86E-02	3,55E-02	3,14E-01	-7,18E-01
$2,0 + 10^{-8}$	-7,24E-01	-7,22E-01	-7,23E-01	-1,72E-02	-3,85E-02	-5,68E-02	2,42E-01	
$2,0 + 10^{-7}$	-6,90E-01	-6,88E-01	-6,89E-01	-1,67E-02	-3,80E-02	-5,61E-02	2,35E-01	
$2,0 + 10^{-6}$	-6,47E-01	-6,45E-01	-6,46E-01	-1,61E-02	-3,73E-02	-5,53E-02	2,26E-01	
$2,0 + 10^{-5}$	-5,90E-01	-5,87E-01	-5,89E-01	-1,52E-02	-3,65E-02	-5,41E-02	2,15E-01	
2,0001	-5,12E-01	-5,08E-01	-5,10E-01	-1,41E-02	-3,53E-02	-5,72E-02	1,96E-01	
2,001	-3,97E-01	-3,93E-01	-3,96E-01	-1,24E-02	-3,36E-02	-4,36E-02	1,80E-01	
2,05	-1,17E-01	-1,19E-01	-1,19E-01	-7,20E-03	-2,83E-02	-4,02E-02	1,14E-01	9,20E-02
2,1	-8,11E-02	-8,44E-02	-8,28E-02	-5,86E-03	-2,69E-02	-4,18E-02	9,80E-02	4,62E-02
2,2	-5,57E-02	-5,95E-02	-5,77E-02	-4,45E-03	-2,55E-02	-4,42E-02	8,08E-02	2,70E-02
2,5	-3,68E-02	-3,87E-02	-3,77E-02	-2,68E-03	-2,37E-02	-5,03E-02	5,49E-02	1,64E-02
3	-2,96E-02	-2,96E-02	-2,95E-02	-1,57E-03	-2,25E-02	-5,51E-02	3,34E-02	1,34E-02
5	-2,68E-02	-2,31E-02	-2,50E-02	-5,04E-04	-2,14E-02	-5,54E-02	4,32E-03	1,29E-02
10	-5,09E-02	-5,07E-02	-5,08E-02	-1,82E-04	-2,13E-02	-1,39E-02	-1,23E-02	2,68E-02



Wykres 4.14. Zmodyfikowana metoda dyskretnego gradientu. Względne odchylenie od okresu teoretycznego jako funkcja ε dla $p_0 = 0.02$, $T_{th} = 6.283342395$, $T(\varepsilon) = T_{avg}(0, 30)$.

Gdy p_0 zbliża się do 2, wówczas obie metody gradientowe stają się dokładniejsze od innych badanych schematów (jedynie dla małych ε, metody projekcyjne są lepsze). Dla p_0 bardzo bliskich tej wartości granicznej dokładność wszystkich metod gwałtownie się pogarsza, a schematy LF i obie metody Surisa dają ruch rotacyjny zamiast oscylacji (tabela 4.3). Zachowanie

dyskretyzacji w najbliższym otoczeniu separatrysy (p_0 = 2) jest dyskutowane dalej. W tym miejscu zauważmy jedynie, że dla p_0 nieco większych od 2, metoda *implicit midpoint* nie odtwarza rotacji, lecz dalej oscyluje wykazując się złym zachowaniem jakościowym.



Wykres 4.15. Względne odchylenie od okresu teoretycznego jako funkcja ε dla p_0 = 1.21, T_{th} = 7.01866131087, $T(\varepsilon) = \overline{T}_{avg}(0, 100, 200)$.

W przypadku ruchu rotacyjnego, względne odchylenie średniego okresu od teorii jest podobne dla wszystkich badanych dyskretyzacji z wyjątkiem metody dyskretnego gradientu, która jest lepsza o jeden lub dwa rzędy wielkości.

4.11 Interesujące przypadki szczególne

W tym podrozdziale zostaną krótko omówione pewne zagadnienia, które warto by było poddać w przyszłości dokładniejszej analizie.

Ekstrapolacja $\varepsilon \to 0$

Dla wszystkich badanych dyskretyzacji oczekujemy, że

$$\lim_{\varepsilon\to 0} T(\varepsilon,\, p_0) = T_{th}(p_0), \qquad \lim_{\varepsilon\to 0} A(\varepsilon,\, p_0) = A_{th}(p_0)\,. \tag{4.70}$$

gdzie $T_{th}(p_0)$, $A_{th}(p_0)$ nie zależą od dyskretyzacji i są równe teoretycznej wartości wyznaczonej ze wzorów analitycznych (wyrażonej poprzez funkcje eliptyczne); porównajmy wykresy 4.12 i 4.13.

Poddajmy analizie ilościowej przypadek pokazany na wykresie 4.12 (dokładny okres $T_{th} \approx 6.28711783$). Fitując wielomian trzeciego stopnia (w rzeczywistości bardzo bliski paraboli) na 12 punktach (ε = 0.01, 0.02,..., 0.11, 0.12) dostajemy

$$\begin{aligned} T &= -0{,}03867\varepsilon^3 + 1{,}310512\varepsilon^2 - 0{,}0001050\varepsilon + 6{,}28711875 \text{ (Sur1)}, \\ T &= -0{,}00909\varepsilon^3 + 0{,}524053\varepsilon^2 - 0{,}0000247\varepsilon + 6{,}28711805 \text{ (Sur2)}, \\ T &= -0{,}00475\varepsilon^3 - 0{,}260242\varepsilon^2 - 0{,}0000130\varepsilon + 6{,}28711794 \text{ (LF)} \end{aligned} \qquad (4.71)$$

Ostatnie wyrazy estymują zupełnie dobrze dokładny okres. Przyjmując za jednostkę 10^{-7} wyznaczono ich bezwzględne błędy odpowiednio jako 9.2, 2.2 i 0.9. Są one porównywalne z odchyleniami od teorii dla ε = 0.001 (wynoszącymi odpowiednio 13.1, 5.2 i -2.6). Odchylenia od teorii przy ε = 0.01 (wynoszące odpowiednio 1307.2, 523.3 i 260.6) są większe o dwa rzędy wielkości. Zmodyfikowana metoda dyskretnego gradientu (z poprawką δ) bije pozostałe dyskretyzacje na głowę: jej odchylenie przy ε = 0.01 wynosi tylko 1.3 (w tych samych jednostkach).

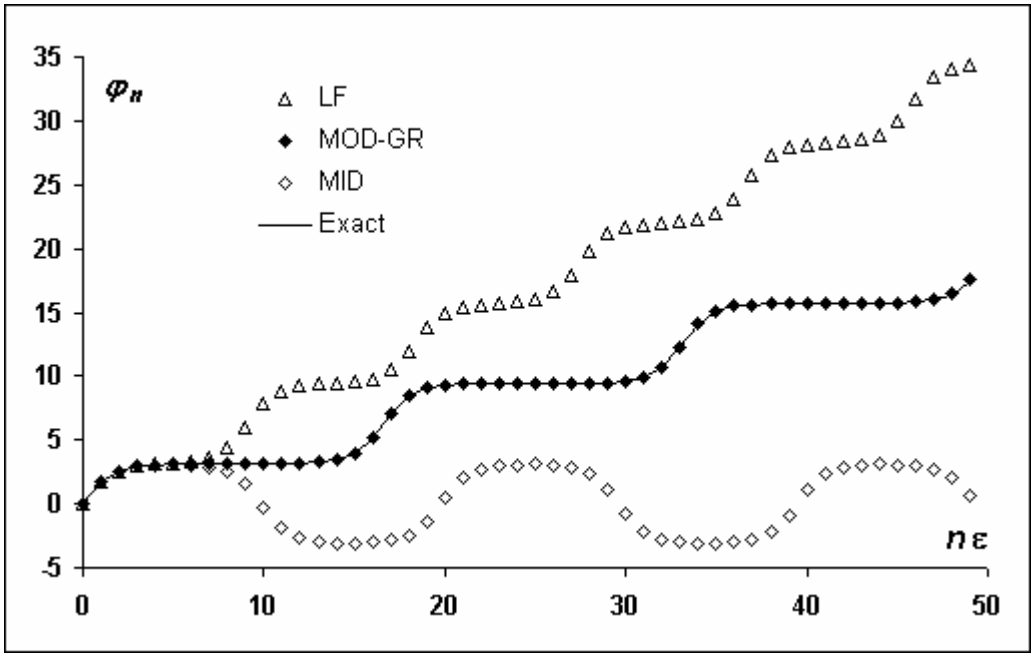
Zachowanie w pobliżu separatrysy

Separatrysa jest granicą pomiędzy ruchem oscylacyjnym i rotacyjnym. Tablica 4.3 prezentuje wartości okresu dla ruchu w pobliżu separatrysy, to znaczy dla $p_0 \approx 2$. Jest to zakres parametrów najtrudniejszy dla dokładnych symulacji numerycznych. Schematy gradientowe i rzutowane dają przyzwoite wyniki, zwłaszcza dla małych ε, i są znacznie lepsze od wszystkich pozostałych metod. Dla ruchu rotacyjnego w pobliżu separatrysy nawet metody rzutowane stają się mniej dokładne i tylko schematy gradientowe dają stosunkowo dobre wyniki ilościowe (tablica 4.3).

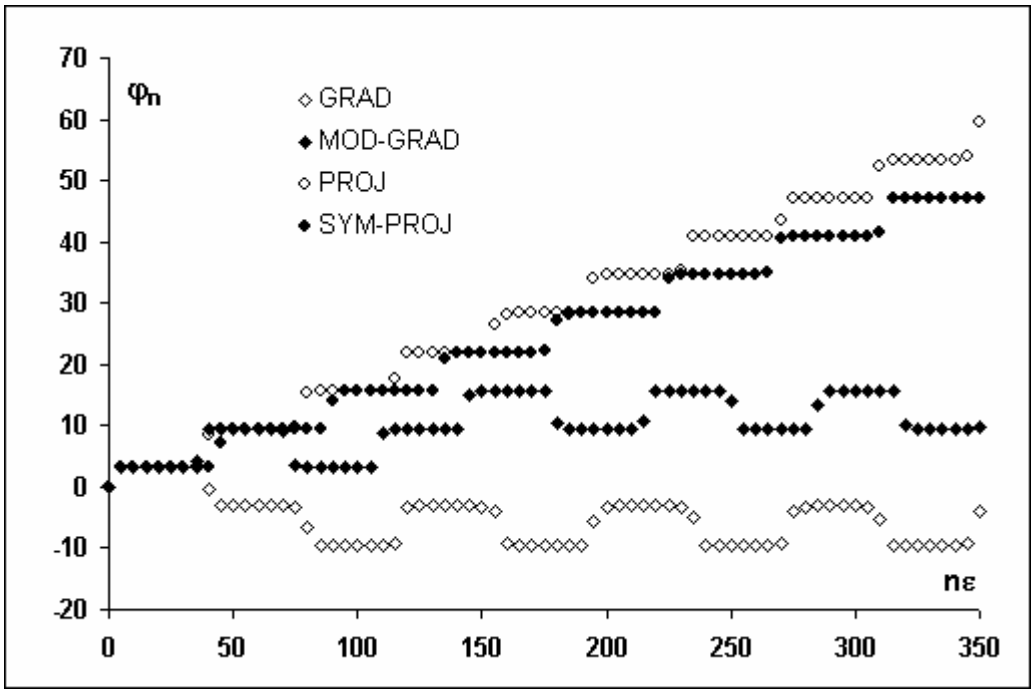
Pozostałe dyskretyzacje mogą generować złe wyniki nie tylko ilościowe, ale nawet jakościowe. Przykładowo, LF i oba schematy Surisa zaczynają symulować ruch rotacyjny przy pewnych $p_0 < 2$ dla dostatecznie dużych ε (np. dla p_0 = 1.99 od ε = 0.5, a dla p_0 = 1.99999 od ε = 0.02). W tym samym czasie metoda *implicit midpoint* generuje oscylacje dla pewnych $p_0 > 2$ (np. dla p_0 = 2.000001 od ε = 0.02, a dla p_0 = 2.001 od ε = 0.5). Dla pewnych warunków początkowych metoda LF generuje trajektorie chaotyczne. Nawet w przypadku dobrego zachowania jakościowego metody te dają bardzo duże błędy względne, zwłaszcza dla większych ε (dla ε = 0.5 i $|\, p_0 - 2\,| \le 0.001$ schematy LF, *implicit midpoint* i oba Surisa osiągają błędy względne rzędu 30% - 70% i więcej.

Jeśli $p_0 = 2$, wówczas (w przypadku ciągłym) mamy ruch wzdłuż separatrysy, to znaczy $\varphi \to \pi$ dla $t \to \infty$. Przy większych ε (poczynając od

$\varepsilon = 0.2$) zachowanie to nie jest odtwarzane przez żadną z badanych dyskretyzacji.

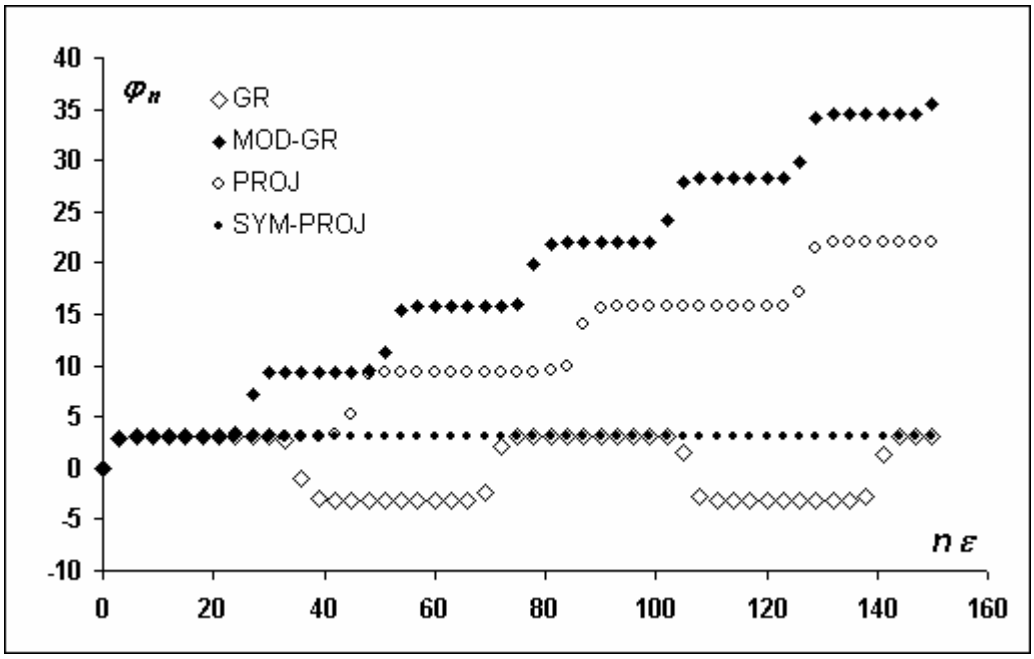


Wykres 4.16. φ_n dla $p_0 = 2.000001$, $\varepsilon = 0.1$. Okres rozwiązania dokładnego $T_{th} = 16.58809538$, średni okres zmodyfikowanego dyskretnego gradientu: $T = 16.56380722$.



Wykres 4.17. φ_n dla $p_0 = 2$, $\varepsilon = 0.1$, błąd zaokrąglenia $\Delta = 10^{-16}$.

Ciekawe wyniki dają oba schematy gradientowe (wykres 4.17). Standardowy dyskretny gradient generuje oscylacje, ale wcześniej wykonał dwie wsteczne rotacje. Podobnie zachowuje się zmodyfikowany dyskretny gradient, a obraz ich ruchu zależy od ε oraz wyboru błędu zaokrąglenia. W każdym przypadku dla obu schematów gradientowych mamy pewną liczbę wyglądających chaotycznie przeskoków pomiędzy oscylacjami i rotacjami w obu kierunkach. Jakościowo takie zachowanie należy uznać za prawidłowe. Odzwierciedla ono fakt, że stan równowagi w punkcie $\varphi = \pi$ jest nietrwały. W tym samym czasie dyskretyzacje rzutowane (zupełnie dobrze jakościowo opisujące ruch w pobliżu separatrysy) generują stosunkowo powolny ruch rotacyjny (podobnie jak standardowa metoda LF i obydwa schematy Surisa). Jednakże dla bardzo małych ε (tj. $\varepsilon \leq$ 0.00025) metoda rzutowania symetrycznego wydaje się generować właściwe zachowanie jakościowe i jest o wiele lepsza niż inne rozpatrywane schematy numeryczne (wykres 4.18).



Wykres 4.18. φ_n dla $p_0 = 2$, $\varepsilon = 0.00025$, błąd zaokrąglenia $\Delta = 10^{-18}$.

Korzyści wynikające z nowej metody dyskretyzacji

Metoda dyskretnego gradientu skorygowana współczynnikiem δ okazała się być bardzo wydajna przy numerycznej estymacji okresu (dla stosunkowo małych amplitud). Zakres tych „małych" amplitud jest całkiem duży, sięgający do $\varphi \approx \pi/4$, co odpowiada $p_0 < 0.8$. Obejmuje zatem również przypadki, które nie mogą być przybliżane drganiami liniowymi. Nawet dla $p_0 \approx 0.8$ nowa metoda jest kilka razy lepsza niż najlepsze z rozpatrywanych schematów, a dla mniejszych p_0 staje się lepsza nawet o cztery rzędy wielkości (np. dla $p_0 = 0.02$ względne odchylenie innych dyskretyzacji jest większe o czynnik przynajmniej 0.5×10^4 (tabela 4.3)).

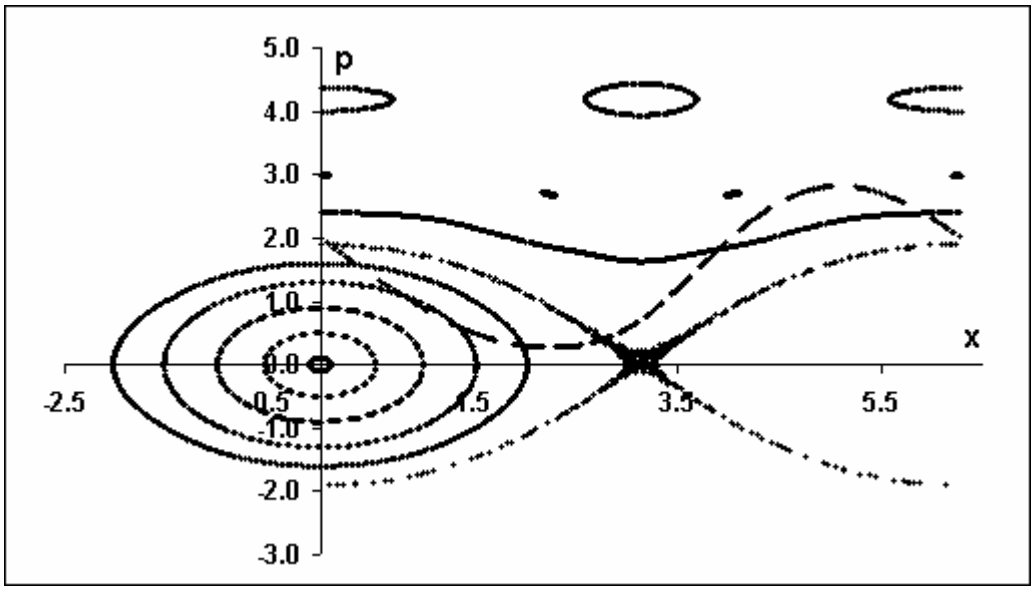
Wykres 4.12 ($p_0 = 0.1$) pokazuje jak dokładny jest okres drgań tej metody w porównaniu z okresami innych schematów numerycznych. Podobnie, wykres 4.14 prezentuje względne odchylenie od teorii dla $p_0 = 0.02$ i szerokiego zakresu zmienności ε. Widzimy, że nawet dla $\varepsilon = 1$, względne odchylenie wynosi tylko 10^{-5}! Przy małych wartościach ε odchylenie to osiąga 10^{-9} i mniej.

Metoda ta pracuje również bardzo dobrze przy dużych amplitudach, jednak dla p_0 większych niż 1.4, lepszy jest zwykły dyskretny gradient, a dla $p_0 \approx 1.2$ i $p_0 \approx 1.6$ nie do pobicia są odpowiednio schematy LF i *implicit midpoint*, dla których są to amplitudy „rezonansowe". W przypadku $p_0 > 2$ poprawka δ ma negatywny wpływ na dokładność dyskretyzacji gradientowej (najlepszej dla ruchu obrotowego), jednak utrzymuje się ona na poziomie pozostałych badanych schematów numerycznych.

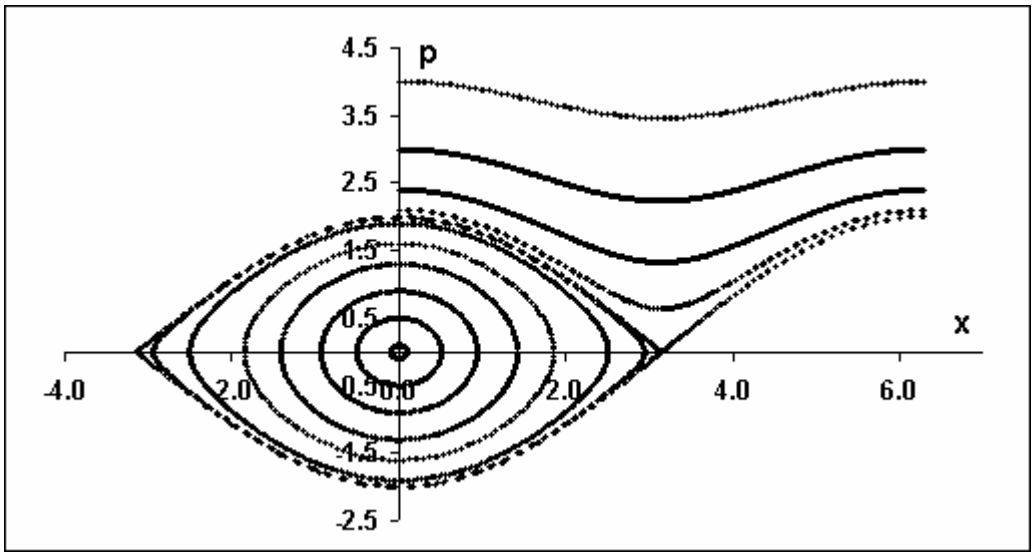
W bliskim otoczeniu separatrysy zmodyfikowana metoda dyskretnego gradientu zachowuje się podobnie jak zwykły dyskretny gradient i doskonale odwzorowuje jakościowe cechy ruchu. Co więcej, również jej wyniki ilościowe są bardzo dobre (tabela 4.3). Wykres 4.16 porównuje zachowanie tej metody ze schematami LF i *implicit midpoint* dla $p_0 = 2.000001$. Punkty generowane przez zmodyfikowaną metodę dyskretnego gradientu praktycznie pokrywają się z rozwiązaniem dokładnym (względny błąd okresu wynosi 0.59%), prawie tak dobrze, jak w przypadku dyskretnego gradientu (względny błąd 0.25%). Widzimy na nim, że metoda LF daje dobre zachowanie jakościowe, ale z okresem dwukrotnie mniejszym od teoretycznego. *Implicit midpoint* daje niewłaściwe zachowanie jakościowe: oscylacje zamiast rotacji.

4.12 Krzywe fazowe

Przedstawimy tu wybrane wykresy fazowe uzyskane z pomocą dyskretyzacji LF i metody dyskretnego gradientu. Oba schematy dają podobny obraz przestrzeni fazowej z tą różnicą, że dyskretny gradient działa zasadniczo przy dowolnym pędzie początkowym p_0 i kroku czasowym ε, zaś LF traci stabilność w pobliżu separatrysy i przy większych ε. Ilustrują to wykresy 4.19 i 4.20.

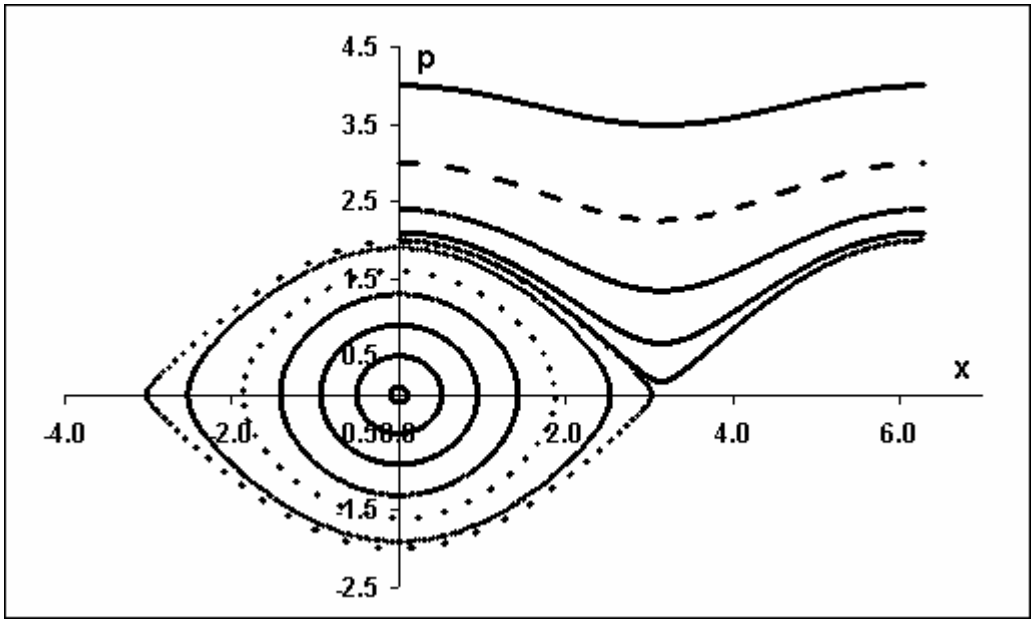


Wykres 4.19. Rodzina krzywych fazowych dyskretyzacji LF w obszarze niestabilności, $\varepsilon = 0.75$, $t = 0$, $p_0 \in \{0.2; 0.5; 0.9; 1.3; 1.6; 1.9; 2.000001; 2.1; 2.4; 3.0; 4.0\}$ (modulo 2π dla ruchu rotacyjnego).

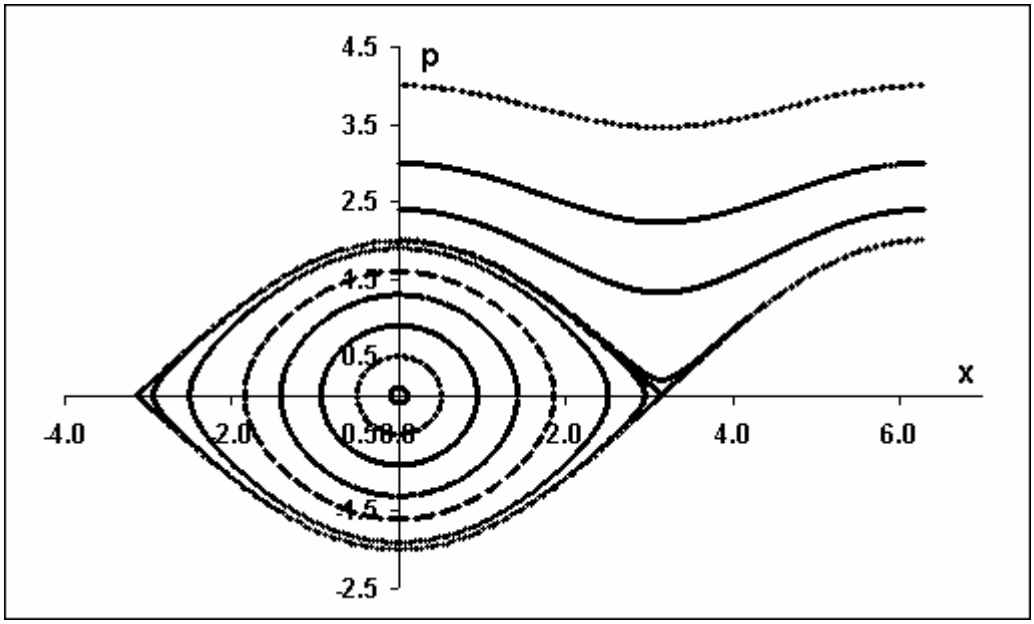


Wykres 4.20. Rodzina krzywych fazowych standardowego dyskretnego gradientu, $\varepsilon = 0.75$, $t = 0$, $p_0 \in \{0.2; 0.5; 0.9; 1.3; 1.6; 1.9; 2.000001; 2.1; 2.4; 3.0; 4.0\}$ (modulo 2π dla ruchu rotacyjnego).

Chociaż LF jest metodą symplektyczną, a dyskretny gradient nie, wykresy 4.21 i 4.22 pokazują, że w obu przypadkach obraz przestrzeni fazowej jest równie stabilny w czasie.



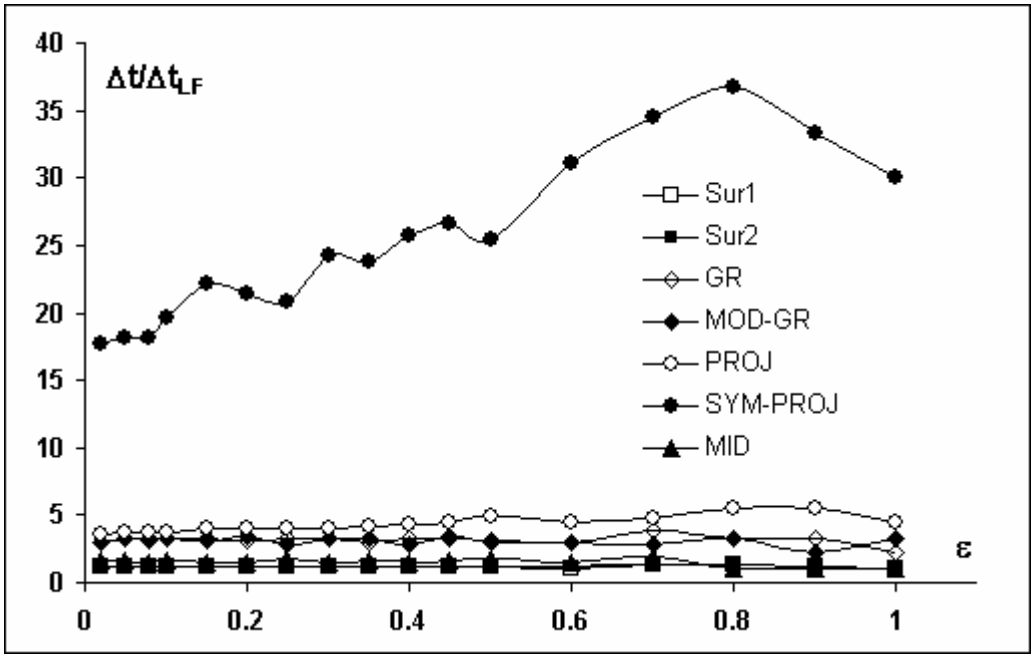
Wykres 4.21. Rodzina krzywych fazowych dyskretyzacji LF, $\varepsilon = 0.2$, $t = 10^6$, $p_0 \in \{0.2; 0.5; 0.9; 1.3; 1.6; 1.9; 1.99; 1.999999; 2.1; 2.4; 3.0; 4.0\}$ (modulo 2π dla ruchu rotacyjnego).



Wykres 4.22. Rodzina krzywych fazowych standardowego dyskretnego gradientu, $\varepsilon = 0.2$, $t = 10^6$, $p_0 \in \{0.2; 0.5; 0.9; 1.3; 1.6; 1.9; 1.99; 1.999999; 2.000001; 2.1; 2.4; 3.0; 4.0\}$ (modulo 2π dla ruchu rotacyjnego).

4.13 Wydajność badanych dyskretyzacji

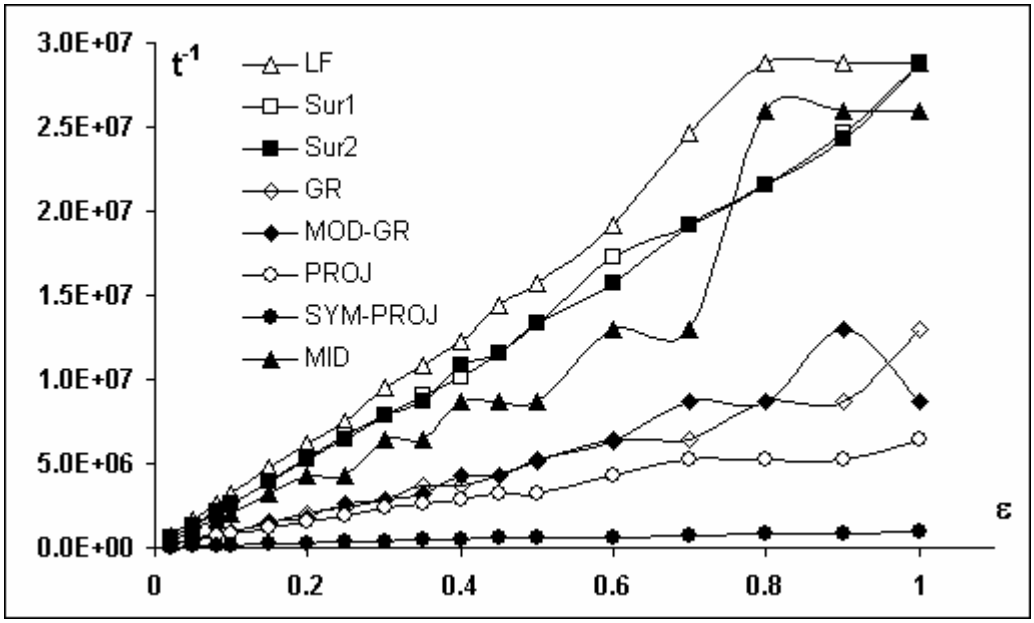
Jest oczywiste, że dyskretyzacje gradientowe, rzutowane i *implicit midpoint* wymagają większego nakładu pracy niż reszta badanych schematów. Dodatkowo często można je realizować przy pomocy różniących się szybkością algorytmów. Przykładowo, jeśli priorytetem jest szybkość działania – lepiej do rozwiązywania równań wykorzystać metodę Newtona. Doświadczenie pokazuje jednak, że wprowadza ona pewien dryf wyników w dłuższym okresie czasu. Dlatego, gdy zależy nam na dokładności, lepiej jest wykorzystać np. wolniej działającą metodę bisekcji. Przeprowadzono pomiary czasów wykonania stosownych procedur dla wszystkich omawianych schematów numerycznych parametryzując je czasem ewolucji układu i wykorzystując w implementacjach metodę Newtona. Wykres 4.23 pozwala ocenić względną pracochłonność wszystkich metod jako funkcję ε dla przykładowej wartości p_0.



Wykres 4.23. Względne czasy wykonania badanych dyskretyzacji dla $p_0 = 0.1$ (LF = 1).

Obie metody Surisa (o równej w zasadzie pracochłonności) są 20-30% wolniejsze od schematu LF, *implicit midpoint* działa od 1,5 do 2 razy wolniej, metody gradientowe i rzutowana standardowo mają czasy wykonania od 3 do 6 razy dłuższe, na końcu jest metoda naturalna rzutowana symetrycznie, która daje czasy od 15 do nawet 35 razy dłuższe (przy małych prędkościach początkowych i dużych ε). Charakterystyczna jest zależność szybkości działania wszystkich bardziej złożonych metod od ε.

Gdyby wykonanie każdego kroku trwało zawsze tyle samo, (co z dość dużą dokładnością zachodzi w przypadku metod dających się wyrazić jawnie za pomocą formuły matematycznej), czasy działania dyskretyzacji $t(\varepsilon)$ powinny zachowywać się jak funkcje postaci $t(\varepsilon) = a\varepsilon^{-b}$, gdzie $b = 1$. Wówczas odwrotność tej funkcji postaci $t^{-1}(\varepsilon) = \frac{1}{a}\varepsilon^{b}$ może definiować wydajność dyskretyzacji wyrażoną w arbitralnych jednostkach, a jej przebieg powinien być zbliżony do linii prostej. Rzeczywiste przebiegi funkcji $t^{-1}(\varepsilon)$ przedstawiono na wykresie 4.24.



Wykres 4.24. Porównanie wydajności badanych dyskretyzacji dla p_0 = 0.1 (LF = 1).

Współczynniki funkcji potęgowych postaci $t(\varepsilon) = \alpha\varepsilon^{\beta}$ dopasowanych do wybranych krzywych z wykresu 4.24 zamieszczone w tabeli 4.4 pokazują, że w odchylenia współczynnika β od jedności dla większości dyskretyzacji są niewielkie. Dotyczy to zwłaszcza metody LF, ale także Sur1 i Sur2. Współczynnik β najbardziej odbiega od jedności w metodzie rzutowania symetrycznego, która charakteryzuje się największą złożonością obliczeniową. Czas wykonania jednego kroku w tej metodzie zdecydowanie nie jest stały i zależy, przy zadanej dokładności, od takich globalnych parametrów jak p_0 i ε, ale także chwilowych wartości położenia i pędu.

Tabela 4.4. Współczynniki dopasowania funkcji potęgowych postaci $t(\varepsilon) = \alpha\varepsilon^{\beta}$ do krzywych wydajności dla wybranych dyskretyzacji (p_0 = 0,1; t = 3000)

Dyskretyzacja	α	$\Delta\alpha$	β	$\Delta\beta$
LF	$3{,}18\cdot10^{7}$	$9{,}03\cdot10^{4}$	0,998	0,00506
Sur2	$2{,}67\cdot10^{7}$	$9{,}37\cdot10^{4}$	1,006	0,00655
MOD-GR	$1{,}06\cdot10^{7}$	$9{,}23\cdot10^{4}$	1,016	0,01570
PROJ	$1{,}00\cdot10^{6}$	$8{,}36\cdot10^{4}$	0,805	0,12760

Omówione pomiary wydajności badanych schematów numerycznych wystarczają, aby postawić i odpowiedzieć na pytanie, czy koszty uzyskania lepszego odwzorowania amplitudy lub okresu drgań przy danej wartości kroku czasowego ε przy pomocy zaawansowanych metod nie są zbyt wysokie. Być może bardziej ekonomiczne jest zastosowanie dyskretyzacji najprostszej i najszybszej (LF) z odpowiednio małym krokiem czasowym? Odpowiedź znajdziemy w tabeli 4.5, która zawiera porównanie efektywności metody LF (szybkiej i dobrej dla rozsądnych parametrów wyjściowych) z wprowadzoną w tym rozdziale metodą zmodyfikowanego dyskretnego gradientu. Za miarę dokładności przyjęto względne odchylenie od okresu teoretycznego. W trzech pierwszych kolumnach są dane dyskretyzacji LF: krok czasowy ε, odpowiadające mu względne odchylenie okresu $\Delta T/T$ oraz czas wykonania odpowiedniej procedury. W czterech dalszych kolumnach mamy względne odchylenie $\Delta T/T$ okresu metody konkurencyjnej przy kroku ε, krok czasowy $\varepsilon' > \varepsilon$, przy którym metoda konkurencyjna osiąga dokładność metody LF, czas pracy tej metody t' przy nowym kroku ε' i na końcu wartość ułamka t'/t (im mniejszy, tym lepiej dla metody konkurującej).

Nietrudno sprawdzić, że dyskretyzacja LF w szerokim zakresie prędkości początkowych (od 0,5 do 1,5) ewentualne braki dokładności nadrabia szybkością działania. Po drodze przytrafia się jej p_0 „rezonansowe" o wartości leżącej w pobliżu 1.21, przy której dodatkowo osiąga najlepszą dokładność odwzorowania okresu. Jednak w obszarze $p_0 < 0{,}5$ zmodyfikowany dyskretny gradient swoją dokładnością zaczyna kompensować mniejszą szybkość działania. Efekt ten potęguje się przy malejących p_0 (zmodyfikowany dyskretny gradient staje się o ponad rząd wielkości bardziej efektywny niż LF). W pobliżu prędkości 1,6 najbardziej efektywna jest metoda *implicit midpoint*, a dla wyższych prędkości ponownie LF oraz obie metody Surisa (nie dotyczy to obszaru w pobliżu separatrysy, w którym metody te dają złe wyniki).

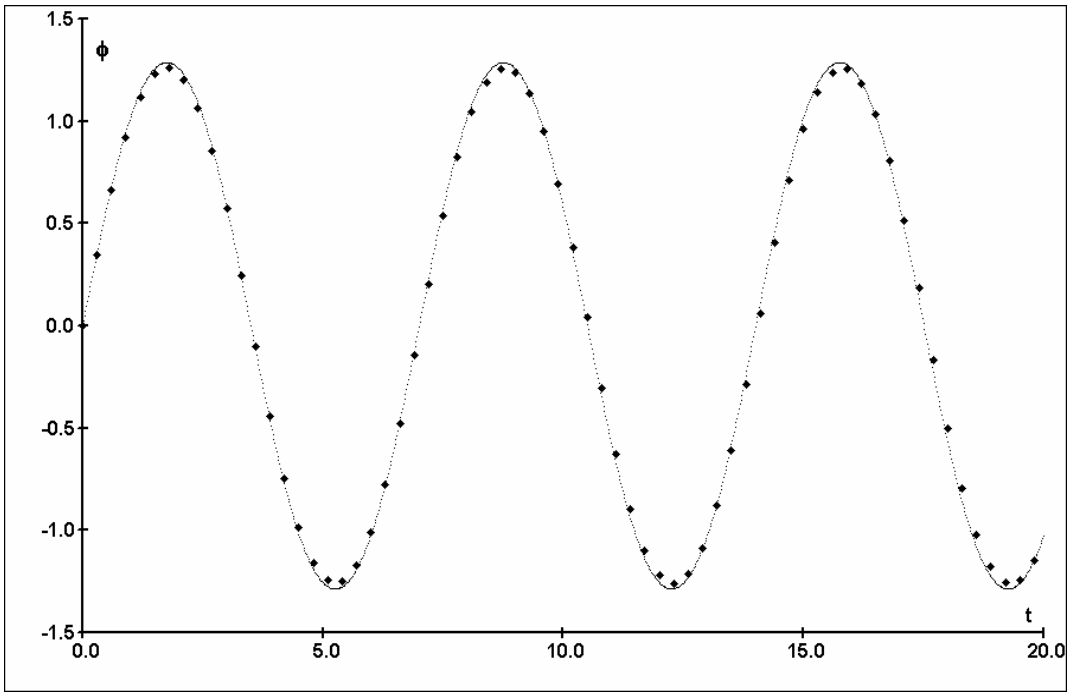
Tabela 4.5. Porównanie efektywności (czasów wykonania) wybranych schematów numerycznych z efektywnością dyskretyzacji naturalnej na przykładzie odchyleń średnich początkowych okresów od teorii dla małych prędkości początkowych

$p_0 = 0{,}02$						
LF			MOD-GR			
ε	$\Delta T / T$	t	$\Delta T / T$	ε′	t′	t′/t
0,02	1,67E-05	1,581E-06	1,25E-07	0,7185	1,257E-07	0,0795
$p_0 = 0{,}05$						
LF			MOD-GR			
ε	$\Delta T / T$	t	$\Delta T / T$	ε′	t′	t′/t
0,02	1,66E-05	1,573E-06	3,14E-08	0,7239	1,309E-07	0,0832
$p_0 = 0{,}1$						
LF			MOD-GR			
ε	$\Delta T / T$	t	$\Delta T / T$	ε′	t′	t′/t
0,02	1,66E-05	1,576E-06	1,09E-07	0,2689	4,888E-07	0,3102
0,05	1,04E-04	6,263E-07	5,18E-07	0,6939	1,924E-07	0,3072
0,08	2,65E-04	3,902E-07	1,38E-06	0,9847	1,364E-07	0,3495
$p_0 = 0{,}3$						
LF			MOD-GR			
ε	$\Delta T / T$	t	$\Delta T / T$	ε′	t′	t′/t
0,02	1,59E-05	1,562E-06	7,45E-07	0,0913	1,099E-06	0,7036
0,05	9,94E-05	6,215E-07	4,71E-06	0,2281	4,375E-07	0,7039
0,08	2,55E-04	3,873E-07	1,21E-05	0,3708	2,683E-07	0,6928
0,1	3,98E-04	3,095E-07	1,89E-05	0,4624	2,149E-07	0,6944
0,15	8,96E-04	2,058E-07	4,19E-05	0,7214	1,374E-07	0,6673
0,2	1,60E-03	1,541E-07	7,56E-05	0,9709	1,019E-07	0,6611
$p_0 = 0{,}5$						
LF			MOD-GR			
ε	$\Delta T / T$	t	$\Delta T / T$	ε′	t′	t′/t
0,02	1,45E-05	1,595E-06	2,13E-06	0,0521	1,993E-06	1,2490
0,05	9,07E-05	6,279E-07	1,31E-05	0,1294	7,797E-07	1,2417
0,08	2,32E-04	3,892E-07	3,35E-05	0,2101	4,730E-07	1,2152
0,1	3,63E-04	3,102E-07	5,23E-05	0,2635	3,745E-07	1,2075
0,15	8,17E-04	2,053E-07	0,000118	0,3972	2,453E-07	1,1949
0,2	1,45E-03	1,532E-07	0,000209	0,5333	1,811E-07	1,1818

Jeśli chodzi o dyskretyzacje rzutowane, to nie są one w stanie zbliżyć się do efektywności innych metod, mimo że w pobliżu $\dot{\alpha}_0 \approx 0{,}5$ dyskretyzacja LF rzutowana symetrycznie ma rezonansową wartość p_0 i jest najdokładniejsza.

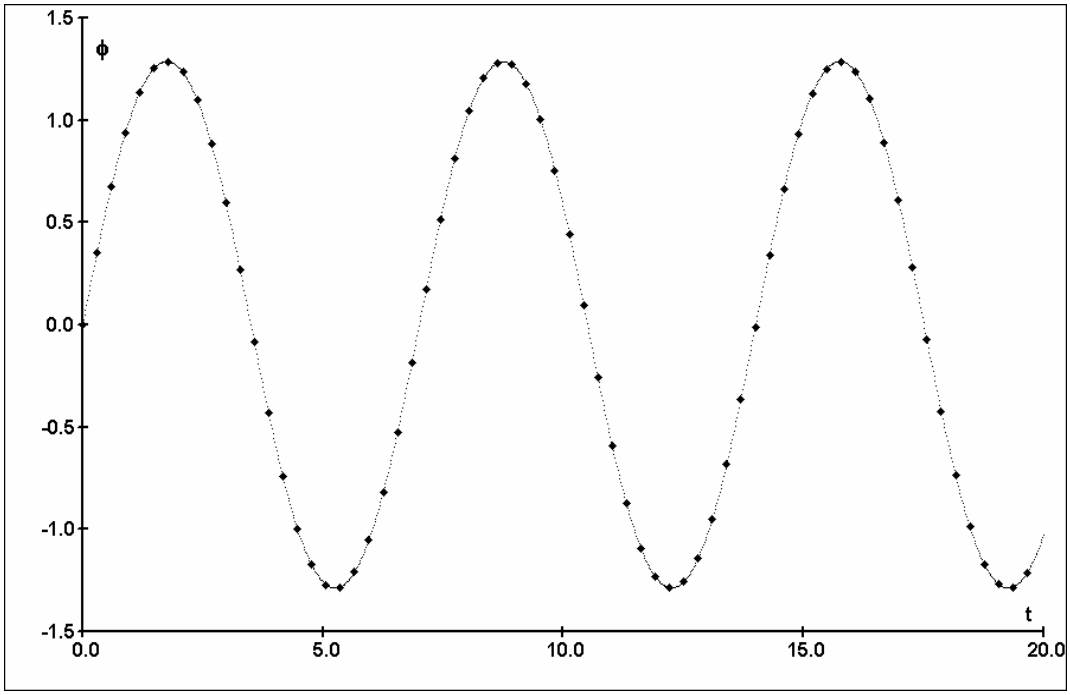
4.14 Numeryczne modyfikowanie badanych dyskretyzacji

Okazuje się, że z pomocą każdej z badanych dyskretyzacji działającą przy (rozsądnych) warunkach początkowych (p_0, ε) można bardzo dokładnie odwzorować przebieg teoretyczny poprzez zastosowanie do niej nieco zmienionej wartości $p_0' = w_{p0} p_0$ i przeskalowania osi czasu z pomocą nowego kroku czasowego $\varepsilon' = w_\varepsilon \varepsilon$ przy pomocy pary współczynników (w_{p0}, w_ε). Za przykład niech posłuży dyskretyzacja Sur1, której przebiegi przed i po modyfikacji zaprezentowano na wykresach 4.25 i 4.26.



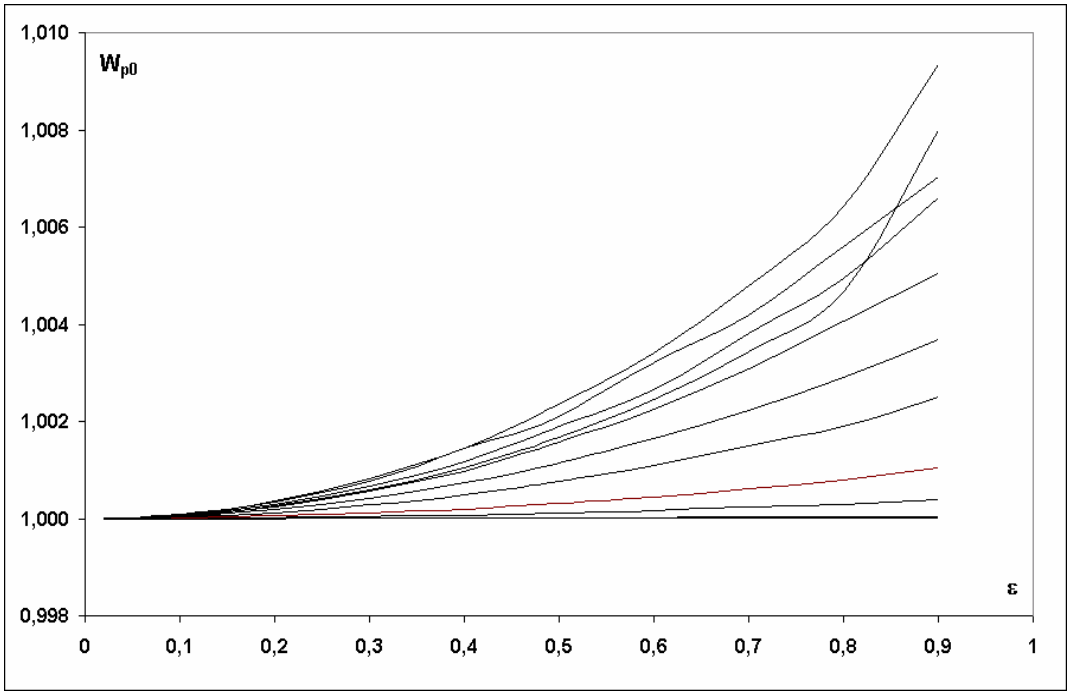
Wykres 4.25. Przebieg dyskretyzacji Sur1, p_0 = 1.2; ε = 0.3

Współczynnik w_{p0} służy do zmodyfikowania amplitudy drgań, natomiast w_ε do zmodyfikowania ich okresu. Do utworzenia nowej, zmodyfikowanej dyskretyzacji, o znacznie większej od pierwowzoru dokładności działania potrzebna jest znajomość dwóch funkcji $w_{p0}(p_0, \varepsilon)$ i $w_\varepsilon(p_0, \varepsilon)$. Można je wyznaczyć eksperymentalnie (stablicować) dla rozsądnych parametrów ruchu i wykorzystywać później posiłkując się interpolacją. Wykresy 4.27, 4.28, 4.29 i 4.30 przedstawiają wyznaczone numerycznie rodziny funkcji $w_{p0}(\varepsilon)$ i $w_\varepsilon(\varepsilon)$ parametryzowane wartością p_0 dla dyskretyzacji Sur1 i zmodyfikowanej metody dyskretnego gradientu (minimalizowano sumę kwadratów odchyleń od teorii w przedziale [10, 30]).

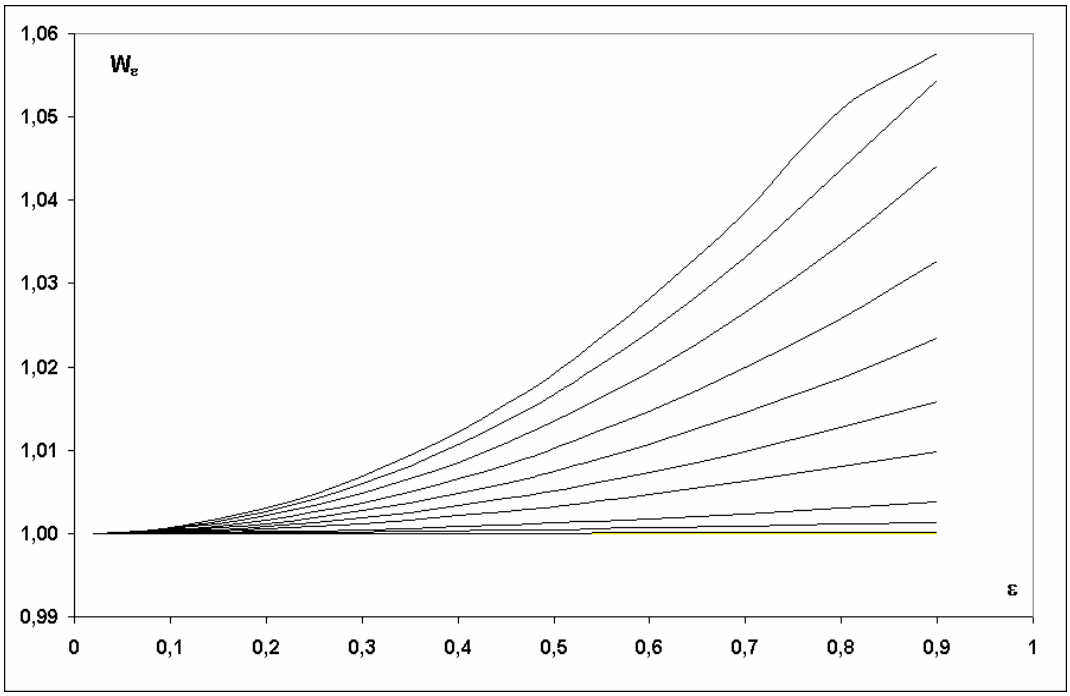


Wykres 4.26. Przebieg zmodyfikowanej dyskretyzacji Sur1 (w_α = 1.01798; w_ε = 0.9926), p_0 = 1.2; ε = 0.3.

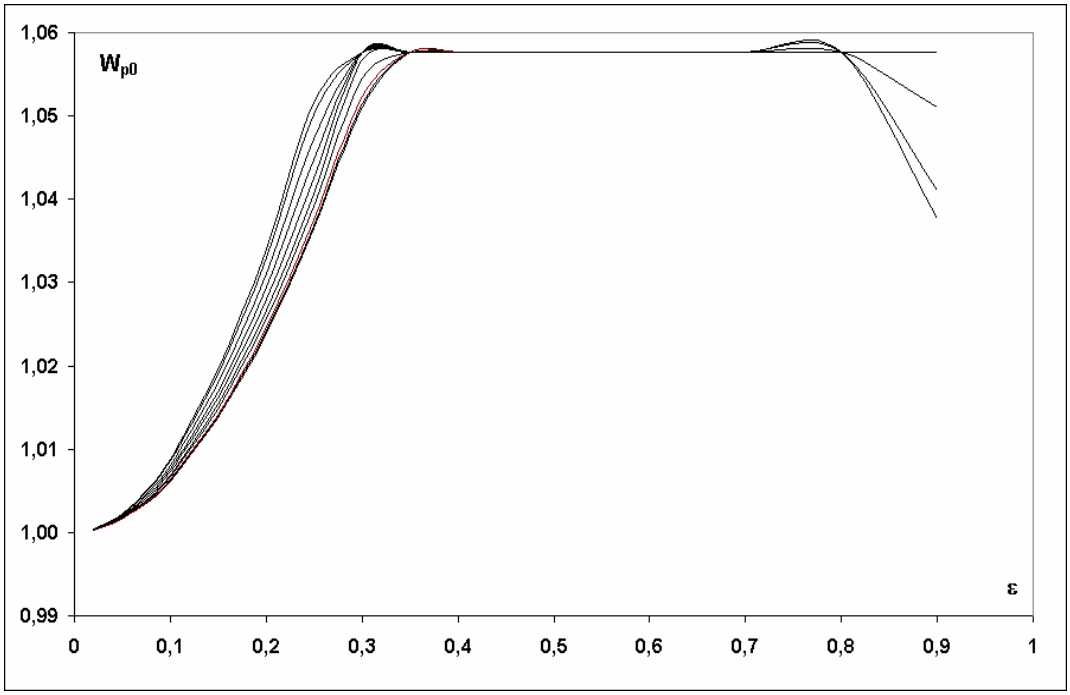
Oczekujemy, że $w_{p0}(\varepsilon) \to 1$ i $w_\varepsilon(\varepsilon) \to 1$, gdy $\varepsilon \to 0$ niezależnie od p_0, co obserwujemy na wykresach. Ponieważ ograniczono się do „rozsądnych" wartości parametrów p_0 i ε, przebiegi $w_{p0}(p_0, \varepsilon)$ i $w_\varepsilon(p_0, \varepsilon)$ nie różnią się od jedności o więcej niż kilka procent. Widać też, że poprawianie numeryczne dyskretyzacji Sur1, charakteryzującej się wyjątkowo dużą stabilnością pracy, jest trudniejsze ze względu na bardziej złożoną postać przedstawionych na wykresach 4.29 i 4.30 funkcji.



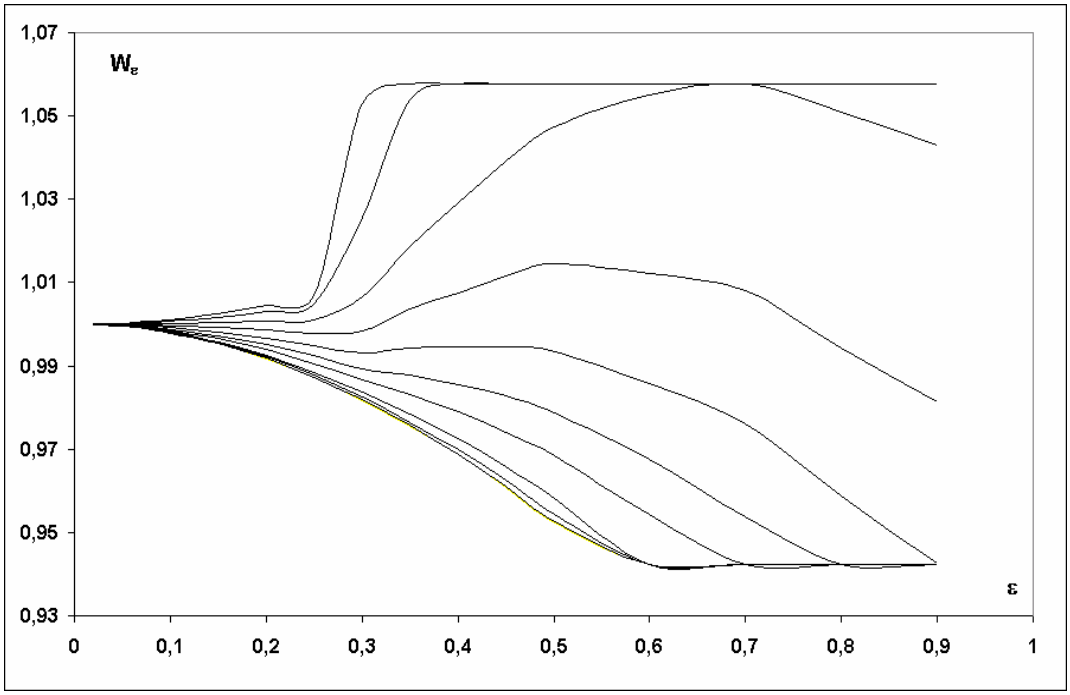
Wykres 4.27. Przebiegi funkcji $w_{p0}(\varepsilon)$ dla metody modyfikowanego dyskretnego gradientu dla $p_0 \in \{0.02; 0.05; 0.1; 0.3; 0.5; 0.8; 1.0; 1.2; 1.4; 1.6; 1.8; 1.9\}$ Większe wartości współczynnika odpowiadają większym p_0.



Wykres 4.28. Rodzina krzywych $w_{\varepsilon}(\varepsilon)$ dla metody modyfikowanego dyskretnego gradientu dla $p_0 \in \{0.02; 0.05; 0.1; 0.3; 0.5; 0.8; 1.0; 1.2; 1.4; 1.6; 1.8; 1.9\}$. Większe wartości współczynnika odpowiadają większym p_0.



Wykres 4.29. Przebiegi funkcji $w_{p0}(\varepsilon)$ dla metody Sur1 dla $p_0 \in \{0.02; 0.05; 0.1; 0.3; 0.5; 0.8; 1.0; 1.2; 1.4; 1.6; 1.8; 1.9\}$. Większe wartości współczynnika odpowiadają większym p_0.



Wykres 4.30. Przebiegi funkcji $w_{\varepsilon}(\varepsilon)$ dla metody Sur1 dla $p_0 \in \{0.02; 0.05; 0.1; 0.3; 0.5; 0.8; 1.0; 1.2; 1.4; 1.6; 1.8; 1.9\}$. Większe wartości współczynnika odpowiadają większym p_0.

4.15 Podsumowanie

Wszystkie metody rozpatrywane w tym rozdziale charakteryzują się bardzo wysoką stabilnością generowanego ruchu okresowego (zakładając, że ε nie jest zbyt wielkie). Są one znacznie bardziej stabilne niż np. niesymplektyczna metoda Rungego-Kutty nawet wysokiego rzędu. Średni okres jest praktycznie stały w bardzo długim przedziale czasu z dokładnością przynajmniej 10^{-7} (testowano dla kilku milionów okresów). Okres i amplituda, jako funkcje czasu, wykazują regularne, małe oscylacje (większe dla metod projekcyjnych). Efekt ten został wyjaśniony poprzez analizę wymiernych przybliżeń (z możliwie małymi mianownikami) liczb rzeczywistych T/ε i $T/(2\varepsilon)$ (podrozdział 9).

Głównym celem tego rozdziału było porównanie kilku schematów numerycznych. Poniżej wyliczono ich najbardziej istotne cechy.

- Stabilność
  - wszystkie rozpatrywane metody (najsłabsze są tu metody rzutowane) dają bardzo stabilne wartości okresu i amplitudy,
  - metody rzutowane dają okresy i amplitudy stabilne po uśrednieniu, jednak z dużymi oscylacjami wokół średniej (przy większych wartościach ε, wykres 4.5).
- Dokładność okresu

  Poza najbliższym otoczeniem separatrysy (tabela 4.3):
  - wszystkie metody (z niewielkimi wyjątkami) mają względne odchylenia tego samego rzędu (zależne od ε),
  - dla małych p_0 zmodyfikowana metoda dyskretnego gradientu jest lepsza o cztery rzędy wielkości od innych metod,
  - w ruchu rotacyjnym metoda dyskretnego gradientu jest generalnie najlepsza (zwłaszcza dla dużych ε i p_0) przewyższając inne metody nawet do dwóch rzędów wielkości.

  W pobliżu separatrysy (tabela 4.3):
  - obydwie metody gradientowe dają najlepsze rezultaty,
  - obydwie metody rzutowane dają dobre wyniki dla $p_0 < 2$,
  - najlepszą symulację ruchu wzdłuż separatrysy daje metoda rzutowania symetrycznego (dla bardzo małych ε); patrz wykres 4.18,
  - LF, *implicit midpoint* i obydwie metody Surisa dają złe rezultaty,
- Dokładność amplitudy
  - Dla większych ε wszystkie metody mają dokładność tego samego rzędu (tabela 4.2). Metody gradientowe są nieco lepsze, podczas gdy LF i obie metody Surisa są gorsze od innych schematów.

– Dla mniejszych ε możemy podzielić metody na dwie grupy: mniej dokładne (LF i obie metody Surisa) i dokładniejsze (schematy gradientowe i rzutowane), lepsze o trzy rzędy wielkości. *Implicit midpoint* należy do pierwszej grupy przy większych wartościach p_0 i do drugiej przy małych.

Standardowa metoda LF, choć niecałkowalna, jest zupełnie dobra w porównaniu z innymi, bardziej wyszukanymi dyskretyzacjami. Jej dokładność może zostać poprawiona poprzez użycie metod projekcyjnych, które wymuszają zachowanie całki energii. Rzutowanie pracuje bardzo dobrze dla małych wartości kroku czasowego, przy dużych natomiast generuje znaczne fluktuacje okresu i amplitudy. W każdym przypadku rzutowanie pozwala uzyskać znacznie większą dokładność odwzorowania średniej amplitudy. Średni okres jest podobny do wartości generowanych przez standardowy LF: nieco lepszy w przypadku ruchu oscylacyjnego, ale trochę gorszy w ruchu obrotowym.

Stwierdzono zaskakujące rezonanse dla $p_0 \approx 1.21$ (dla metody LF) i $p_0 \approx 1.6$ (dla metody *implicit midpoint*). W pobliżu tych punktów obie metody osiągają wyjątkowo dużą dokładność odwzorowania okresu (praktycznie dla dowolnego ε), znacznie lepszą niż wszystkie inne metody. Interesujące byłoby wyjaśnienie tego efektu.

Dyskretyzacje znalezione przez Surisa [25] są bardzo stabilne, lecz generują stosunkowo duże odchylenia w porównaniu z innymi metodami. Jest to zaskakujące, gdyż każda z nich jest całkowalna i symplektyczna. Prawdopodobnie duże odchylenia od rozwiązania dokładnego mają charakter systematyczny. Możliwe, że uda się je tak zmodyfikować, aby poprawić ich dokładność, nie psując stabilności.

Metoda dyskretnego gradientu należy (dla dowolnych ε i p_0) do najdokładniejszych metod. W rozdziale tym zaproponowano jej modyfikację, która okazała się skuteczna, zwłaszcza dla ruchu oscylacyjnego. Metoda ta jest wyjątkowo efektywna w przypadku małych drgań. Względne błąd okresu otrzymany z jej pomocą jest o przynajmniej cztery rzędy wielkości mniejszy niż osiągany z pomocą innych rozpatrywanych tu schematów numerycznych.

5 Lokalnie dokładne modyfikacje metody dyskretnego gradientu w przypadku jednowymiarowym

5.1 Wprowadzenie

Metoda dyskretnego gradientu została wprowadzona na początku lat 70-tych ubiegłego wieku w celu numerycznego całkowania równań ruchu układów N ciał w mechanice klasycznej z możliwością zastosowania w dynamice molekularnej i mechanice nieba [37, 54]. W późniejszych latach badania w tym kierunku były kontynuowane [35, 49, 65, 97], ale znacznie większą popularność zdobyły równolegle rozwijane metody symplektyczne [29, 30, 93, 110, 111, 113], zwłaszcza że były wśród nich schematy otwarte (jawne), a poza tym dość szybko znaleziono sposoby na znaczną poprawę dokładności schematów symplektycznych bez utraty ich własności geometrycznych [105, 112].

W rozdziale tym opisany zostanie nowy schemat numeryczny, który nazwany został lokalnie dokładną metodą dyskretnego gradientu (w skrócie: GR-LEX), a także symetryczna (odwracalna w czasie) modyfikacja tego schematu (GR-SLEX).

Główna idea prezentowanego tu podejścia polega na zmodyfikowaniu danego schematu numerycznego (zazwyczaj poprzez zamianę kroku czasowego ε na pewną funkcję zależącą od ε i innych zmiennych niezależnych) w celu uzyskania schematu „lokalnie dokładnego”, to jest takiego, który dla równań zlinearyzowanych staje się dyskretyzacją dokładną [19].

Sprawą wielkiej wagi jest zachowanie „dobrych” własności wyjściowego schematu numerycznego podczas takiej modyfikacji. W przypadku dyskretnego gradientu jego „lokalnie dokładna” modyfikacja prezentowana w tym rozdziale zachowuje całkę energii. W efekcie uzyskany został schemat cechujący się wysoką stabilnością i bardzo dobrym zachowaniem jakościowym przy jednoczesnym zdecydowanym wzroście dokładności.

W rozdziale tym ograniczamy się do przypadku jednowymiarowego:

$$\dot{p} = -V'(x), \quad \dot{x} = p\,, \tag{5.1}$$

gdzie $V(x)$ jest potencjałem, a kropka i prim oznaczają odpowiednio różniczkowanie po t oraz x. W takim przypadku metoda dyskretnego gradientu redukuje się do tak zwanej zmodyfikowanej metody punktu środkowego [54]:

$$\begin{aligned} \frac{x_{n+1} - x_n}{\varepsilon} &= \frac{1}{2}(p_{n+1} + p_n), \\ \frac{p_{n+1} - p_n}{\varepsilon} &= -\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n} \end{aligned} \tag{5.2}$$

przy czym ε oznacza krok czasowy. Można łatwo sprawdzić, że powyższy schemat zachowuje energię całkowitą:

$$\frac{1}{2} p_n^2 + V(x_n) = E = const\,. \tag{5.3}$$

Wystarczy tylko pomnożyć przez siebie oba równania (5.2), zamieniając uprzednio strony jednego z równań. Otrzymamy wtedy po jednej stronie różnicę energii kinetycznych, a po drugiej − różnicę energii potencjalnych.

Zmodyfikowana metoda punktu środkowego w naturalny sposób rozszerza się na przypadek trójwymiarowy i na układy cząstek. Zachowuje ona dokładnie energię całkowitą oraz pędu i momentu pędu układu [41, 54, 97].

W ostatnim czasie metoda dyskretnego gradientu została rozwinięta i rozszerzona w kontekście całkowania geometrycznego [35, 49, 63]. W szczególności Quispel ze współpracownikami skonstruowali schematy numeryczne zachowujące całki ruchu układów równań różniczkowych zwyczajnych [64, 84, 86].

Ogólnie można stwierdzić, że schematy numeryczne zachowujące całki ruchu bardzo dobrze odtwarzają jakościowe cechy równań różniczkowych, do symulacji których są stosowane, jednak nie jest łatwo poprawić ich dokładność. Przedstawione niżej wyniki wskazują nowe sposoby na istotne poprawienie dokładności metody dyskretnego gradientu (która nie jest symplektyczna) bez utraty jej bardzo dobrych cech jakościowych.

5.2 Zmodyfikowany schemat dyskretnego gradientu

W rozdziale 4 porównano szereg dyskretyzacji równania wahadła matematycznego ($V(x) = -\cos(x)$) ze szczególnym naciskiem na ich zachowanie w długim okresie czasu. Dyskretny gradient znalazł się wśród najlepszych metod zwłaszcza dla dużych energii (ruch rotacyjny) i w pobliżu separatrysy. Zaproponowaliśmy tam modyfikację schematu (5.2). Otóż, przy założeniu, że równowaga trwała występuje w punkcie $x = 0$, zamieniono ε na funkcję $\delta_0 = \delta_0(\varepsilon)$:

$$\delta_0 = \frac{2}{\omega_0} \tan \frac{\omega_0 \varepsilon}{2}, \tag{5.4}$$

gdzie $\omega_0 = \sqrt{V''(0)}$. Modyfikacja ta, nazwana MOD-GR, stosowana była do wyznaczania kolejnych punktów dyskretyzacji (jej krok czasowy pozostał równy ε). Uzasadnieniem dla tej zmiany była chęć zachowywania przez dyskretyzację (prawie dokładnie) małych oscylacji wokół $x = 0$, gdzie wahadło może być traktowane jak oscylator harmoniczny. A w tym przypadku istnieje tak zwana

dokładna dyskretyzacja, (rozdział 3). Zmodyfikowana metoda dyskretnego gradientu pozwoliła uzyskać dokładność o 4 rzędy wielkości lepszą od innych rozważanych schematów w przypadku małych oscylacji i zachowała porównywalną ze standardowym dyskretnym gradientem dokładność w przypadku innych warunków początkowych (obie metody są rzędu drugiego).

Schemat MOD-GR pojawił się u nas w wyniku czysto fizycznej motywacji (wykorzystanie oscylatora harmonicznego) [21], niezależnie od istniejących podejść numerycznych. Analogiczny lub nawet identyczny wynik można było osiągnąć stosując przynajmniej trzy inne podejścia: niestandardowe schematy różnicowe Mickensa [66, 67], metody trygonometryczne Gautschiego [33, 41] oraz integratory wykładnicze (*exponential integrators*) [70].

Modyfikacja (5.2) przypomina podejście Mickensa (w którym także ma miejsce zastępowanie ε pewną funkcją $\delta(\varepsilon)$), jednak wybór funkcji $\delta(\varepsilon)$ dokonywany jest u nas w inny sposób (bardziej precyzyjny i mniej intuicyjny). Co więcej, Mickens rozważa wyłącznie stałe funkcje $\delta(\varepsilon)$. W niniejszej pracy, poczynając od tego rozdziału, wprowadzamy znacznie ogólniejszą klasę funkcji, dopuszczając w zasadzie dowolną zależność funkcji δ od zmiennych i parametrów.

5.3 Metoda lokalnie dokładnego dyskretnego gradientu i jej symetryczna modyfikacja

Głównym wynikiem tego rozdziału jest jeszcze inna (dużo lepsza) modyfikacja standardowego dyskretnego gradientu [22, 24], która została nazwana lokalnie dokładnym dyskretnym gradientem (GR-LEX):

$$\begin{aligned} &\frac{x_{n+1}-x_n}{\delta_n} = \frac{1}{2}(p_{n+1}+p_n), \\ &\frac{p_{n+1}-p_n}{\delta_n} = -\frac{V(x_{n+1})-V(x_n)}{x_{n+1}-x_n} \end{aligned} \tag{5.5}$$

gdzie δ_n jest funkcją zdefiniowaną wzorami

$$\begin{aligned} &\delta_n = \frac{2}{\omega_n}\tan\frac{\varepsilon\omega_n}{2}, \text{ (jeżeli } V''(x_n)>0), \\ &\delta_n = \varepsilon, \text{ (jeżeli } V''(x_n)=0), \\ &\delta_n = \frac{2}{\omega_n}\tanh\frac{\varepsilon\omega_n}{2}, \text{ (jeżeli } V''(x_n)<0), \end{aligned} \tag{5.6}$$

przy czym ε oznacza krok czasowy, oraz

$$\omega_n = \sqrt{V''(x_n)}\,. \tag{5.7}$$

Wzory (5.6) nakładają ograniczenie na krok czasowy, mianowicie $\varepsilon\omega_n < \pi$. Warunek ten nie jest bardzo restrykcyjny, przykładowo dla wahadła matematycznego mamy $\omega_n \le \omega_0$ i tym samym $\varepsilon < \frac{1}{2}T_0$ (T_0 jest okresem małych drgań względem położenia równowagi).

W celu wyprowadzenia lokalnie dokładnego schematu numerycznego (5.5) użyjemy dokładnej dyskretyzacji klasycznego oscylatora harmonicznego ze stałą siłą wymuszającą

$$\ddot{\xi} + \omega^2 \xi = a, \quad \dot{\xi} = p\,, \tag{5.8}$$

gdzie $\omega > 0$ i a jest stałą. Równanie to posiada (podobnie jak wszystkie układy opisane równaniami różniczkowymi zwyczajnymi [1, 82]) następującą dyskretyzację dokładną [17]:

$$\begin{aligned} &\xi_{n+1} - 2\cos\omega\varepsilon\,\xi_n + \xi_{n-1} = \left(\frac{2}{\omega}\sin\frac{\omega\varepsilon}{2}\right)^2 a, \\ &p_n = \frac{\omega}{\sin\omega\varepsilon}(\xi_{n+1} - \xi_n\cos\omega\varepsilon) - \frac{a}{\omega}\tan\frac{\omega\varepsilon}{2}, \end{aligned} \tag{5.9}$$

którą można zapisać w postaci schematu jednokrokowego

$$\begin{aligned} &\xi_{n+1} = \xi_n\cos\omega\varepsilon + \frac{\sin\omega\varepsilon}{\omega}\,p_n + \frac{2a}{\omega^2}\left(\sin\frac{\omega\varepsilon}{2}\right)^2, \\ &p_{n+1} = p_n\cos\omega\varepsilon - \xi_n\omega\sin\omega\varepsilon + \frac{a}{\omega}\sin\omega\varepsilon. \end{aligned} \tag{5.10}$$

Dyskretyzacja (5.10) nosi nazwę dokładnej, gdyż $\xi_n = \xi(n\varepsilon)$, $p_n = p(n\varepsilon)$, gdzie $\xi(t)$, $p(t)$ spełniają układ równań (5.8) (w szczególności $\xi_0 = \xi(0)$, $p_0 = p(0)$). Przypadek $\omega^2 \le 0$ może być potraktowany analogicznie (na przykład dla $\omega^2 < 0$ możemy formalnie położyć $\omega = i\sqrt{|\omega^2|}$ i zamiast funkcji trygonometrycznych pojawią się wówczas funkcje hiperboliczne).

Aby uzyskać schemat (5.5) zastąpimy ε we wzorze (5.2) przez zmienną δ_n zależną nie tylko od ε, ale również od x_n. Postać funkcji δ_n wyznaczymy żądając, aby zmodyfikowany schemat (5.5) był *lokalnie dokładny*. Mówiąc o lokalnej dokładności będziemy mieli na myśli, że linearyzacja schematu (5.5) wokół punktu x_n, czyli

$$\begin{aligned}
&\frac{x_{n+1}-x_n}{\delta_n}=\frac{1}{2}(p_{n+1}+p_n),\\
&\frac{p_{n+1}-p_n}{\delta_n}=-V'(x_n)-\frac{1}{2}(x_{n+1}-x_n)V''(x_n),
\end{aligned}\tag{5.11}$$

jest zgodna z *dokładną dyskretyzacją* zlinearyzowanego wokół x_n układu równań (5.1), który przyjmie postać

$$\frac{dp}{dt}=-V'(x_n)-V''(x_n)\xi,\quad \frac{d\xi}{dt}=p,\tag{5.12}$$

przy czym $\xi = x - x_n$ i x_n jest ustalone (jest traktowane jako stała). Porównując (5.12) z (5.8) dostajemy

$$\omega^2=V''(x_n),\ \ a=-V'(x_n).\tag{5.13}$$

Co więcej, mamy też $\xi_n = x_n - x_n = 0$ i $\xi_{n+1} = x_{n+1} - x_n$. Dla ustalenia uwagi ograniczymy się do przypadku $\omega^2 > 0$. Wówczas równania (5.10) przyjmują postać

$$\begin{aligned}
&x_{n+1}=x_n+\frac{\sin\omega\varepsilon}{\omega}p_n+\frac{2a}{\omega^2}\left(\sin\frac{\omega\varepsilon}{2}\right)^2,\\
&p_{n+1}=p_n\cos\omega\varepsilon+\frac{a}{\omega}\sin\omega\varepsilon.
\end{aligned}\tag{5.14}$$

Zatem dokładna dyskretyzacja (5.12) dana jest przez (5.14) z ω oraz a zdefiniowanymi równaniami (5.13). Aby porównać (5.11) z (5.14), (5.13) przepiszmy (5.11) jak następuje:

$$\begin{aligned}
&x_{n+1}=x_n+\frac{\delta_n p_n}{1+\frac{1}{4}\delta_n^2V''(x_n)}-\frac{\frac{1}{2}\delta_n^2V'(x_n)}{1+\frac{1}{4}\delta_n^2V''(x_n)},\\
&p_{n+1}=\frac{1-\frac{1}{4}\delta_n^2V''(x_n)}{1+\frac{1}{4}\delta_n^2V''(x_n)}p_n-\frac{\delta_nV'(x_n)}{1+\frac{1}{4}\delta_n^2V''(x_n)},
\end{aligned}\tag{5.15}$$

Z porównania ostatnich członów (5.15) z (5.14), (5.13) dostajemy bezpośrednio $\delta_n=\frac{2}{\omega}\tan\frac{\varepsilon\omega}{2}$. Stąd wynika, że $\delta_n^2V''(x_n)=4\tan^2\frac{\varepsilon\omega}{2}$ i łatwo sprawdzić, że wówczas (5.15) staje się identyczne z (5.14). W ten sposób otrzymujemy pierwsze z równań (5.6). Pozostałe przypadki można wyprowadzić od podstaw w analogiczny sposób. Wygodniej jednak zauważyć, iż trzeci przypadek jest konsekwencją formalnego podstawienia urojonego ω, natomiast przypadek drugi można otrzymać poprzez przejście graniczne $\varepsilon \to 0$.

Należy podkreślić, że δ_n jest funkcją praktycznie stałą, $\delta_n = \varepsilon + O(\varepsilon^3)$, jednak okazuje się, że te bardzo małe zmiany mają zaskakująco silny korzystny wpływ na dokładność otrzymanego schematu numerycznego.

Podstawienie $\varepsilon \to \delta_n$ pracuje bardzo dobrze w przypadku układu równań pierwszego rzędu (5.5). Stosując to podstawienie w równaniu dyskretnym drugiego stopnia na x_n (równanie (4.50) w rozdziale 4), otrzymamy schemat różnicowy, który jest tylko nieco lepszy od tego ze stałą wartością $\delta = \delta_0$ i jest jednocześnie znacznie gorszy od schematu (5.5). Zaznaczmy, że równanie drugiego stopnia wynikające ze schematu (5.5) ma postać

$$\frac{x_{n+1} - x_n}{\delta_n} - \frac{x_n - x_{n-1}}{\delta_{n-1}} = -\frac{\delta_n}{2}\left(\frac{V_{n+1} - V_n}{x_{n+1} - x_n}\right) - \frac{\delta_{n-1}}{2}\left(\frac{V_n - V_{n-1}}{x_n - x_{n-1}}\right) \tag{5.16}$$

gdzie $V_n := V(x_n)$. To równanie zawiera zarówno δ_n jak i δ_{n-1}, zatem nie może być otrzymane z równania (4.50) poprzez proste podstawienie $\varepsilon \to \delta_n$. Jednak dla $\varepsilon \approx 0$ mamy $\delta_n \approx \delta_{n-1} \approx \varepsilon$ i dopiero w tej granicy równania (5.16) i (4.50) stają się identyczne. .

Układ równań (5.1) jest oczywiście symetryczny (odwracalny w czasie), ale jego lokalnie dokładna dyskretyzacja (5.5), (5.6) symetryczna nie jest. Zmieniając nieco definicję ω_n, mianowicie kładąc we wzorach (5.5), (5.6), (5.7)

$$\omega_n = \sqrt{\left|V''\left(\frac{x_n + x_{n+1}}{2}\right)\right|}, \tag{5.17}$$

otrzymujemy schemat numeryczny. który jest odwracalny w czasie. Nazwiemy go *symetryczną modyfikacją lokalnie dokładnego dyskretnego gradientu*, w skrócie GR-SLEX. Wydawałoby się, że powinien on być lepszy od GR-LEX, ale tak nie jest. Eksperymenty numeryczne wskazują na podobną dokładność obu tych metod. W zależności od wyboru parametrów niewielką przewagę osiąga raz jedna, raz druga.

5.4 Rząd rozpatrywanych metod

Rząd schematu numerycznego jest równy N jeśli $\left|\vec{x}_{n+1} - \vec{x}(t+\varepsilon)\right| = O(\varepsilon^{N+1})$, przy założeniu, że $\vec{x}_n = \vec{x}(t)$, gdzie $\vec{x} = \vec{x}(t)$ jest rozwiązanie dokładnym. W naszym przypadku wektorem $\vec{x}$ jest $\vec{x} = (x, p)$.

Układ równań (5.5) (gdzie $x_n \equiv x$ i $p_n \equiv p$ są dane, a $\delta = \delta_n$ jest małym parametrem) w sposób niejawny definiuje x_{n+1} i p_{n+1}. Z tego względu, stosując wzór na pochodną funkcji uwikłanej, możemy otrzymać jawne wyrażenie na x_{n+1} i p_{n+1} w postaci odpowiedniego szeregu Taylora:

$$
\begin{aligned}
x_{n+1} &= x + p\delta - \frac{1}{2}V'\delta^2 - \frac{1}{4}pV''\delta^3 + \frac{1}{24}(3V'V'' - 2p^2V''')\delta^4 + O(\delta^5),\\
p_{n+1} &= p - V'\delta - \frac{1}{2}pV''\delta^2 + \frac{1}{12}(3V'V'' - 2p^2V''')\delta^3\\
&\quad - \frac{1}{24}(4pV'V''' + 3p(V'')^2 - p^3V^{(4)})\delta^4 + O(\delta^5).
\end{aligned}
\tag{5.18}
$$

Natomiast korzystając z równań różniczkowych (5.1) (i ich różniczkowych konsekwencji) możemy rozwinąć w szereg Taylora $x(t + \varepsilon)$ i $p(t + \varepsilon)$:

$$
\begin{aligned}
x(t+\varepsilon) &= x + p\varepsilon - \frac{1}{2}V'\varepsilon^2 - \frac{1}{6}pV''\varepsilon^3 + \frac{1}{24}(V'V'' - p^2V''')\varepsilon^4 + O(\varepsilon^5),\\
p(t+\varepsilon) &= p - V'\varepsilon - \frac{1}{2}pV''\varepsilon^2 + \frac{1}{6}(V'V'' - p^2V''')\varepsilon^3\\
&\quad + \frac{1}{24}(3pV'V''' + p(V'')^2 - p^3V^{(4)})\varepsilon^4 + O(\varepsilon^5).
\end{aligned}
\tag{5.19}
$$

Ostatnim krokiem jest podstawienie odpowiedniego $\delta = \delta(\varepsilon)$ (odpowiadającego metodzie, której rząd badamy), a ściślej rozwinięcia tej funkcji w szereg względem ε, do wzoru (5.18) i porównanie wyniku z szeregiem (5.19).

Metoda dyskretnego gradientu (GR) odpowiada funkcji $\delta(\varepsilon) = \varepsilon$. W takim przypadku

$$
\begin{aligned}
x_{n+1} - x(t+\varepsilon) &= -\frac{1}{12}pV''\varepsilon^3 + O(\varepsilon^4),\\
p_{n+1} - p(t+\varepsilon) &= \frac{1}{12}V'V''\varepsilon^3 + O(\varepsilon^4).
\end{aligned}
\tag{5.20}
$$

Zatem metoda dyskretnego gradientu jest rzędu drugiego dla $V'' \neq 0$.

Natomiast w przypadku $V'' = 0$ (potencjał V zależący liniowo od x) schemat ten jest dokładny, czyli można powiedzieć, że jego rząd jest nieskończony. Interpretacja fizyczna dokładności tego schematu dla potencjału liniowego wynika z prostych („szkolnych") własności ruchu ze stałym przyspieszeniem. Pierwsze z równań (5.2) ilustruje znany fakt, iż prędkość średnia w takim ruchu jest równa średniej arytmetycznej prędkości początkowej i końcowej. Drugie równanie stwierdza po prostu, że przyspieszenie średnie na dowolnym odcinku jest w takim ruchu stałe.

Szereg Taylora dla δ_n danego wzorami (5.6) wyraża się następującym wyrażeniem (takim samym w każdym z trzech podprzypadków):

$$
\delta(\varepsilon) = \varepsilon + \frac{1}{12}V''\varepsilon^3 + \frac{1}{120}(V'')^2\varepsilon^5 + O(\varepsilon^7),
\tag{5.21}
$$

dlatego w przypadku schematu GR-LEX wzory (5.18) mają następujące rozwinięcie w szereg względem parametru ε:

$$
\begin{aligned}
x_{n+1} &= x + p\varepsilon - \frac{1}{2}V'\varepsilon^2 - \frac{1}{6}pV''\varepsilon^3 + \frac{1}{24}(V'V'' - 2p^2V''')\varepsilon^4 + O(\varepsilon^5),\\
p_{n+1} &= p - V'\varepsilon - \frac{1}{2}pV''\varepsilon^2 + \frac{1}{6}(V'V'' - p^2V''')\varepsilon^3\\
&\quad + \frac{1}{24}(4pV'V''' + p(V'')^2 - p^3V^{(4)})\varepsilon^4 + O(\varepsilon^5).
\end{aligned}
\tag{5.22}
$$

Stąd

$$
\begin{aligned}
x_{n+1} - x(t+\varepsilon) &= -\frac{1}{24}p^2V'''\varepsilon^4 + O(\varepsilon^5),\\
p_{n+1} - p(t+\varepsilon) &= \frac{1}{24}pV'V''\varepsilon^4 + O(\varepsilon^5),
\end{aligned}
\tag{5.23}
$$

co pozwala stwierdzić, że metoda GR-LEX jest trzeciego rzędu.

W przypadku schematu MOD-GR parametr δ zależy tylko od ε i jest dany wzorem (5.4), który można łatwo rozwinąć w szereg:

$$
\delta(\varepsilon) = \varepsilon + \frac{1}{12}V_0''\varepsilon^3 + \frac{1}{120}(V_0'')^2\varepsilon^5 + O(\varepsilon^7),
\tag{5.24}
$$

gdzie $V_0'' \equiv V''(0)$. W tym przypadku (5.18) przyjmuje postać:

$$
\begin{aligned}
x_{n+1} &= x + p\varepsilon - \frac{1}{2}V'\varepsilon^2 - \frac{1}{12}p(3V'' - V_0'')\varepsilon^3 + O(\varepsilon^4),\\
p_{n+1} &= p - V'\varepsilon - \frac{1}{2}pV''\varepsilon^2 + \frac{1}{12}(3V'V'' - V'V_0'' - 2p^2V''')\varepsilon^3 + O(\varepsilon^4),
\end{aligned}
\tag{5.25}
$$

co daje w końcu

$$
\begin{aligned}
x_{n+1} - x(t+\varepsilon) &= \frac{1}{12}p(V_0'' - V'')\varepsilon^3 + O(\varepsilon^4),\\
p_{n+1} - p(t+\varepsilon) &= \frac{1}{12}V'(V'' - V_0'')\varepsilon^3 + O(\varepsilon^4).
\end{aligned}
\tag{5.26}
$$

Stąd wynika, że zmodyfikowana metoda dyskretnego gradientu, MOD-GR, jest rzędu drugiego.

Na zakończenie rozpatrzymy schemat GR-SLEX, dla którego ω_n dana jest wzorem (5.17). W tym przypadku mamy:

$$
\delta(\varepsilon) = \varepsilon + \frac{1}{12}V''\varepsilon^3 + \frac{1}{24}pV'''\varepsilon^4 + \frac{1}{24}\varepsilon^5\left(\frac{1}{5}(V'')^2 + \frac{1}{4}p^2V^{(4)} - \frac{1}{2}V'V'''\right) + O(\varepsilon^6)
\tag{5.27}
$$

oraz $x_{n+1} - x(t + \varepsilon) = O(\varepsilon^5)$, a także $p_{n+1} - p(t + \varepsilon) = O(\varepsilon^5)$. Okazuje się zatem, że schemat numeryczny GR-SLEX, zadany równaniami (5.5), (5.6) i (5.17), jest rzędu czwartego.

5.5 Eksperymenty numeryczne

Dokładność prezentowanych wyżej schematów numerycznych (MOD-GR, GR-LEX, GR-SLEX) została przetestowana na przykładzie wahadła matematycznego ($V(x) = -k\cos x$) oraz na potencjale Morse'a ($V(x) = \frac{1}{2}ke^{-2\alpha x} - ke^{-\alpha x}$). Nasze nowe schematy porównaliśmy ze standardowym schematem Störmera-Verleta oraz z metodą dyskretnego gradientu.

Ruch wahadła matematycznego jest zawsze periodyczny, podobnie jak interesujące z praktycznego punktu widzenia przypadki ruchu w potencjale Morse'a, dlatego też skupiliśmy uwagę na na ruchach okresowych, badając stałość ich numerycznego okresu. Sposób obliczania tego okresu został przedstawiony w rozdziale 4. Podobnie jak tam, zakładamy tu dla uproszczenia, że zawsze $x_0 = 0$.

Doświadczenia numeryczne potwierdziły, iż metoda dyskretnego gradientu jest bardzo stabilna. Trzy modyfikacje tej metody, MOD-GR, GR-LEX, i GR-SLEX, także odziedziczyły po niej tę cechę.

Co więcej, w obu przypadkach znane jest rozwiązanie ścisłe, dlatego możliwe było obliczenie względnego odchylenia od teorii okresu drgań jaki miały badane dyskretyzacje.

W obliczeniach przyjęliśmy, że wszystkie stałe występujące w równaniach ruchu są równe jedności ($m = 1$, $\alpha = 1$, $k = 1$). Ścisłe rozwiązanie równania wahadła matematycznego z warunkiem początkowym x(0) = 0, $p(0) = p_0$, dane jest przez

$$\sin\frac{x}{2} = \frac{p_0}{2}\,\mathrm{sn}\left(t, \frac{p_0}{2}\right) \text{ dla } p_0 < 2 \quad \text{oraz} \quad \sin\frac{x}{2} = \frac{p_0}{2}\,\mathrm{sn}\left(\frac{p_0 t}{2}, \frac{2}{p_0}\right) \text{ dla } p_0 > 2,$$

gdzie sn(u, k) oznacza jedną z funkcji eliptycznych Jacobiego (porównaj na przykład [82]). W przypadku granicznym ($p_0 = 2$) mamy $\sin\frac{x}{2} = \tanh t$.

Jeszcze lepszą sytuację mamy w przypadku potencjału Morse'a, gdzie wszystkie rozwiązania wyrażają się przez funkcje elementarne. Ruch periodyczny ($0 \le p_0 < 1$) opisany jest wzorem

$$x(t) = \ln\left(\frac{1 - {p_0}^2\cos\omega t + p_0\sin\omega t}{\omega^2}\right)$$

przy czym $\omega = \sqrt{1 - p_0^2}$. Gdy $p_0^2 > 1$, wówczas, definiując $\omega = \sqrt{\left|p_0^2 - 1\right|}$ dostajemy

$$x(t) = \ln\left(\frac{-1 + {p_0}^2 \cos \omega t + p_0 \sin \omega t}{\omega^2} \right).$$

Trajektoria krytyczna ($p_0 = 1$), czyli ruch wzdłuż separatrysy, zadana jest przez $x(t) = \ln\left(1 + p_0 t + \frac{1}{2} t^2\right)$.

5.5.1 Lokalnie dokładny predyktor

W praktycznej implemetacji schemat (5.5) używany był jako korektor, podczas gdy jako predyktor służyły następujące równania:

$$\begin{aligned} x_{n+1} &= x_n + \frac{\sin(\omega_n \varepsilon)}{\omega_n} p_n - \frac{1 - \cos(\omega_n \varepsilon)}{\omega_n^2} V'(x_n), \quad [V''(x_n) > 0], \\ x_{n+1} &= x_n + \varepsilon p_n - \frac{1}{2} \varepsilon^2 V'(x_n), \quad [V''(x_n) = 0], \\ x_{n+1} &= x_n + \frac{\sinh(\omega_n \varepsilon)}{\omega_n} p_n - \frac{1 - \cosh(\omega_n \varepsilon)}{\omega_n^2} V'(x_n), \quad [V''(x_n) < 0]. \end{aligned} \tag{5.28}$$

Aby otrzymać zależności (5.28), należy wyeliminować p_{n+1} z układu (5.5) i rozwinąć wynik w szereg Taylora względem x_{n+1} - x_n, pozostawiając tylko wyrazy liniowe. Układ (5.28) jest tego samego (trzeciego) rzędu co (5.5), a ponadto obydwa schematy różnicowe są lokalnie dokładne.

Taki sposób otrzymywania schematów numerycznych został zaproponowany już kilkadziesiąt lat temu [81], a obecnie jest rozwijany pod nazwą schematy wykładnicze [71]. Równania (5.28) są bardzo dobrym predyktorem dla małych ε i krótkich czasów, w przypadku dłuższych odcinków czasowych są niewłaściwe i dają rozwiązania o złym zachowaniu jakościowym.

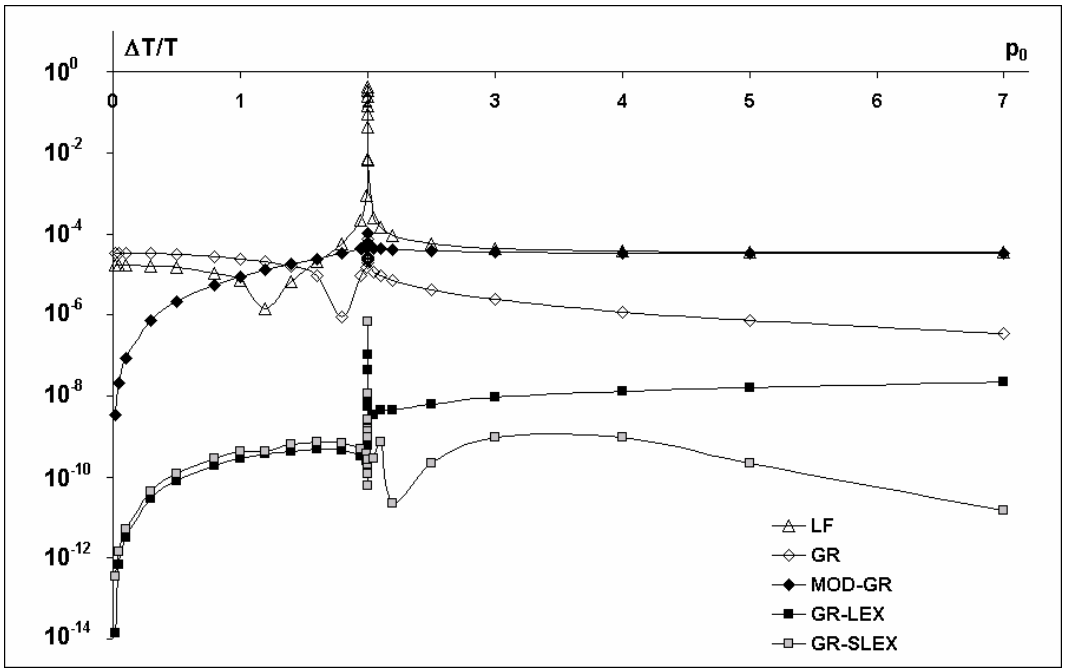
5.5.2 Rozwiązania iteracyjne równań uwikłanych

Schemat numeryczny dyskretnego gradientu i wszystkie jego odmiany dają rozwiązania w postaci niejawnej. Do ich rozwiązywania zastosowano dwie metody: punktu stałego i Newtona. W obu przypadkach warunkiem zakończenia iteracji było osiągnięcie dokładności na poziomie 10^{-16}. W pewnych punktach (więcej szczegółów w rozdziale 4) należących do obszarów, gdzie funkcja $x = x(t)$ jest bardzo spłaszczona, niekorzystne warunki początkowe sprawiały, że osiągalna dokładność procedury iteracyjnej (porównaj rozdział 11) stawała się nieco większa niż 10^{-16}. W efekcie pojawiały się chaotyczne oscylacje ograniczone promieniem zbieżności iteracji (rzędu 10^{-15}). W takich przypadkach

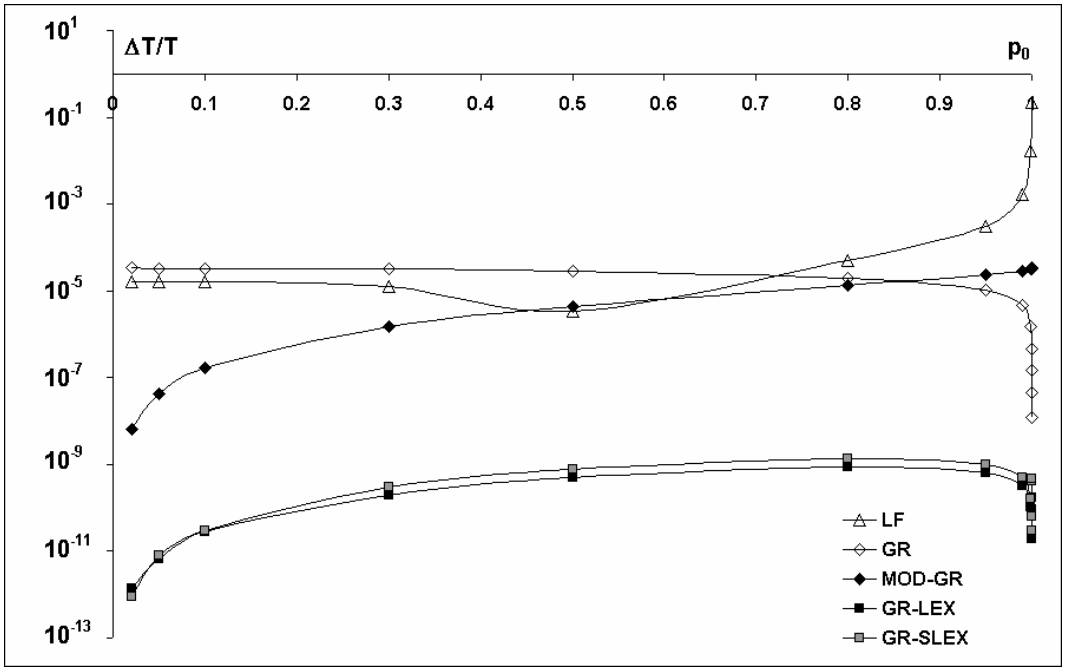
obliczenia były zatrzymywane po wykonaniu 100 iteracji (w metodzie punktu stałego) lub 15 iteracji (w metodzie Newtona). Średnia liczba iteracji przypadająca na jeden krok dyskretyzacji silnie zależy od ε i wykazuje też pewną zależność od pędu początkowego p_0. Przy małych ε (np. ε = 0.02) zakładana dokładność jest osiągana już po 2 krokach w metodzie Newtona i po 5 w metodzie Banacha. Liczba iteracji rośnie wraz z krokiem czasowym i wynosi odpowiednio 3 i 12-13 przy ε = 0.2 oraz 3-4 i 21-24 przy ε = 0.5 (dokładna średnia liczba iteracji zależy od p_0). Okazało się, że metoda Newtona potrzebuje ok. 3.3 raza więcej czasu na jeden krok niż metoda punktu stałego. Z tej przyczyny obie metody pracowały w naszym przypadku ze zbliżoną szybkością. Przy małych krokach czasowych (ε < 0.2) koszt obliczeń dla obu metod był praktycznie taki sam (i przewyższał 6-10 razy analogiczny koszt dla schematu leap-frog). Dla najmniejszych wartości ε (ε ≈ 0.02), metoda punktu stałego okazywała się nieco szybsza od metody Newtona, a dla największych (ε ≈ 0.9) dwukrotnie od niej wolniejsza.

5.5.3 Względne odchylenie od okresu teoretycznego

Dokładność prezentowanych w tym rozdziale schematów numerycznych jest zaskakująco wysoka, zwłaszcza dla małych (ale nie koniecznie bardzo małych) kroków czasowych. Jako przykład zeprezentowany zostanie przypadek ε = 0.02 (wykresy 5.1 i 5.2).



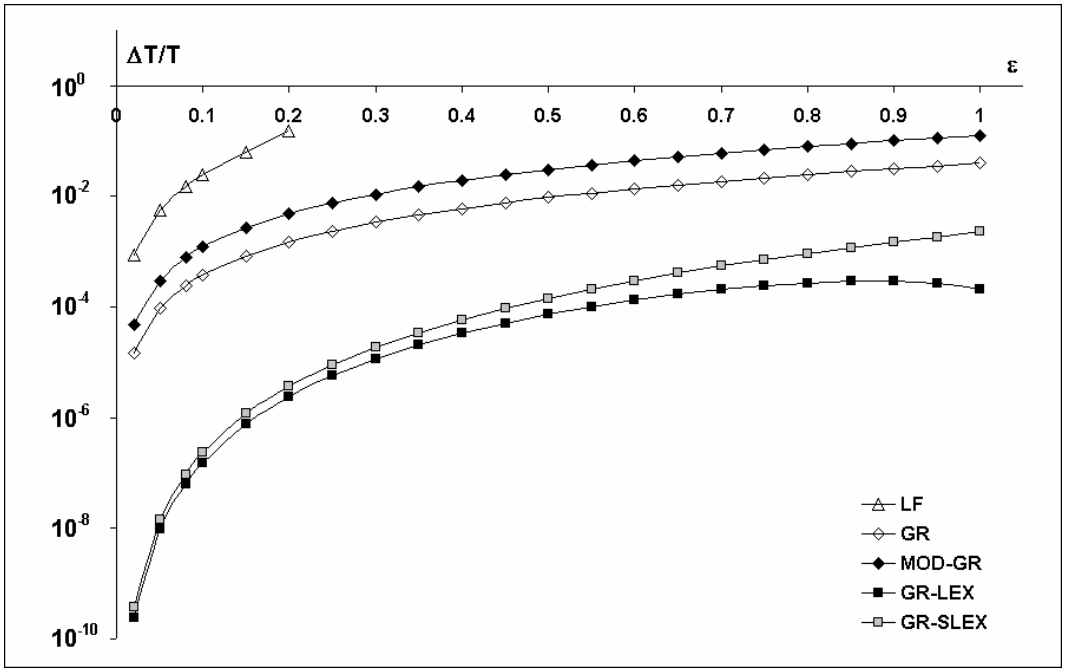
Wykres 5.1. Wahadło matematyczne. Względne odchylenie od okresu teoretycznego jako funkcja p_0 dla ε = 0.02.



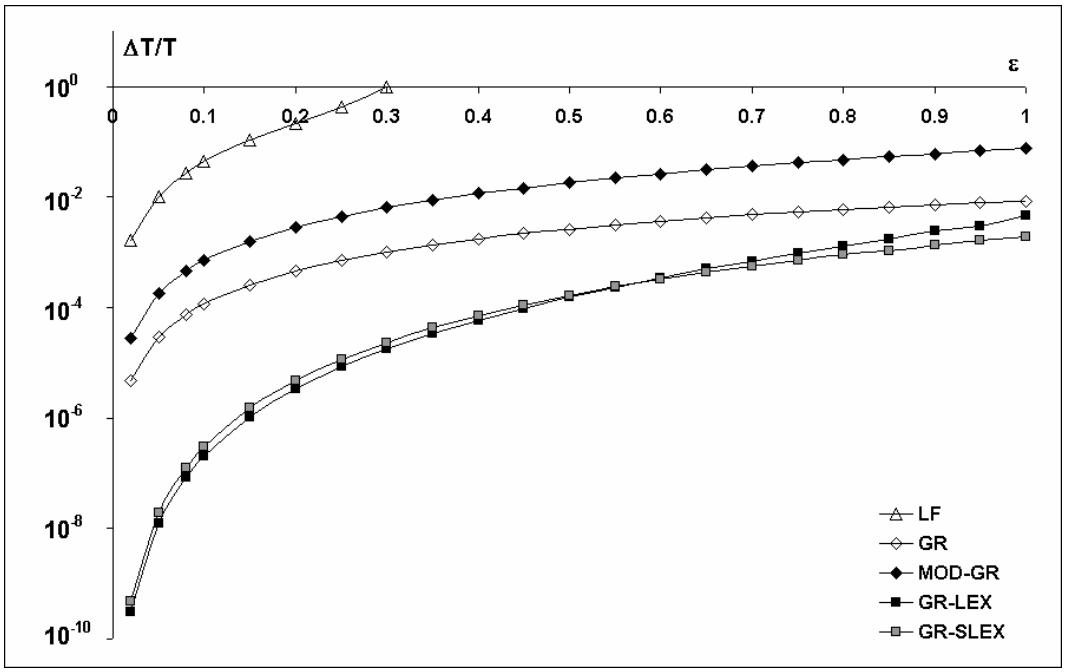
Wykres 5.2. Potencjał Morse'a. Względne odchylenie od okresu teoretycznego jako funkcja p_0 dla ε = 0.02.

Wstępne porównanie wykresów pozwala stwierdzić duże podobieństwo przebiegu krzywych uzyskanych dla obszaru oscylacyjnego wahadła i potencjału Morse'a oraz o ok. dwa rzędy wielkości mniejszą dokładność odwzorowania okresu przez wprowadzone w tym rozdziale schematy całkowania w przypadku drugiego z badanych potencjałów.

Przy małych oscylacjach (p_0 = 0.02) dokładność lokalnie dokładnego dyskretnego gradientu jest o 5 rzędów wielkości w przypadku wahadła i o 4 rzędy wielkości w przypadku potencjału Morse'a większa niż osiągana przez zmodyfikowany dyskretny gradient. Jednocześnie schemat ten osiąga dokładność o 9 (wahadło) i 7 (Morse) rzędów wielkości większą niż metody *leap-frog* i zwykły dyskretny gradient. Z wykresów 5.1 i 5.2 widzimy, że lokalnie dokładny dyskretny gradient i jego symetryczna modyfikacja są znacznie dokładniejsze od każdej z rozpatrywanych metod przy dowolnych warunkach początkowych (zauważmy, że skala na osi pionowej jest logarytmiczna). W przypadku ruchu rotacyjnego wahadła, metody zaproponowane w tym rozdziale są lepsze od pozostałych, uwzględnionych w porównaniach o około 4 rzędy wielkości. Symetryczna modyfikacja lokalnie dokładnego dyskretnego gradientu góruje nad wszystkimi pozostałymi metodami w obszarze dużych energii ($p_0 > 2$ w przypadku wahadła).



Wykres 5.3. Wahadło matematyczne. Względne odchylenie od okresu teoretycznego jako funkcja ε dla p_0 = 1.99 (okres teoretyczny T_{th} = 14.787500329575).



Wykres 5.4. Potencjał Morse'a. Względne odchylenie od okresu teoretycznego jako funkcja ε dla p_0 = 0.99 (okres teoretyczny T_{th} = 44.540319718441).

Na wykresach 5.3 i 5.4 zaprezentowano zależność względnego odchylenia okresu badanych schematów numerycznych od teorii dla dużych drgań (p_0 = 1.99 dla wahadła i p_0 = 0.99 dla potencjału Morse'a).

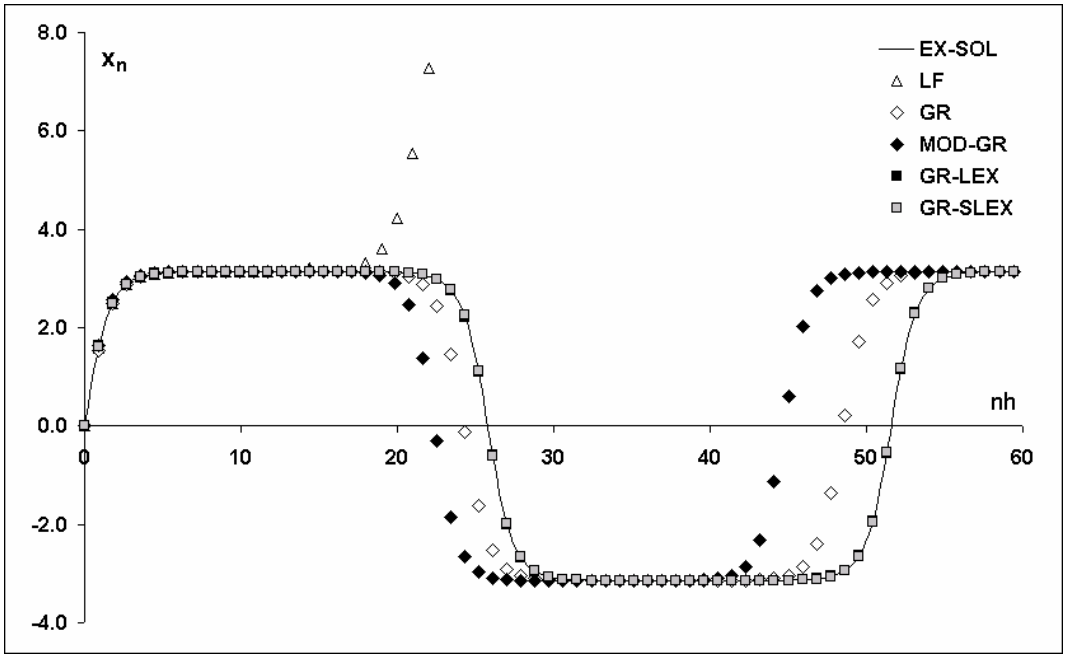
Lokalnie dokładny dyskretny gradient jest ponownie najlepszy. Obserwujemy też nieco mniejszą dokładność jego symetrycznej modyfikacji, chociaż w przypadku potencjału Morse'a przy dużych ε sytuacja się odwraca. Standardowy dyskretny gradient pod względem dokładności pozostaje daleko w tyle i jedynie w przypadku dużych kroków czasowych zbliża się do dwóch prezentowanych w tym rozdziale metod. Odnotujmy też ograniczoną stosowalność schematu *leap-frog*, który przy stosunkowo małych wartościach ε przestaje nawet jakościowo odtwarzać oscylacje.

Biorąc pod uwagę koszt obliczeniowy metod (który w praktyce redukuje się do wybierania mniejszego kroku czasowego dla schematu *leap-frog*) stwierdzamy, że po skorygowaniu metoda ta daje podobne rezultaty jak zwykły dyskretny gradient, jednak dwie wprowadzone w tym rozdziale metody są nadal znacznie lepsze. Tylko w pewnych szczególnych warunkach (np. „rezonansowa" wartość p_0 = 1.21 dla wahadła (rozdział 4)) skorygowany schemat *leap-frog* może być z nimi porównywalny.

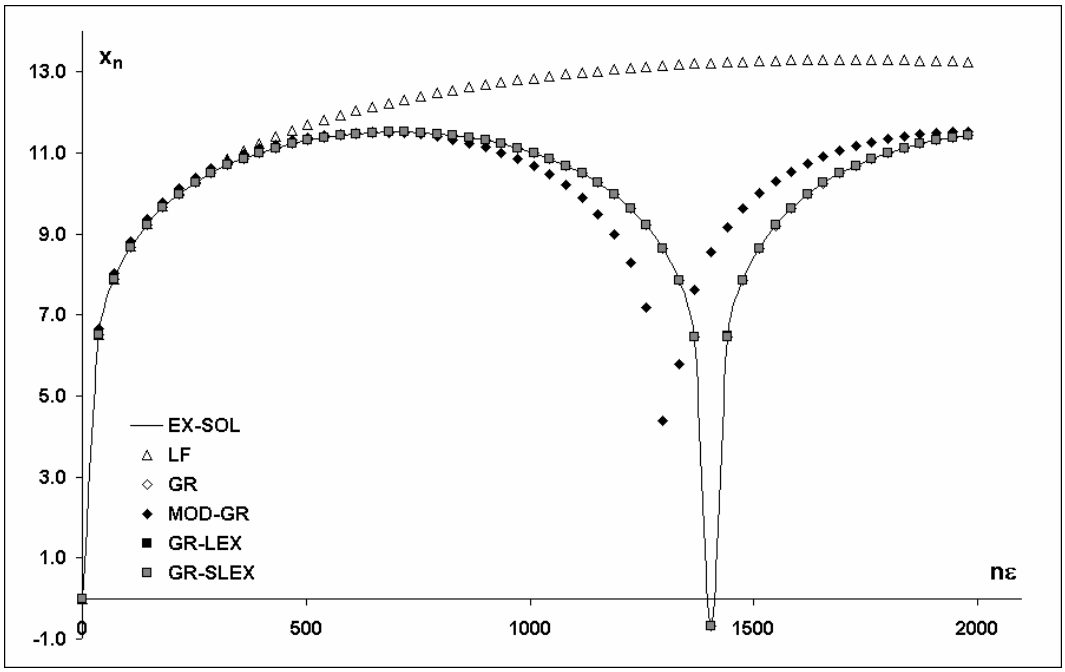
5.5.4 Bliskie otoczenie separatrysy i trajektorii krytycznej

Najtrudniejszym obszarem do symulacji numerycznej jest sąsiedztwo separatrysy ($p_0 \approx 2$, wykres 5.5) dla wahadła i trajektorii krytycznej ($p_0 \approx 1$, wykres 5.6) dla potencjału Morse'a.

Obydwa wykresy pokazują, że standardowy dyskretny gradient stosunkowo dobrze zachowuje się w tym obszarze (rozdział 4), co szczególnie dobrze jest widoczne na wykresie dla potencjału Morse'a, gdzie p_0 nie zbliża się tak bardzo do trajektorii krytycznej, jak w przypadku wahadła. Lokalnie dokładny dyskretny gradient wraz z symetryczną modyfikacją dają praktycznie takie same wyniki (na wykresie 5.6 dodatkowo pokrywające się prawie z dyskretnym gradientem). Podkreślmy, że chociaż w przypadku wahadła trajektoria jest bardzo bliska separatrysie ($|p_0 - 2| = 10^{-10}$) i ε jest bardzo duży, to prezentowane tu metody symulują bardzo dokładnie jego ruch. Punkty dyskretne x_n leżą praktycznie na krzywej ciągłej reprezentującej rozwiązanie dokładne. Pozostałe dwie metody gradientowe również dają dobre wyniki (przynajmniej jakościowo) podczas, gdy schemat *leap-frog* na wykresie 5.5 zupełnie nie odtwarza teorii, a na wykresie 5.6 bardzo się z nią rozmija.



Wykres 5.5. Wahadło matematyczne. x_n jako funkcja n bardzo blisko separatrysy (p_0 = 1.9999999999), ε = 0.9 dla schematów gradientowych i ε = 0.001 dla schematu leap-frog. (okres teoretyczny T_{th} = 51.59687914). Linia ciągła odpowiada rozwiązaniu teoretycznemu.



Wykres 5.6. Potencjał Morse'a. x_n jako funkcja n bardzo blisko trajektorii krytycznej (p_0 = 0.99999), ε = 0.9 dla schematów gradientowych i ε = 0.01 dla schematu leap-frog, (okres teoretyczny T_{th} = 1404.9664586). Linia ciągła odpowiada rozwiązaniu teoretycznemu.

Koszt obliczeniowy prezentowanych tu nowych niejawnych schematów gradientowych jest większy od kosztu schematów jawnych. Szacunkowo, w przypadku danych z wykresów 5.5 i 5.6, koszt jednego kroku dyskretyzacji dla metod gradientowych (iteracja Newtona) jest ok. 13 razy większy niż dla schematu *leap-frog* (w iteracji metodą punktu stałego ten współczynnik jest bliski 29). Okazuje się jednak, że dla omawianych tu warunków początkowych ruchu schemat *leap-frog* nie może być istotnie poprawiony nawet poprzez zdecydowane zmniejszenie kroku czasowego. Zwróćmy uwagę, że na wykresie 5.5 i 5.6 krok czasowy metody *leap-frog* jest odpowiednio 900 i 90 razy mniejszy niż w przypadku pozostałych metod, co sprawia, że jej koszt obliczeniowy staje się bardzo wysoki w porównaniu ze schematem dyskretnego gradientu oraz jego modyfikacjami.

5.6 Podsumowanie

Zaproponowane w tym rozdziale schematy numeryczne GR-LEX i GR-SLEX, zadane wzorami (5.5) i (5.6) (lokalnie dokładny dyskretny gradient i jego symetryczna modyfikacja, odpowiadające odpowiednio równaniom (5.7) i (5.17)) wykazały się wieloma zaletami:

(i) dokładnym zachowaniem całki energii (równanie 5.3),
(ii) wyższym rzędem (odpowiednio trzecim i czwartym) w porównaniu z metodą dyskretnego gradientu (rząd drugi),
(iii) wysoką stabilnością i dokładnością,
(iv) bardzo dobrym zachowaniem jakosciowym rozwiązania numerycznego w długim okresie czasu.

Zatem modyfikacje te w istotny sposób poprawiają metodę dyskretnego gradientu, zachowując przy tym wszystkie zalety tej metody. W pracy tej rozpatrujemy dokładnie tylko przypadek jednowymiarowy. Przypadek wielowymiarowy wychodzi poza zakres pracy, ale analogiczne algorytmy lokalnie dokładne istnieją i w tym przypadku [19].

Podkreślmy jednak, że schematy GR-LEX i GR-SLEX, podobnie jak wszystkie metody gradientowe, nie są symplektyczne ani też nie zachowują objętości przestrzeni fazowej. Co więcej, metoda GR-LEX nie jest odwracalna w czasie. Symetryczna (odwracalna w czasie) metoda GR-SLEX mimo, iż ma wyższy rząd (czwarty), nie wykazuje się większą precyzją niż metoda GR-LEX (w ruchu oscylacyjnym jest zwykle nawet mniej dokładna). Może to sugerować, że lokalna dokładność ma istotne zalety i warto ją stosować nawet kosztem złamania symetrii schematu numerycznego. Niewątpliwie kwestia ta zasługuje na dalsze badania.

6 Dyskretyzacje równań Lotki-Volterry zachowujące trajektorie

6.1 Wprowadzenie

W rozdziale tym będzie rozważany najprostszy 2 wymiarowy model Lotki-Volterry, zadany równaniami

$$\begin{aligned} \dot{x} &= Ax + Bxy \\ \dot{y} &= Cy + Dxy \end{aligned} \tag{6.1}$$

gdzie $x = x(t) \in \mathbf{R}$, $y = y(t) \in \mathbf{R}$ natomiast A, B, C, D są stałymi. W latach 20-tych ubiegłego wieku Alfred Lotka zastosował ten model do opisu inwazji pasożytów, a Vito Volterra zinterpretował przy jego pomocy dane o stanie zarybienia Morza Adriatyckiego [41]. Model ten, jako stosunkowo prosty układ nieliniowy, dość często pojawia się w biologii i ekologii (np. do opisu współistnienia dwóch gatunków: drapieżników i ich ofiar), a także fizyce i chemii (np. przy opisie przebiegu czasowego reakcji chemicznych).

Wiele badań [69, 76, 90, 91, 92, 96] poświęcono symplektycznym dyskretyzacjom równań Lotki-Volterry, które dają prawidłowe jakościowe zachowanie rozwiązań. Znacznie mniej uwagi poświęcono schematom numerycznym zachowującym energię. Schemat tego typu, inny od dyskretyzacji przedstawionych w niniejszej pracy, można znaleźć w monografii [1].

Celem tego rozdziału jest pokazanie na przykładzie modelu (6.1) przewagi jaką mają odpowiednio zmodyfikowane (ale zachowujące energię) metody dyskretnego gradientu nad schematami symplektycznymi. Przedstawione zostaną dyskretyzacje, które nie tylko zachowują dokładnie (z dokładnością do błędów numerycznych) wszystkie trajektorie, ale odtwarzają też z bardzo dużą dokładnością ewolucję czasową układu.

6.2 Metoda dyskretnego gradientu

Układ (6.1) posiada niezmiennik (całkę energii) dany wzorem

$$H(x, y) = A\ln y + By - C\ln x - Dx \tag{6.2}$$

i można go zapisać w tzw. niekanonicznej postaci hamiltonowskiej (zwanej też układem Poissona):

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & xy \\ -xy & 0 \end{pmatrix} \nabla H(x, y) \tag{6.3}$$

gdzie $\nabla H(x,y) = \left(\frac{\partial H}{\partial x}, \frac{\partial H}{\partial y} \right)^T$, a macierz poprzedzająca gradient hamiltonianu nazywa się macierzą Poissona (patrz np. [10, 41]). Bezpośrednim rachunkiem można sprawdzić, że układ (6.3) jest identyczny z (6.1).

Macierz Poissona musi być antysymetryczna oraz spełniać dodatkowe warunki wynikające z tożsamości Jacobiego dla nawiasu Poissona. Warto jednak zaznaczyć, że metoda dyskretnego gradientu [65, 66] nie wymaga założenia hamiltonowskości i spełnienia tych dodatkowych warunków. Wystarczy, aby macierz stojąca w miejscu macierzy Poissona była antysymetryczna.

W przypadku $H(x,y) = T(y) + V(x)$ standardowa metoda dyskretnego gradientu prowadzi do równań

$$\begin{aligned} \frac{x_{n+1} - x_n}{\varepsilon} &= x_n y_n \frac{T(y_{n+1}) - T(y_n)}{y_{n+1} - y_n}, \\ \frac{y_{n+1} - y_n}{\varepsilon} &= -x_n y_n \frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n}. \end{aligned} \tag{6.4}$$

Podstawiając $T(y) = A\ln y + By$ i $V(x) = -C\ln x - Dx$ (odpowiadające równaniom Lotki-Volterry), otrzymujemy

$$\begin{aligned} \frac{x_{n+1} - x_n}{\varepsilon} &= x_n y_n \left(B + \frac{A \ln \left| \frac{y_{n+1}}{y_n} \right|}{y_{n+1} - y_n} \right), \\ \frac{y_{n+1} - y_n}{\varepsilon} &= x_n y_n \left(D + \frac{C \ln \left| \frac{x_{n+1}}{x_n} \right|}{x_{n+1} - x_n} \right). \end{aligned} \tag{6.5}$$

Eksperymenty numeryczne zostały przeprowadzone dla układu

$$\begin{aligned} \dot{x} &= x(y-2), \\ \dot{y} &= y(1-x), \end{aligned} \tag{6.6}$$

(czyli A = -2, B = 1, C = 1, D = -1), który był standardowym przykładem rozważanym w monografii [41]. W tym przypadku:

$$H(x,y) = x + y - \ln|x| - 2\ln|y|. \tag{6.7}$$

6.3 Schematy numeryczne zachowujące trajektorie

Rozważmy następującą klasę schematów numerycznych

$$\begin{aligned} \frac{x_{n+1}-x_n}{\delta} &= x_n y_n \frac{T(y_{n+1})-T(y_n)}{y_{n+1}-y_n}, \\ \frac{y_{n+1}-y_n}{\delta} &= -x_n y_n \frac{V(x_{n+1})-V(x_n)}{x_{n+1}-x_n}, \end{aligned} \tag{6.8}$$

gdzie δ jest dowolną funkcją od $x_n, x_{n+1}, y_n, y_{n+1}, \varepsilon,$ spełniającą warunek

$$\lim_{\varepsilon\to 0}\frac{\delta(x_n, x_{n+1}, y_n, y_{n+1}, \varepsilon)}{\varepsilon} = 1, \tag{6.9}$$

który gwarantuje, że schemat ten jest zgodny.

Twierdzenie 1. *Dowolna dyskretyzacja należąca do klasy (6.8) zachowuje dokładnie (z dokładnością do błędów zaokrągleń) wszystkie trajektorie układu równań (*6.1*).*

Dowód: Po odwróceniu kolejności w jednym z równań, mnożymy oba równania (6.8) stronami. W wyniku otrzymujemy, że $H(x_n, y_n) = const$, czyli energia układu jest ściśle zachowana. W przypadku jednowymiarowym poziomica energii, czyli krzywa $H(x, y) = const$, jest torem ruchu. Zatem punkty dyskretne należą do tego toru ruchu, o ile tylko należy do niego punkt początkowy. □

Twierdzenie to może być łatwo uogólnione na dowolny układ jednowymiarowy dopuszczający sformułowanie w postaci Poissona. Dowód jest identyczny. Schemat dyskretnego gradientu i jego δ-modyfikacje ściśle zachowują wszystkie trajektorie w przestrzeni fazowej.

Twierdzenie 2. *Niech*

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & f(x,y) \\ -f(x,y) & 0 \end{pmatrix} \nabla H(x,y) \tag{6.10}$$

gdzie $\nabla H(x,y) = \left(\frac{\partial H}{\partial x}, \frac{\partial H}{\partial y}\right)^T$. *Wówczas schemat numeryczny*

$$\begin{aligned} \frac{x_{n+1}-x_n}{\delta_n} &= f(x_n, y_n)\frac{\Delta_2 H}{y_{n+1}-y_n}, \\ \frac{y_{n+1}-y_n}{\delta_n} &= -f(x_n, y_n)\frac{\Delta_1 H}{x_{n+1}-x_n}, \end{aligned} \tag{6.11}$$

gdzie

$$\Delta_1 H + \Delta_2 H = H(x_{n+1}, y_{n+1}) - H(x_n, y_n), \tag{6.12}$$

*zachowuje dokładnie wszystkie trajektorie w przestrzeni (x, y) układu (*6.10*).*

I w tym przypadku δ_n może zależeć w dowolny sposób od ε, x_n, y_n, x_{n+1}, y_{n+1} (również poprzez funkcję *f*, jeśli to konieczne) o ile tylko spełnia warunek (6.14).

W przypadku rozważanego przez nas hamiltonianu separowalnego $H(x, y) = T(y) + V(x)$ dyskretyzacja (6.11) sprowadza się do

$$\begin{aligned} \frac{x_{n+1} - x_n}{\delta_n} &= f(x_n, y_n)\frac{T(y_{n+1}) - T(y_n)}{y_{n+1} - y_n}, \\ \frac{y_{n+1} - y_n}{\delta_n} &= -f(x_n, y_n)\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n}. \end{aligned} \tag{6.13}$$

Wstępne testy pokazały, że własności numeryczne schematu (6.5) nie są zbyt dobre. Kluczem do sukcesu okazuje się zamiana zmiennych prowadząca do równoważnego układu cechującego się znacznie lepszymi własnościami numerycznymi.

Podstawiamy $x = e^p$, $y = e^q$, transformując w ten sposób (6.1) do postaci

$$\begin{aligned} \dot{q} &= A + Be^p, \\ \dot{p} &= C + De^q. \end{aligned} \tag{6.14}$$

Jest to kanoniczny układ hamiltonowski, gdyż $\dot{q} = \frac{\partial H}{\partial p}$, $\dot{p} = -\frac{\partial H}{\partial q}$, gdzie

$$H(q, p) = Ap + Be^p - Cq - De^q. \tag{6.15}$$

Jeśli $\frac{A}{B} < 0$ i $\frac{C}{D} < 0$, wówczas układ (6.14) ma punkt równowagi dla

$$p = \ln\left|\frac{A}{B}\right|, \quad q = \ln\left|\frac{C}{D}\right|. \tag{6.16}$$

Zauważmy, że $H(q, p) = T(p) + V(q)$, gdzie $T(p) = Ap + Be^p$ i $V(q) = -Cq - De^q$. Schemat dyskretnego gradientu zadany jest teraz równaniami.

$$\begin{aligned} \frac{q_{n+1} - q_n}{\varepsilon} &= \frac{T(p_{n+1}) - T(p_n)}{p_{n+1} - p_n}, \\ \frac{p_{n+1} - p_n}{\varepsilon} &= -\frac{V(q_{n+1}) - V(q_n)}{q_{n+1} - q_n}, \end{aligned} \tag{6.17}$$

a po podstawieniu $T(p) = Ap + Be^p$ i $V(q) = -Cq - De^q$ otrzymujemy:

$$\frac{q_{n+1}-q_n}{\varepsilon} = A + B\frac{e^{p_{n+1}}-e^{p_n}}{p_{n+1}-p_n},$$
$$\frac{p_{n+1}-p_n}{\varepsilon} = C + D\frac{e^{q_{n+1}}-e^{q_n}}{q_{n+1}-q_n}. \tag{6.18}$$

6.4 Metoda lokalnie dokładnego dyskretnego gradientu

Lokalnie dokładne schematy numeryczne dla równań Lotki-Volterry otrzymujemy tak samo jak w rozdziale 5. Dla równań (6.10) wynikiem jest schemat numeryczny postaci

$$\frac{x_{n+1}-x_n}{\delta(\bar{x},\bar{y})} = f(\bar{x},\bar{y})\frac{T(y_{n+1})-T(y_n)}{y_{n+1}-y_n},$$
$$\frac{y_{n+1}-y_n}{\delta(\bar{x},\bar{y})} = -f(\bar{x},\bar{y})\frac{V(x_{n+1})-V(x_n)}{x_{n+1}-x_n}, \tag{6.19}$$

gdzie

$$\delta = \frac{2}{\omega}\tan\frac{\omega\varepsilon}{2}, \quad \omega = \sqrt{T''(\bar{y})V''(\bar{x})}\,, \tag{6.20}$$

a konkretnie, dla równań (6.5),

$$\omega = \sqrt{\frac{-AC}{\bar{x}^2\bar{y}^2}}\,. \tag{6.21}$$

Podobnie jak w rozdziale 5, mamy dwa naturalne sposoby wyboru $\bar{x}$ i $\bar{y}$. Pierwszy to $\bar{x} = x_n$, $\bar{y} = y_n$, prowadzący do schematu GR-LEX, zaś drugi jest jego „symetryczną modyfikacją” $\bar{x} = \frac{1}{2}(x_n + x_{n+1})$, $\bar{y} = \frac{1}{2}(y_n + y_{n+1})$, dającą w wyniku schemat GR-SLEX.

W zmiennych *q* i *p*, które okazały się być bardziej odpowiednie do zastosowań numerycznych, klasa lokalnie dokładnych schematów dyskretnego gradientu ma postać

$$\frac{q_{n+1}-q_n}{\delta_n} = A + B\frac{e^{p_{n+1}}-e^{p_n}}{p_{n+1}-p_n},$$
$$\frac{p_{n+1}-p_n}{\delta_n} = C + D\frac{e^{q_{n+1}}-e^{q_n}}{q_{n+1}-q_n}, \tag{6.22}$$

gdzie

$$\delta_n = \frac{2}{\omega}\tan\frac{\omega\varepsilon}{2}, \quad \omega_n = \sqrt{T''(\bar{p})V''(\bar{q})}\,. \tag{6.23}$$

Kładąc $\overline{q} = q_n$ i $\overline{p} = p_n$ dostajemy schemat GR-LEX, gdzie

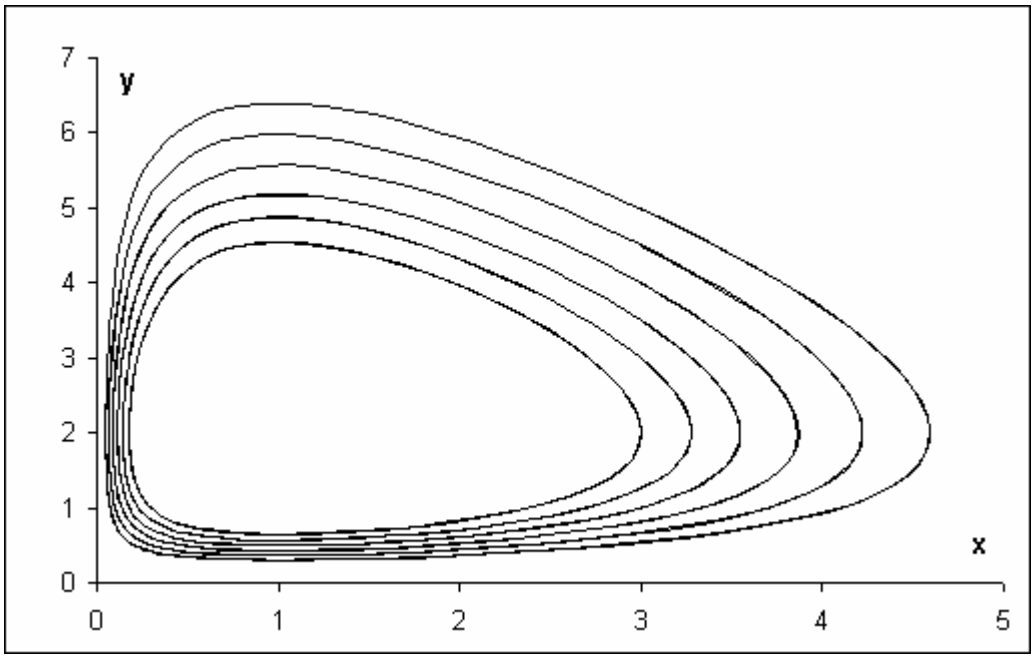
$$\omega = \sqrt{-BD}\exp\frac{p_n + q_n}{2}, \tag{6.24}$$

przyjmując z kolei („symetryczna modyfikacja”, czyli GR-SLEX) $\overline{q} = \frac{1}{2}(q_n + q_{n+1})$ i $\overline{p} = \frac{1}{2}(p_n + p_{n+1})$, otrzymujemy

$$\omega_n = \sqrt{-BD}\exp\frac{p_n + p_{n+1} + q_n + q_{n+1}}{4}. \tag{6.25}$$

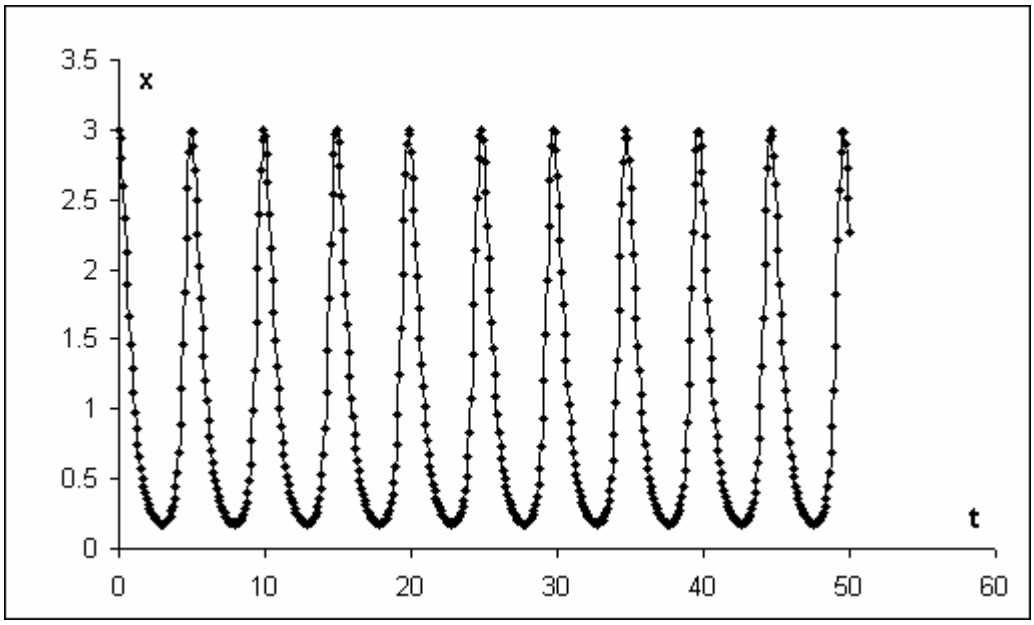
6.5 Eksperyment numeryczny

W doświadczeniu przyjęto A = -2, B = 1, C = 1, D = -1 i badano trajektorie takiego układu oraz przebiegi czasowe (okresy drgań) wykorzystując równania (6.17) i (6.22). Wszystkie testowane dyskretyzacje, na mocy twierdzeń z rozdziału 6.3, ściśle zachowują dokładne trajektorie. Ich przykładowy obraz umieszczono na wykresie 6.1.



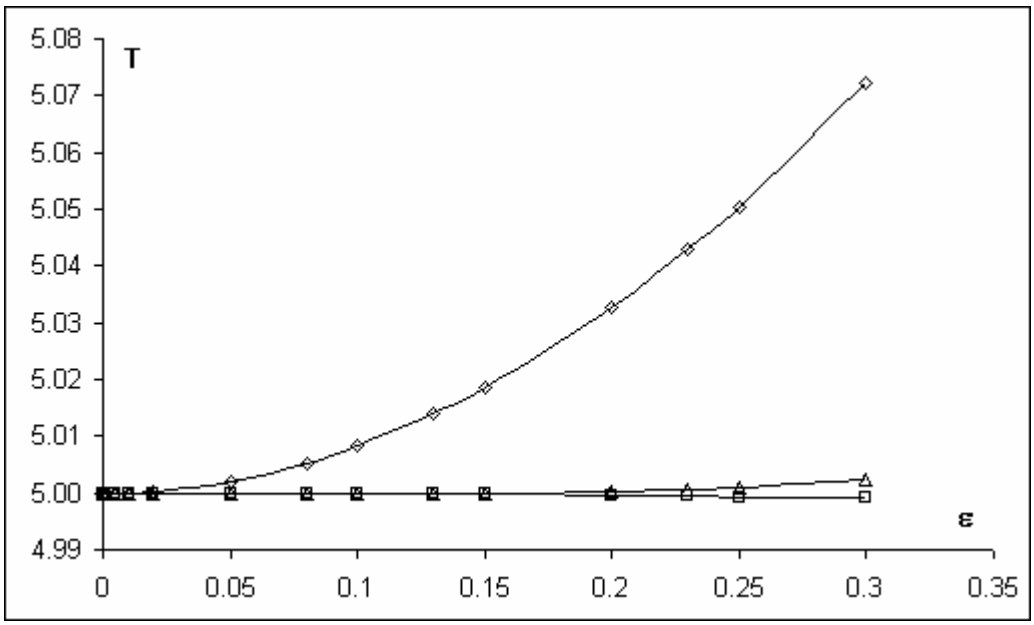
Wykres 6.1. Typowe (dokładne) trajektorie w badanym modelu Lotki-Volterry, metoda dyskretnego gradientu, x_0 = 3.0, $y_0 \in$ {2.0; 3.0; 3.5; 4.5; 5.0} (ε = 0.1).

Znajomość dokładnego kształtu krzywych $y(x)$ nie daje informacji o ewolucji czasowej układu $x(t)$ i $y(t)$. Można z góry założyć, że metoda dyskretnego gradientu odtwarza dobrze zachowanie układu tylko jakościowo.

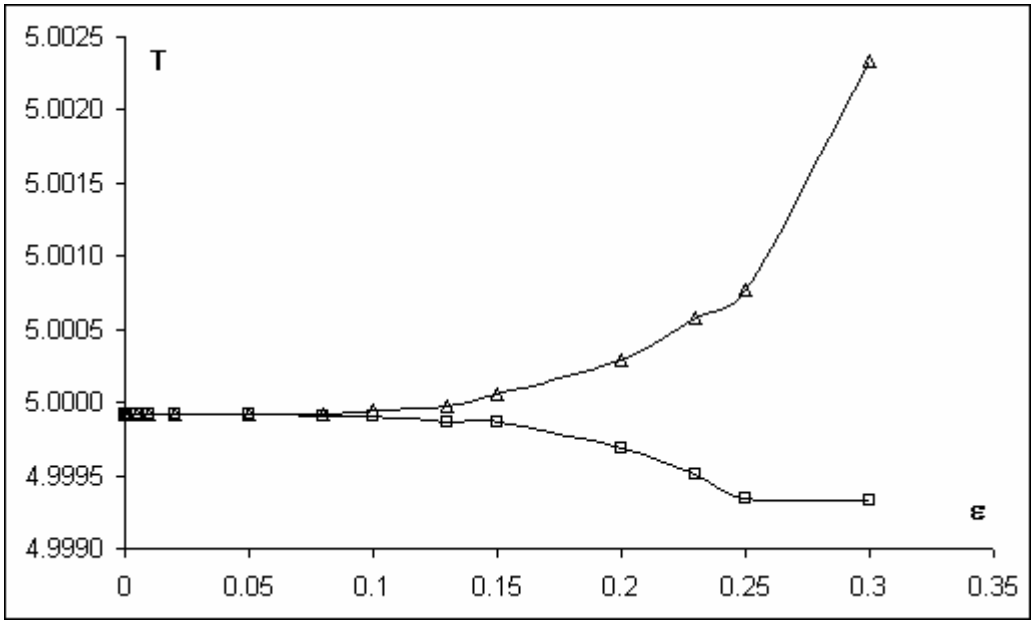


Wykres 6.2. Typowy obraz oscylacji czasowych w badanym modelu Lotki-Volterry, metoda dyskretnego gradientu, $x_0 = 3.5$, $y_0 = 2.0$ ($\varepsilon = 0.1$).

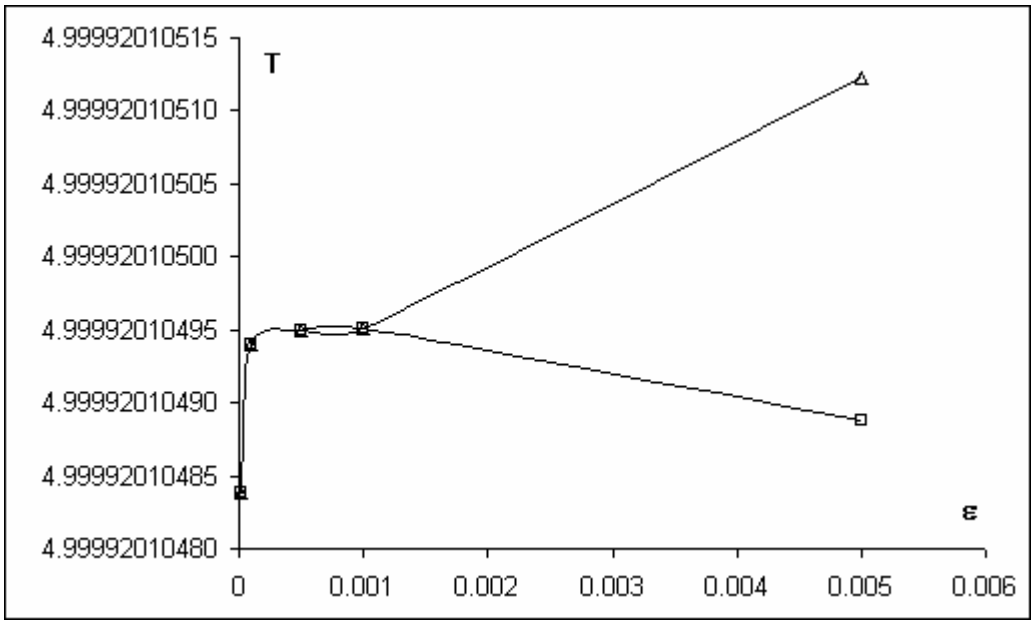
Kolejne trzy wykresy pokazują dokładność działania metody dyskretnego gradientu, lokalnie dokładnego dyskretnego gradientu i jego symetrycznej modyfikacji na przykładzie średniego okresu drgań.



Wykres 6.3. Średni okres drgań modelu w funkcji ε. Romby – dyskretny gradient; kwadraty – lokalnie dokładny dyskretny gradient; trójkąty – jego symetryczna modyfikacja; $x_0 = 3.0$, $y_0 = 1.5$.



Wykres 6.4. Średni okres drgań modelu w funkcji ε. Kwadraty – lokalnie dokładny dyskretny gradient; trójkąty – jego symetryczna modyfikacja; $x_0 = 3.0$, $y_0 = 1.5$.



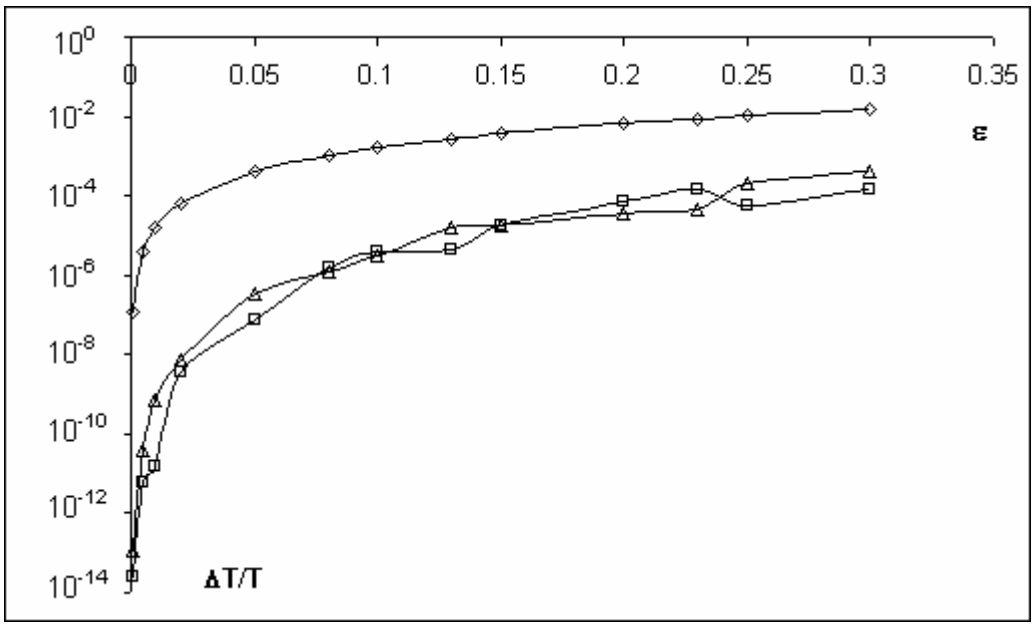
Wykres 6.5. Średni okres drgań modelu w funkcji ε. Kwadraty – lokalnie dokładny dyskretny gradient; trójkąty – jego symetryczna modyfikacja; $x_0 = 3.0$, $y_0 = 1.5$.

Wykresy 6.3 do 6.5 pokazują dobre asymptotyczne zachowanie badanych dyskretyzacji i jednocześnie ogromną przewagę w dokładności odtwarzania przebiegu czasowego wykazywaną przez zmodyfikowane wersje dyskretnego gradientu. Standardowy dyskretny gradient zbiega wraz ze zmniejszaniem ε do tej samej wartości okresu drgań jednak znacznie wolniej (tabela 6.1).

Tabela 6.1. Średni okres drgań badanych dyskretyzacji w granicy małych ε, $x_0 = 3.0$, $y_0 = 1.5$.

ε	grad	gzd2	gzd2s
0.00001	4.99992010492491	4.99992010483907	4.99992010483907
0.0001	4.99992011327211	4.99992010493938	4.99992010493938
0.0005	4.99992031328001	4.99992010495026	4.99992010495014
0.001	4.99992093826957	4.99992010495077	4.99992010495111
0.005	4.99994093770013	4.99992010488738	4.99992010512322
0.01	5.00000343265346	4.99992010379249	4.99992010759092

Na podstawie danych z tabeli 6.1, możemy przyjąć, że znamy okres drgań układu z dokładnością na poziomie 10^{-13}. Posługując się tą wartością w zastępstwie okresu faktycznego, który jest nieznany, wyznaczono względne odchylenia od okresu „teoretycznego" w powyższym sensie. Przykładowe wyniki znajdziemy na wykresie 6.6.



Wykres 6.6. Względne odchylenie okresu badanych dyskretyzacji od „teorii" jako funkcja ε. Romby – dyskretny gradient; kwadraty – lokalnie dokładny dyskretny gradient; trójkąty – jego symetryczna modyfikacja; $x_0 = 3.0$, $y_0 = 2.0$.

Lokalnie dokładna modyfikacja metody dyskretnego gradientu przewyższa ją o 6-7 rzędów wielkości przy małych wartościach ε i o 2 rzędy wielkości przy dużych ε. Sytuacja jest podobna dla innych warunków początkowych.

6.6 Podsumowanie

W rozdziale tym zajęliśmy się 2-wymiarowym modelem Lotki-Volterry, na którym przetestowano 3 dyskretyzacje: standardową metodę dyskretnego gradientu oraz dwie nowe metody wprowadzone w rozdziale 5 – schemat lokalnie dokładnego dyskretnego gradientu i jego symetryczną modyfikację. Wszystkie przedstawione integratory zachowują trajektorie równań Lotki-Volterry w przestrzeni fazowej, a obie metody lokalnie dokładne, GR-LEX i GR-SLEX, pozwalają na bardzo dokładne obliczanie ewolucji czasowej.

Stwierdzono, że dla celów eksperymentalnych korzystna jest zamiana zmiennych określona wzorami (6.14), która zdecydowanie poprawia stabilność rozwiązań numerycznych.

Testy pozwoliły na bardzo dokładne wykreślenie zależności czasowych oraz typowych trajektorii układu w przestrzeni fazowej. Okazało się, że schematy lokalnie dokładne pozwalają bez trudu wyznaczać okres oscylacji z dokładnością 10^{-13}.

7 Algorytmy dyskretnego gradientu wyższych rzędów dla jednowymiarowych układów hamiltonowskich

W rozdziale tym zostaną wprowadzone metody dyskretnego gradientu wyższych rzędów. Schematy dyskretnego gradientu są użyteczne w całkowaniu numerycznym dynamicznych układów wielu ciał [35], [41], [49], [54], [98]. Zachowują dokładnie zarówno całkowitą energię jak i moment pędu. Ostatnio metody dyskretnego gradientu zostały rozwinięte w kontekście numerycznego całkowania geometrycznego [64]. Quispel ze współpracownikami skonstruowali integratory zachowujące wszystkie całki ruchu dowolnego układu równań różniczkowych zwyczajnych [65], [66], [85], [87]. Podobne idee zostały wykorzystane w dziedzinie dynamiki molekularnej i cieczy spinowych.

Ogólnie geometryczne integratory numeryczne bardzo dobrze zachowują jakościowe cechy symulowanych równań różniczkowych, lecz nie jest łatwo zwiększyć ich dokładność nie tracąc własności geometrycznych. Algorytmy symplektyczne mogą być poprawiane poprzez odpowiednie metody podziału [9], [30], [63], [78], [79], [106], [113], okazuje się, że metodę dyskretnego gradientu (niesymplektyczną) również można poprawić nie tracąc jej cennych własności jakościowych (rozdział 5).

W tym miejscu zajmiemy się dalszym istotnym ulepszeniem omawianej metody poprzez skonstruowanie schematów dyskretnego gradientu dowolnego rzędu N dla jednowymiarowego układu Hamiltonowskiego w postaci

$$\dot{p} = -V'(x), \quad \dot{x} = p\,, \tag{7.1}$$

gdzie $V(x)$ jest potencjałem, a kropka i prim oznaczają odpowiednio różniczkowanie po t oraz x. W takim przypadku metoda dyskretnego gradientu redukuje się do tak zwanej zmodyfikowanej metody punktu środkowego:

$$\begin{aligned} \frac{x_{n+1} - x_n}{h} &= \frac{1}{2}(p_{n+1} + p_n), \\ \frac{p_{n+1} - p_n}{h} &= -\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n}, \end{aligned} \tag{7.2}$$

przy czym h oznacza krok czasowy.

Zmodyfikowana metoda punktu środkowego została w sposób naturalny rozszerzona do przypadku trzech wymiarów i układów cząstek, dokładnie zachowując całkowitą energię, pęd i moment pędu układu [54].

7.1 Schematy dyskretnego gradientu *N*-tego rzędu

Rozpatrujemy następującą rodzinę niestandardowych schematów numerycznych parametryzowanych przez funkcję δ:

$$\begin{aligned} \frac{x_{n+1}-x_n}{\delta} &= \frac{1}{2}(p_{n+1}+p_n), \\ \frac{p_{n+1}-p_n}{\delta} &= -\frac{V(x_{n+1})-V(x_n)}{x_{n+1}-x_n}, \end{aligned} \tag{7.3}$$

przy czym δ może zależeć od dowolnych zmiennych i parametrów włączając w to h, x_n, p_n, x_{n+1}, p_{n+1}. Można łatwo sprawdzić, że każdy schemat z tej rodziny zachowuje energię całkowitą:

$$\frac{1}{2}p_n^2 + V(x_n) = E = const\,. \tag{7.4}$$

W rozdziale 5 rozpatrywana była δ w postaci

$$\delta = \frac{2}{\omega}\tan\frac{h\omega}{2}, \quad \omega = \sqrt{V''(\bar{x})}\,, \tag{7.5}$$

gdzie $\bar{x}$ mogło zależeć od x_n, x_{n+1}, ale zwykle nie zależało od h. Biorąc $\bar{x} = x_0$ (przy $V'(x_0) = 0$) dostajemy zmodyfikowaną metodę dyskretnego gradientu (MOD-GR). Kładąc $\bar{x} = x_n$ i $\bar{x} = \frac{1}{2}(x_n + x_{n+1})$ dostajemy schemat lokalnie dokładnego lokalnego gradientu (GR-LEX) i jego symetryczną modyfikację (GR-SLEX) (rozdział 5). Te trzy schematy numeryczne są odpowiednio rzędu drugiego, trzeciego i czwartego.

Pokażemy, że rodzina integratorów numerycznych (7.3) zawiera w sobie schematy numeryczne dowolnego rzędu. Ich jawne postacie zostaną podane do rzędu 11 włącznie.

Układ równań (7.3) (gdzie $x_n \equiv x$, $p_n \equiv p$ są dane, a $\delta_n = \delta$ jest małym parametrem) w sposób pośredni definiuje x_{n+1} i p_{n+1}. Dlatego też stosując różniczkowanie niejawne, możemy zapisać odpowiednie rozwinięcia w szereg Taylora:

$$\begin{aligned} x_{n+1} &= x + p\delta - \frac{1}{2}V'\delta^2 - \frac{1}{4}pV''\delta^3 + \frac{1}{24}(3V'V'' - 2p^2V''')\delta^4 + O(\delta^5), \\ p_{n+1} &= p - V'\delta - \frac{1}{2}pV''\delta^2 + \frac{1}{12}(3V'V'' - 2p^2V''')\delta^3 \\ &\quad - \frac{1}{24}(4pV'V''' + 3p(V'')^2 - p^3V^{(4)})\delta^4 + O(\delta^5). \end{aligned} \tag{7.6}$$

Załóżymy teraz, że x_{n+1} i p_{n+1} są zgodne z rozwiązaniem dokładnym aż do wyrazów rzędu N, to znaczy, że ich rozwinięcia w szereg Taylora mają co

najmniej N pierwszych wyrazów identycznych jak w szeregu Taylora (7.21). Obliczymy pierwszych N składników rozwinięcia Taylora funkcji δ używając pierwszego z równań (7.3) tj.

$$\delta = \frac{2(x_{n+1} - x_n)}{(p_{n+1} + p_n)}. \tag{7.7}$$

Wielomian wynikowy N-tego stopnia oznaczymy przez δ_N, a jego współczynniki przez a_k:

$$\delta_N = \delta_N(x, p, h) = \sum_{k=1}^{N} a_k(x, p)h^k = h + \sum_{k=3}^{N} a_k(x, p)h^k, \tag{7.8}$$

gdzie $a_1 = h$, $a_2 = 0$ oraz a_k (dla $k \geq 3$) są wielomianami ze względu na p ze współczynnikami zależnymi od x poprzez pochodne V. Oznaczając przez a z indeksem dolnym różniczkowanie względem x (i używając skrótów typu $V_{4x} \equiv V_{xxxx}$) przedstawimy pewną liczbę współczynników a_k w postaci jawnej:

$$a_3 = \frac{1}{12}V_{xx}, \quad a_4 = \frac{1}{24}pV_{xxx}, \tag{7.9}$$

$$a_5 = \frac{1}{240}(2V_{xx}^2 - 4V_xV_{xxx} + 3p^2V_{4x}), \tag{7.10}$$

$$a_6 = \frac{1}{1440}\left((5V_{xx}V_{xxx} - 15V_xV_{4x})p + 4V_{5x}p^3\right), \tag{7.11}$$

$$a_7 = \frac{1}{20160}(a_{70} + a_{72}p^2 + a_{74}p^4), \tag{7.12}$$

$$\begin{aligned}
a_8 &= \frac{1}{40320}(a_{81}p + a_{83}p^3 + a_{85}p^5),\\
a_9 &= \frac{1}{725760}(a_{90} + a_{92}p^2 + a_{94}p^4 + a_{96}p^6),\\
a_{10} &= \frac{1}{7257600}(a_{101}p + a_{103}p^3 + a_{105}p^5 + a_{107}p^7),\\
a_{11} &= \frac{1}{159667200}(a_{110} + a_{112}p^2 + a_{114}p^4 + a_{116}p^6 + a_{118}p^8),
\end{aligned} \tag{7.13}$$

przy czym współczynniki a_{jk}, a_{jkm} zależą od x poprzez pochodne V następująco:

$$\begin{aligned}
a_{70} &= 17V_{xx}^3 + 45V_x^2V_{4x} - 44V_xV_{xx}V_{xxx},\\
a_{72} &= 20V_{xxx}^2 - 12V_xV_{4x} - 72V_xV_{5x},\\
a_{74} &= 10V_{6x},
\end{aligned} \tag{7.14}$$

$$\begin{aligned}
a_{81} &= 21V_{xx}^2V_{xxx} - 42V_xV_{xxx}^2 + 63V_x^2V_{5x},\\
a_{83} &= 14V_{xxx}V_{4x} - 21V_{xx}V_{5x} - 35V_xV_{6x},\\
a_{85} &= 3V_{7x},
\end{aligned} \tag{7.15}$$

$$
\begin{aligned}
a_{90} &= 62V_{xx}^4 - 228V_xV_{xx}^2V_{xxx} + 168(V_x^2V_{xxx}^2 - V_x^3V_{5x}) + 90V_x^2V_{xx}V_{4x},\\
a_{92} &= 75V_{xx}V_{xxx}^2 + 81V_{xx}^2V_{4x} - 462V_xV_{xxx}V_{4x} + 360V_xV_{xx}V_{5x} + 420V_x^2V_{6x},\\
a_{94} &= 42V_{4x}^2 - 120(V_{xx}V_{6x} + V_xV_{7x}),\\
a_{96} &= 7V_{8x},
\end{aligned}
\tag{7.16}
$$

$$
\begin{aligned}
a_{101} &= 460V_{xx}^3V_{xxx} - 1170V_xV_{xx}V_{xxx}^2 - 630V_xV_{xx}^2V_{4x} + 2385V_x^2V_{xxx}V_{4x}\\
&\quad - 945V_x^2V_{xx}V_{5x} - 1260V_x^3V_{6x},\\
a_{103} &= 150V_{xxx}^3 + 15V_{xx}V_{xxx}V_{4x} - 945V_xV_{4x}^2 - 456V_xV_{xxx}V_{5x} + 483V_{xx}^2V_{5x}\\
&\quad + 1785V_xV_{xx}V_{6x} + 1080V_x^2V_{7x},\\
a_{105} &= 126V_{4x}V_{5x} - 114V_{3x}V_{6x} - 261V_{xx}V_{7x} - 189V_xV_{8x},\\
a_{107} &= 8V_{9x},
\end{aligned}
\tag{7.17}
$$

$$
\begin{aligned}
a_{110} &= 1382V_{xx}^5 - 6448V_xV_{xx}^3V_{3x} + 4140V_x^2V_{xx}^2V_{4x} + 840V_x^3V_{xx}V_{5x}\\
&\quad + 8280V_x^3V_{3x}V_{4x} + 7368V_x^2V_{xx}V_{3x}^2 + 3150V_x^4V_{6x},\\
a_{112} &= 3240V_{xx}^2V_{3x}^2 - 6480V_xV_{3x}^3 + 696V_{xx}^3V_{4x} + 4872V_xV_{xx}V_{3x}V_{4x}\\
&\quad + 15660V_x^2V_{4x}^2 - 9144V_xV_{xx}^2V_{5x} + 11988V_x^2V_{3x}V_{5x}\\
&\quad - 21000V_x^2V_{xx}V_{6x} - 10800V_x^3V_{7x},\\
a_{114} &= 1710V_{3x}^2V_{4x} - 1803V_{xx}V_{4x}^2 - 8676V_xV_{4x}V_{5x} + 1260V_{xx}V_{3x}V_{5x}\\
&\quad + 4770V_{xx}^2V_{6x} + 3060V_xV_{3x}V_{6x} + 11400V_xV_{xx}V_{7x} + 4725V_x^2V_{8x},\\
a_{116} &= 336V_{5x}^2 - 780V_{3x}V_{7x} - 980V_{xx}V_{8x} - 560V_xV_{9x} + 120V_{4x}V_{6x},\\
a_{118} &= 18V_{10x}.
\end{aligned}
\tag{7.18}
$$

Schemat numeryczny (7.3), gdzie $\delta = \delta_N$ jest zdefiniowana przez (7.8), będzie oznaczany przez GR-*N*. Metoda GR-*N* jest (co najmniej) rzędu *N*. Warto zauważyć, że metody GR-1 i GR-2 są zgodne ze standardowym dyskretnym gradientem (GR) danym przez (7.2) (w szczególności GR-1 jest rzędu 2). Ponadto, jeśli potencjał *V* jest liniowy względem *x*, wówczas dowolna metoda GR-*N* jest dokładna (jej rząd staje się nieskończony).

7.2 Standardowe metody *N*-tego rzędu

Wyprowadzimy teraz jawne schematy numeryczne dowolnego rzędu posługując się rozwinięciem w szereg Taylora $x(t + h)$ i $p(t + h)$:

$$
x(t+h) = \sum_{k=0}^{N} \frac{h^k}{k!} \frac{d^k x(t)}{dt^k}, \qquad p(t+h) = \sum_{k=0}^{N} \frac{h^k}{k!} \frac{d^k p(t)}{dt^k},
\tag{7.19}
$$

gdzie wszystkie pochodne mogą być zastąpione funkcjami x, p poprzez zastosowanie równania (7.1) i jego pochodnych (np. $\ddot{p} = -V''(x)\dot{x} = -V''(x)p$). Dostajemy wówczas

$$\begin{aligned} x(t+h) &= x + ph - \frac{1}{2}V'h^2 - \frac{1}{6}pV''h^3 + \frac{1}{24}(V'V'' - V'''p^2)h^4 + O(h^5), \\ p(t+h) &= p - V'h - \frac{1}{2}pV''h^2 + \frac{1}{6}(V'V'' - V'''p^2)h^3 \\ &\quad + \frac{1}{24}(3pV'V''' + p(V'')^2 - p^3V^{(4)})h^4 + O(h^5). \end{aligned} \tag{7.20}$$

Dlatego też rozwinięcie Taylora może być przedstawione w formie

$$x(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} b_k(x,p), \qquad p(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} c_k(x,p), \tag{7.21}$$

gdzie $b_k = \dfrac{d^k}{dt^k}x$, $c_k = \dfrac{d^k}{dt^k}p$, a pochodne te obliczamy stosując (7.1). Na przykład $b_0 = x$, $b_1 = \dot{x} = p$ i $b_2 = \ddot{x} = \dot{p} = -V'(x)$. Ogólnie

$$b_{k+1} = \frac{d}{dt}b_k = \frac{\partial b_k}{\partial x}\dot{x} + \frac{\partial b_k}{\partial p}\dot{p} = p\frac{\partial b_k}{\partial x} - V'(x)\frac{\partial b_k}{\partial p}. \tag{7.22}$$

Z kolei $p = \dot{x}$ implikuje

$$c_k = \frac{d}{dt}b_k = b_{k+1}. \tag{7.23}$$

Współczynniki b_k (k = 1, 2, …, 11), obliczone rekurencyjnie z (7.22) mają postać

$$\begin{aligned} &b_0 = x, \quad b_1 = p, \quad b_2 = -V_x, \quad b_3 = -pV_{xx}, \\ &b_4 = V_xV_{xx} - p^2V_{xxx}, \\ &b_5 = p(V_{xx}^2 + 3V_xV_{xxx}) - p^3V_{4x}, \\ &b_6 = -3V_x^2V_{xxx} - V_xV_{xx}^2 + p^2(5V_{xx}V_{xxx} + 6V_xV_{4x}) - p^4V_{5x}, \\ &b_7 = -p(V_{xx}^2 + 18V_xV_{xx}V_{xxx} + 15V_x^2V_{4x}) \\ &\quad + p^3(5V_{xxx}^2 + 11V_{xx}V_{4x} + 10V_xV_{5x}) - p^5V_{6x}, \end{aligned} \tag{7.24}$$

$$\begin{aligned} b_8 &= V_xV_{xx}^3 + 18V_x^2V_{xx}V_{xxx} + 15V_x^3V_{4x} \\ &\quad - p^2(21V_{xx}^2V_{xxx} + 33V_xV_{xxx}^2 + 81V_xV_{xx}V_{4xx} + 45V_x^2V_{5x}) \\ &\quad + p^4(21V_{3x}V_{4x} + 21V_{xx}V_{5x} + 15V_xV_{6x}) - p^6V_{7x}, \end{aligned} \tag{7.25}$$

$$\begin{aligned} b_9 &= p(V_{xx}^4 + 81V_xV_{xx}^2V_{3x} + 84V_x^2V_{3x}^2 + 225V_x^2V_{xx}V_{4x} + 105V_x^3V_{5x}) \\ &\quad - p^3(75V_{xx}V_{3x}^2 + 102V_{xx}^2V_{4x} + 231V_xV_{3x}V_{4x} + 225V_xV_{xx}V_{5x} \\ &\quad + 105V_x^2V_{6x}) + p^5(21V_{4x}^2 + 42V_{3x}V_{5x} + 36V_{xx}V_{6x} + 21V_xV_{7x}) \\ &\quad - p^7V_{8x}, \end{aligned} \tag{7.26}$$

$$
\begin{aligned}
b_{10} = &-(V_x V_{xx}^4 + 81 V_x^2 V_{xx}^2 V_{3x} + 84 V_x^3 V_{3x}^2 + 225 V_x^3 V_{xx} V_{4x} \\
&+ 105 V_x^4 V_{5x}) + p^2 (85 V_{xx}^3 V_{3x} + 555 V_x V_{xx} V_{3x}^2 + 837 V_x V_{xx}^2 V_{4x} \\
&+ 1086 V_x^2 V_{3x} V_{4x} + 1305 V_x^2 V_{xx} V_{5x} + 420 V_x^3 V_{6x}) \\
&- p^4 (75 V_{3x}^3 + 585 V_{xx} V_{3x} V_{4x} + 336 V_x V_{4x}^2 + 357 V_{xx}^2 V_{5x} \\
&+ 696 V_x V_{3x} V_{5x} + 645 V_x V_{xx} V_{6x} + 210 V_x^2 V_{7x}) + p^6 (84 V_{4x} V_{5x} \\
&+ 78 V_{3x} V_{6x} + 57 V_{xx} V_{7x} + 28 V_x V_{8x}) - p^8 V_{9x},
\end{aligned}
\tag{7.27}
$$

$$
\begin{aligned}
b_{11} = &-p(V_{xx}^5 + 336 V_x V_{xx}^3 V_{3x} + 1524 V_x^2 V_{xx} V_{3x}^2 + 2430 V_x^2 V_{xx}^2 V_{4x} \\
&+ 2565 V_x^3 V_{3x} V_{4x} + 3255 V_x^3 V_{xx} V_{5x} + 945 V_x^4 V_{6x}) \\
&+ p^3 (810 V_{xx}^2 V_{3x}^2 + 855 V_x V_{3x}^3 + 922 V_{xx}^3 V_{4x} + 2430 V_x^2 V_{4x}^2 \\
&+ 4875 V_x V_{xx}^2 V_{5x} + 5175 V_x^2 V_{3x} V_{5x} + 5145 V_x^2 V_{xx} V_{6x} \\
&+ 7296 V_x V_{xx} V_{3x} V_{4x} + 1260 V_x^3 V_{7x}) - p^5 (810 V_{3x}^2 V_{4x} \\
&+ 921 V_{xx} V_{4x}^2 + 1995 V_{xx} V_{3x} V_{5x} + 1872 V_x V_{4x} V_{5x} + 1002 V_{xx}^2 V_{6x} \\
&+ 1809 V_x V_{3x} V_{6x} + 1407 V_x V_{xx} V_{7x} + 378 V_x^2 V_{8x}) + p^7 (84 V_{5x}^2 \\
&+ 162 V_{4x} V_{6x} + 135 V_{3x} V_{7x} + 85 V_{xx} V_{8x} + 36 V_x V_{9x}) - p^9 V_{10x}.
\end{aligned}
\tag{7.28}
$$

Zatem dla dowolnego N otrzymaliśmy następujący, zadany w sposób jawny, schemat numeryczny oznaczany przez TAY-N (schemat Taylora N-tego rzędu):

$$
x_{n+1} = \sum_{k=0}^{N} \frac{h^k}{k!} b_k(x_n, p_n), \qquad p_{n+1} = \sum_{k=0}^{N} \frac{h^k}{k!} c_k(x_n, p_n), \tag{7.29}
$$

gdzie b_k i c_k są zdefiniowane przez (7.22), (7.23) oraz, w szczególnych przypadkach, przez (7.24), (7.25), (7.26), (7.27) i (7.28). Jawne integratory TAY-N zostaną użyte do porównań z metodami dyskretnego gradientu wyższych rzędów. Co więcej są one dobrymi kandydatami na predyktory, gdy metody gradientowe (7.3) są używane w charakterze korektorów.

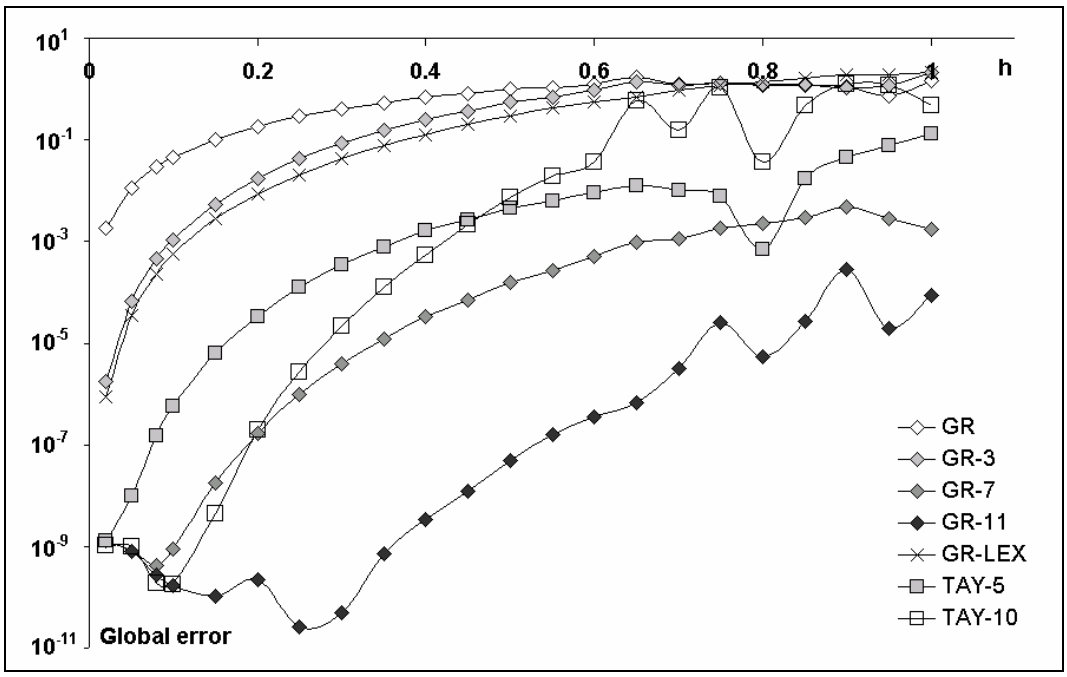
7.3 Eksperymenty numeryczne

W poprzednich rozdziałach tej pracy porównaliśmy szereg dyskretyzacji na przykładzie potencjału wahadła matematycznego ($V(x) = -k\cos(x)$) oraz potencjału Morse'a ($V(x) = \frac{1}{2} k e^{-2\alpha x} - k e^{-\alpha x}$). Najlepsze okazały się schematy lokalnie dokładnego dyskretnego gradientu (GR-LEX i GR-SLEX). W niektórych testach ich dokładność przewyższała o wiele rzędów wielkości osiąganą przez schematy standardowe, takie jak leap-frog, metodę niejawnego punktu środkowego (implicit midpoint) czy też zwykły dyskretny gradient (GR).

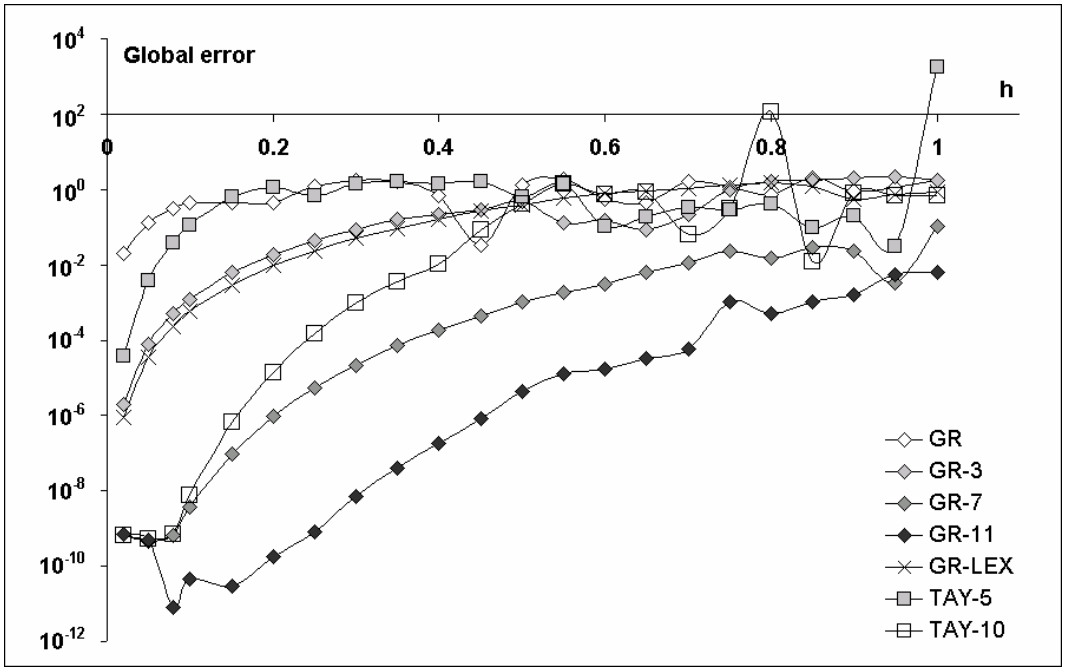
Porównamy teraz schemat GR-LEX (GR-SLEX daje bardzo zbliżone rezultaty) z algorytmami wyższych rzędów wprowadzonych w tym rozdziale, tj. GR-*N* oraz TAY-*N* wykorzystując te same potencjały (w obu przypadkach znane są rozwiązania dokładne). Dla uproszczenia zawsze zakładamy $x_0 = 0$ (położenie początkowe w punkcie równowagi trwałej). Szczegóły techniczne obliczania okresu zostały wyjaśnione w rozdziale 4.8, a szczegóły procedur iteracyjnych w rozdziale 11.3. Warto zaznaczyć, że δ_N zadana przez (7.8) zależy od x_n, p_n, lecz nie zależy od x_{n+1}, p_{n+1}, co oznacza, że parametr ten jest obliczany tylko raz w każdym kroku.

7.3.1 Błąd globalny

Wykresy 7.1 i 7.2 pokazują zależność błędu globalnego rozwiązania numerycznego w zależności od kroku czasowego (liczonego po upływie $120T_{th}$). Schemat GR-3 daje praktycznie takie same wyniki jak GR-LEX. Są one lepsze od standardowej metody GR o kilka rzędów wielkości, zwłaszcza dla mniejszych *h*. Widzimy też, że GR-*N* (dla $N \geq 5$) są dokładniejsze od GR-LEX o kilka rzędów wielkości i zachowują precyzję działania również w przypadku dużych wartości kroku czasowego. Dodatkowo dla małych kroków czasowych ($h \leq 0.1$) dokładność GR-7 i GR-11 praktycznie nie zależy od ich wartości (a nawet nieco maleje).

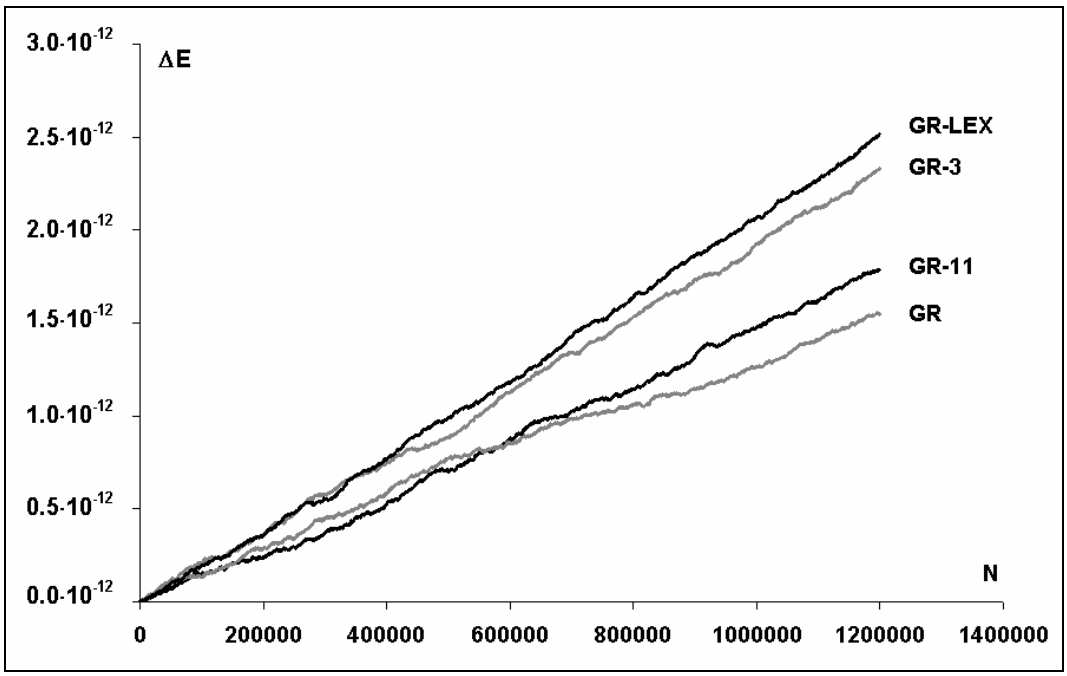


Wykres 7.1. Błąd globalny po upływie $t = 120T_{th}$ jako funkcja kroku czasowego *h* dla wahadła matematycznego, $p_0 = 1.8$ ($T_{th} = 9.12219655$).



Wykres 7.2. Błąd globalny po upływie $t = 120T_{th}$ jako funkcja kroku czasowego h dla potencjału Morse'a, $p_0 = 0.8$ ($T_{th} = 10.471\,975\,51$).

Schemat TAY-10 dorównuje dokładnością algorytmom GR-7 i GR-11 przy małych wartościach h. Przedział jego konkurencyjności rośnie wraz z przesuwaniem się do obszaru drgań bliskich harmonicznym (zmniejszaniem p_0).

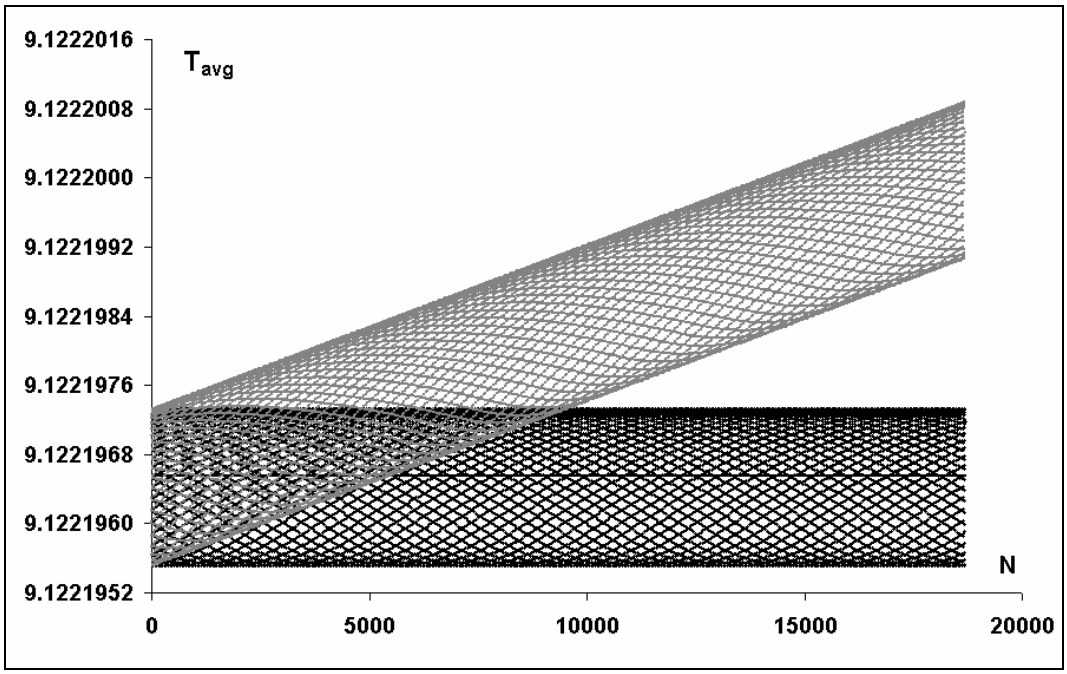


Wykres 7.3. Błąd energii jako funkcja czasu ($t = Nh$, $h = 0.25$) dla wahadła matematycznego, $p_0 = 1.8$ ($E_{ex} = 0.62$).

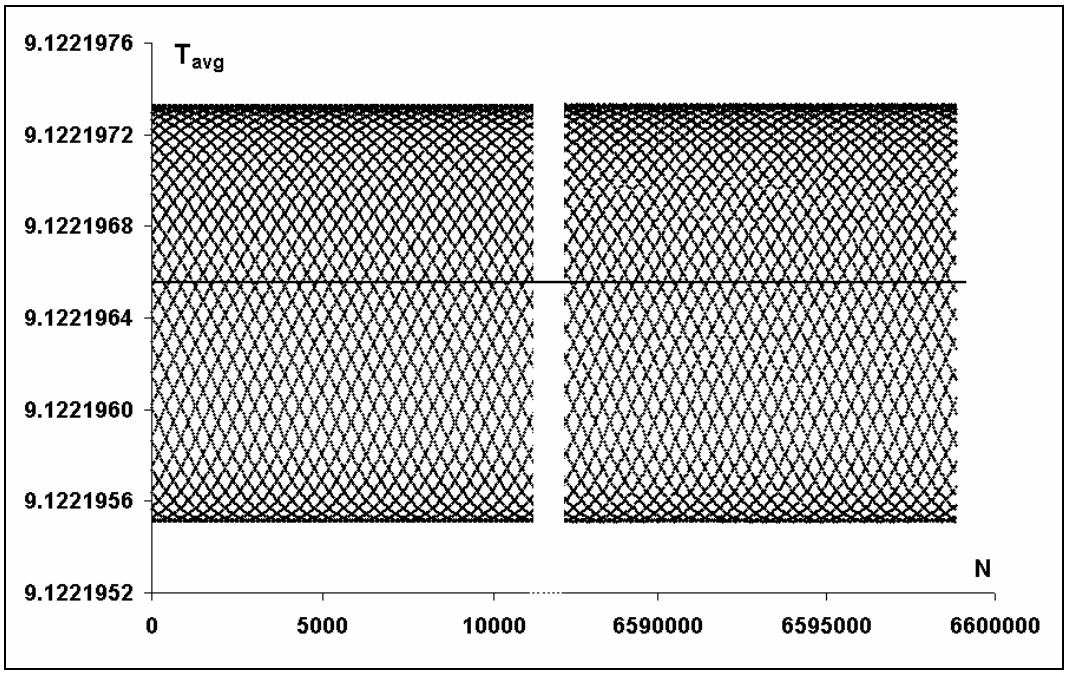
Teoretycznie wszystkie schematy gradientowe (7.3) dokładnie zachowują energię, jednak błędy numeryczne wprowadzają w sposób nieunikniony niewielkie, ale narastające odchylenie od wartości teoretycznej (wykres 7.3). Przyrost ten jest z grubsza liniowy i osiąga $\Delta E \approx 10^{-12}$ dla $t \approx 200000$.

7.3.2 Stabilność oscylacji i błąd względny okresu

Wszystkie schematy gradientowe mają stały okres drgań w bardzo długim okresie czasu. Stabilność okresu standardowej metody dyskretnego gradientu była testowana szczegółowo w rozdziale 4.8. Wyniki tam przedstawione odnoszą się również do pozostałych metod należących do tej rodziny. Na wykresie 7.4 porównano średnie okresy (ściśle $T_{avg}(N, 20)$, porównaj rozdział 4.8) rozwiązań numerycznych otrzymanych przy pomocy GR-7 i TAY-10. Dla czasów, które nie są bardzo duże, w obydwu przypadkach średni okres oscyluje wokół wartości teoretycznej T_{th}. Jednak dokładność schematu TAY-10 powoli, ale systematycznie maleje, podczas gdy schemat gradientowy oscyluje w ten sam sposób przez bardzo długi czas (miliony okresów – porównaj wykres 7.5).

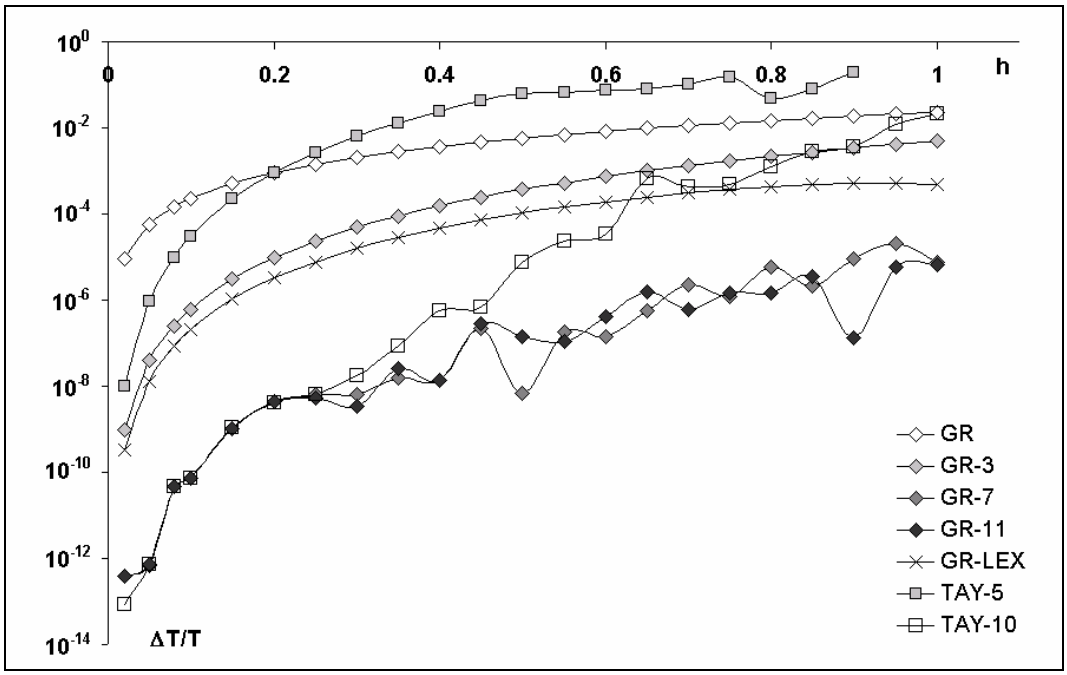


Wykres 7.4. Średni okres jako funkcja czasu (*N* jest liczbą półokresów) dla wahadła matematycznego $p_0 = 1.8$ ($T_{th} = 9.12219655$). Czarne punkty – GR-7, jasne punkty – TAY-10, pozioma linia – okres teoretyczny.



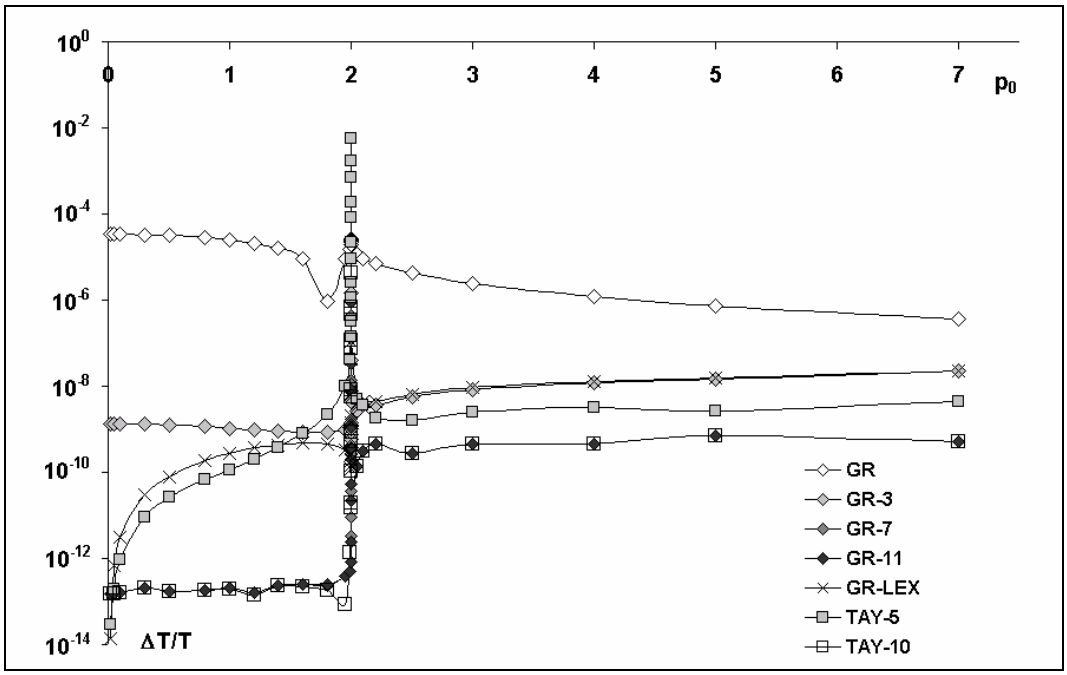
Wykres 7.5. Średni okres jako funkcja czasu (*N* jest liczbą półokresów) dla wahadła matematycznego, p_0 = 1.8, h = 0.25 (T_{th} = 9.12219655), schemat GR-11. Linia pozioma – okres teoretyczny.

Wykresy 7.6 i 7.7 przedstawiają względny błąd okresu (dokładniej wielkości $\overline{T}_{avg}(0, 100, 200)$, podobnie jak w rozdziałach 4.8 i 5.5). W zarysie chodzi o pewien rodzaj uśredniania dotyczący pierwszych 200 okresów i porównanie wyniku z okresem dokładnym T_{th}.



Wykres 7.6. Względy błąd okresu wahadła matematycznego jako funkcja *h* dla p_0 = 1.95 (T_{th} = 11.657 585 28).

Wykres 7.6 prezentuje zależność względnego błędu okresu od kroku czasowego h. Widać, że GR-7 daje świetne wyniki przewyższając swą dokładnością o 3-4 rzędy wielkości metodę GR-LEX. Dokładności GR-11 i TAY-10 są (dla $p_0 = 1.95$ i $h < 0.3$) zasadniczo takie same, jednak precyzja działania drugiej z nich spada wyraźnie wraz ze wzrostem h.

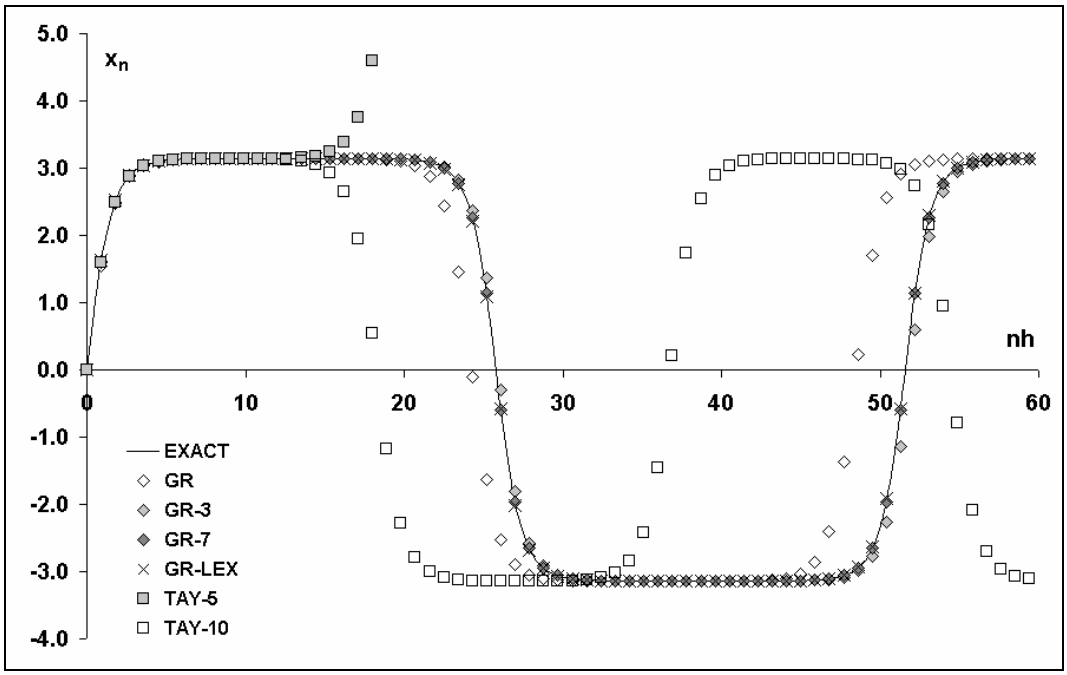


Wykres 7.7. Względy błąd okresu wahadła matematycznego jako funkcja p_0 dla $h = 0.02$ ($T_{th} = 6.283\,342\,4$).

Zwiększanie rzędu GR-N przy małych wartościach h (wykres 7.7) zwiększa dokładność tylko do pewnego momentu. Widzimy, że GR-7 daje praktycznie takie same wyniki jak GR-11 i TAY-10 tj. 10^{-13} dla oscylacji i 10^{-10} w przypadku ruchu rotacyjnego. Wyjątkiem jest wąski obszar $p_0 \approx 2$, gdzie dokładność jest mniejsza dla każdego schematu numerycznego. Przy $p_0 < 2$ (oscylacje) GR-7 przewyższa dokładnością GR o 7-9 rzędów wielkości. GR-LEX i TAY-5 również osiągają porównywalną dokładność, lecz tylko dla małych wartości p_0. Przy $p_0 > 2$ schemat GR-LEX daje takie same wyniki jak GR-3, które są w przybliżeniu o rząd wielkości gorsze od uzyskiwanych przez TAY-5 i o dwa rzędy od GR-7, GR-11 i TAY-10, które na wykresie 7.7 w wielu punktach się pokrywają. Należy pamiętać, że wraz z upływem czasu błędy generowane przez schematy TAY-N będą systematycznie rosły, co nie dotyczy w praktyce (wykresy 7.4, 7.5) schematów gradientowych.

7.3.3 Sąsiedztwo separatrysy

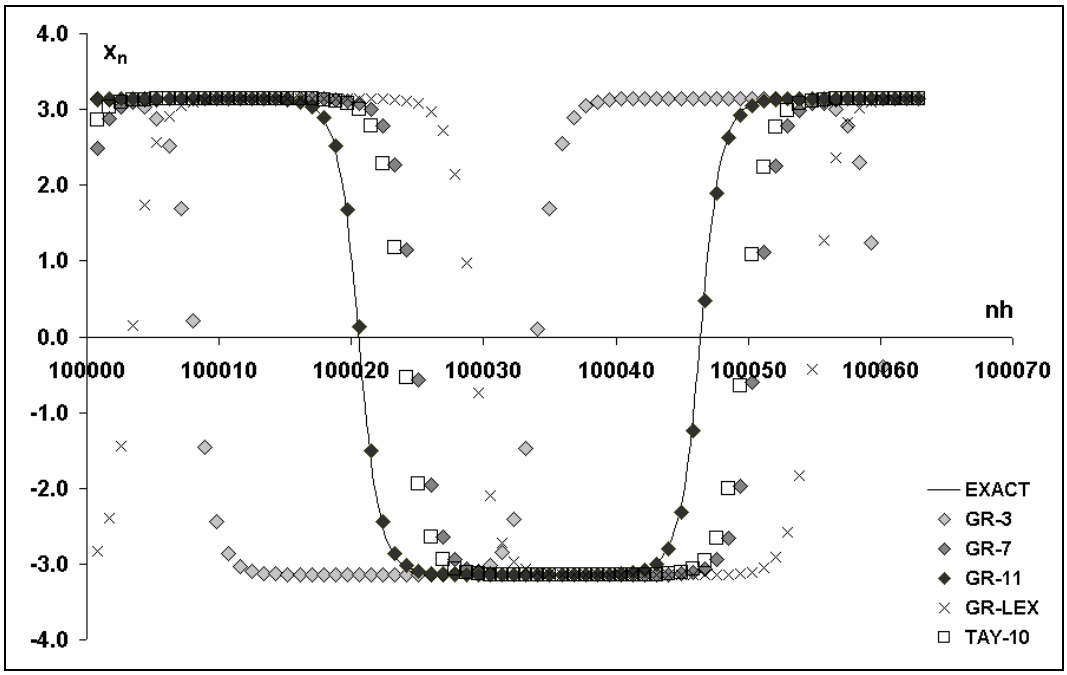
Najtrudniejsze do numerycznej symulacji jest sąsiedztwo separatrysy ($p_0 \approx 2$ dla wahadła matematycznego). Standardowa metoda dyskretnego gradientu okazuje się stosunkowo dobra w tym regionie (rozdział 4.11), a schematy lokalnie dokładne (rozdział 5.5) pracują niemal perfekcyjnie. W tym miejscu wybrano do porównań również GR-3, GR-7, GR-11 oraz TAY-5 i TAY-10. Schematy taylorowskie okazały się najgorsze: TAY-10 nie jest w stanie odtworzyć nawet zachowania jakościowego (wykres 7.8). Zwykła dyskretyzacja gradientowa (GR) daje dobre zachowanie jakościowe i jest zdecydowanie dokładniejsza od TAY-10 pracującego z krokiem o połowę mniejszym. Widzimy, że w pierwszym okresie GR-3, GR-7 i GR-LEX dają podobne wyniki. Podkreślmy, że rozwiązanie dokładne jest bardzo bliskie separatrysy ($|p_0 - 2| = 10^{-10}$) i krok czasowy jest bardzo duży, mimo to wszystkie ulepszone schematy gradientowe symulują bardzo dokładnie ruch wahadła.



Wykres 7.8. x_n jako funkcja czasu ($t = nh$) bardzo blisko separatrysy (p_0 = 1.999 999 999 9), h = 0.09 dla TAY-5, h = 0.45 dla TAY-10, h = 0.9 dla pozostałych dyskretyzacji. Linia ciągła odpowiada rozwiązaniu dokładnemu (T_{th} = 51.596 879 14).

Wykres 7.9 pokazuje tę samą sytuację po upływie znacznie dłuższego czasu (t > 100 000). Zwróćmy uwagę, że krok czasowy dla TAY-10 (h = 0.09) jest dużo mniejszy niż krok, z jakim pracują pozostałe dyskretyzacje (h = 0.9).

Mimo to TAY-10 jest tylko minimalnie lepsza od GR-7 i zdecydowanie gorsza od GR-11. GR-7 okazuje się dokładniejsza od GR-LEX.



Wykres 7.9. x_n jako funkcja czasu ($t = nh$) bardzo blisko separatrysy (p_0 = 1.999 999 999 9), h = 0.09 dla TAY-10 i h = 0.9 dla pozostałych dyskretyzacji. Linia ciągła odpowiada rozwiązaniu dokładnemu (T_{th} = 51.596 879 14).

7.4 Podsumowanie

Przedstawione tu integratory numeryczne GR-*N* mają podobne zalety jak metody GR-LEX i GR-SLEX opisane w rozdziale 5: zachowują dokładnie całkę energii, są bardzo stabilne i doskonale sprawdzają się w długich okresach czasu. Można je konstruować dla dowolnego, zadanego *N*.

Posiadając zalety wcześniej wprowadzonych schematów jednocześnie istotnie poprawiają dokładność metod dyskretnego gradientu (przynajmniej w przypadku jednowymiarowym). Integratory GR-*N* ($N \geq 7$) są znacznie lepsze od GR-LEX w większości testowanych przypadków. Jedynie w obszarze małych p_0 schematy lokalnie dokładne są porównywalne z metodami gradientowymi wysokich rzędów.

Trzeba zaznaczyć, że schematy (7.3) podobnie jak wszystkie metody gradientowe nie są symplektyczne, ani też nie zachowują objętości w przestrzeni fazowej. Co więcej, integratory GR-*N* nie są odwracalne w czasie. Dlatego wydaje się, że zachowywanie energii i wysoki rząd są wystarczające do zapewnienia ich prezentowanych wyżej doskonałych cech jakościowych i ilościowych.

8 Dokładna dyskretyzacja jednowymiarowego oscylatora anharmonicznego

8.1 Ścisłe rozwiązanie oscylatora anharmonicznego

Rozważmy ruch zadany następującym równaniem Newtona

$$\ddot{x} = -\alpha x + \beta x^3, \tag{8.1}$$

gdzie α, β - stałe. Zasada zachowania energii dana jest przez

$$\frac{1}{2}p^2 + \frac{1}{2}\alpha x^2 - \frac{1}{4}\beta x^4 = E, \tag{8.2}$$

przy czym $p = \dot{x}$. W rozdziale tym skupimy się na ruchu oscylacyjnym.

8.1.1 Przypadek $\alpha > 0$ i $\beta > 0$

Przy zadanej energii układu E i warunku początkowym $x_0 = 0$, $p_0 = \sqrt{2E}$, sprawdzimy, że dokładnym rozwiązaniem równania (8.1) jest

$$x = A\,\mathrm{sn}(\omega t; k), \tag{8.3}$$

gdzie sn jest sinusem eliptycznym Jacobiego, zaś stałe A, ω, k wyrażają się w odpowiedni sposób przez α, β, E (zauważmy, że warunek początkowy $x_0 = 0$ jest spełniony automatycznie). Skorzystamy z kilku tożsamości spełnianych przez funkcje eliptyczne Jacobiego [15, 28, 95]:

$$\mathrm{sn}^2 u + \mathrm{cn}^2 u = 1, \quad \mathrm{dn}^2 u + k^2 \mathrm{sn}^2 u = 1, \quad \frac{d}{du}\mathrm{sn}\, u = \mathrm{cn}\, u\, \mathrm{dn}\, u, \tag{8.4}$$

z których bezpośrednio wynika wzór

$$\left(\frac{d}{du}\mathrm{sn}\, u\right)^2 = (1 - \mathrm{sn}^2 u)(1 - k^2 \mathrm{sn}^2 u). \tag{8.5}$$

Z postulatu $x = A\,\mathrm{sn}\, u$ (gdzie $u = \omega t$) mamy

$$\dot{x}^2 = A^2\omega^2(1 - \mathrm{sn}^2 u)(1 - k^2 \mathrm{sn}^2 u) = A^2\omega^2\left(1 - \frac{x^2}{A^2}\right)\left(1 - \frac{k^2 x^2}{A^2}\right). \tag{8.6}$$

Z drugiej strony z wyjściowego równania (8.2) wynika

$$\dot{x}^2 = \frac{1}{2}\beta x^4 - \alpha x^2 + 2E. \tag{8.7}$$

Porównanie dwóch ostatnich wzorów prowadzi do związków

$$\beta A^2 = 2k^2\omega^2, \quad \alpha = \omega^2(1 + k^2), \quad 2E = A^2\omega^2. \tag{8.8}$$

Równania te pozwalają na wyznaczenie parametrów rozwiązania A, ω, k poprzez stałe opisujące potencjał i energię:

$$k=\frac{\sqrt{\frac{4\beta E}{\alpha^2}}}{1+\sqrt{1-\frac{4\beta E}{\alpha^2}}},\qquad A=\frac{2\sqrt{\frac{E}{\alpha}}}{\sqrt{1+\sqrt{1-\frac{4\beta E}{\alpha^2}}}},\qquad \omega=\sqrt{\frac{\alpha}{2}\left(1+\sqrt{1-\frac{4\beta E}{\alpha^2}}\right)}. \tag{8.9}$$

Zatem teoretyczne rozwiązanie równania (8.1) jest kompletne, wystarczy podstawić (8.9) do (8.3). Zauważmy, że znalezione rozwiązanie jest funkcją tylko jednego parametru E (lub $p_0=\sqrt{2E}$).

Przypadek graniczny (separatrysa) odpowiada energii $E=\frac{\alpha^2}{4\beta}$ (lub $p_0=\frac{\alpha}{\sqrt{2\beta}}$). Wówczas rozwiązanie wyraża się przez funkcje elementarne:

$$x(t)=2\sqrt{\frac{\alpha}{\beta}}\tanh\left(\sqrt{\frac{\alpha}{2}}t\right). \tag{8.10}$$

8.1.2 Przypadek $\alpha>0$ i $\beta<0$

Poszukamy dokładnego rozwiązania równania (8.1) przy zadanej energii układu E i warunku początkowym $x_0 = 0$, $p_0=\sqrt{2E}$ w postaci

$$x = A\,\mathrm{cn}(\mu t; k) \tag{8.11}$$

(początek ruchu w punkcie największego wychylenia). Wprowadzając oznaczenia

$$s=\mathrm{sn}\,u,\quad c=\mathrm{cn}\,u,\quad d=\mathrm{dn}\,u,\quad k'^2=1-k^2, \tag{8.12}$$

możemy krótko zapisać związki pomiędzy funkcjami eliptycznymi

$$s^2+c^2=1,\quad d^2+k^2s^2=1 \tag{8.13}$$

oraz wynikające z nich tożsamości

$$d^2+k^2=1+k^2c^2,\quad d^2=c^2+(1-k^2)s^2=c^2+k'^2s^2. \tag{8.14}$$

Wzory na pochodne funkcji eliptycznych:

$$s'=cd,\quad c'=-sd,\quad d'=-k^2sc. \tag{8.15}$$

Podnosząc pochodne do kwadratu i wykorzystując (8.13) otrzymujemy

$$\begin{aligned} s'^2&=(1-s^2)(1-k^2s^2),\\ c'^2&=(1-c^2)(1-k^2c^2),\\ d'^2&=(1-d^2)(d^2-1+k^2). \end{aligned} \tag{8.16}$$

Korzystając z drugiego z równań (8.16) w postaci

$$c'^2 = (1-c^2)(1-k^2c^2) = -k^2c^4 + (k^2 - k'^2)c^2 + k'^2, \tag{8.17}$$

oraz podstawiając $c = x/A$ na mocy równania (8.11), otrzymujemy:

$$\dot{x}^2 = A^2\mu^2 c'^2 = A^2\mu^2\left(-\frac{k^2x^4}{A^4} + \frac{(k^2 - k'^2)x^2}{A^2} + k'^2\right) \tag{8.18}$$

czyli

$$\dot{x}^2 = -\frac{k^2\mu^2x^4}{A^4} - \mu^2(k'^2 - k^2)x^2 + k'^2A^2\mu^2. \tag{8.19}$$

Porównujemy to z (8.9) i dostajemy:

$$\frac{k^2\mu^2}{A^2} = \frac{1}{2}|\beta|, \qquad \mu^2(k'^2 - k^2) = \alpha, \qquad k'^2A^2\mu^2 = 2E. \tag{8.20}$$

Wzory skrajne dają:

$$k^2k'^2\mu^4 = E|\beta|, \quad A^4 = \frac{4Ek^2}{|\beta|k'^2}. \tag{8.21}$$

Zatem

$$A^2 = \sqrt{\frac{4Ek^2}{|\beta|k'^2}}, \quad \mu^2 = \sqrt{\frac{E|\beta|}{k^2k'^2}}, \quad \frac{\alpha}{k'^2 - k^2} = \frac{\sqrt{E|\beta|}}{k'k}. \tag{8.22}$$

Naszym zadaniem jest obliczenie A, μ oraz k, k' mając dane α, β, E. Łatwo sprawdzić, że $(k'^2 - k^2)^2 + (2kk') \equiv (k'^2 + k^2) = 1$. Zatem istnieje φ (z pierwszej ćwiartki, bo wobec $\alpha > 0$ musi być $k' > k$):

$$k'^2 - k^2 = \cos\varphi, \quad 2kk' = \sin\varphi. \tag{8.23}$$

Wobec tego (łącząc (8.23) z ostatnim z równań (8.22))

$$\tan\varphi = 2\frac{\sqrt{E|\beta|}}{\alpha}. \tag{8.24}$$

Znamy również $\cos\varphi = \dfrac{1}{\sqrt{1+\tan^2\varphi}}$. Z pierwszego z równań (8.23) mamy (bo $k'^2 = 1 - k^2$):

$$k = \sqrt{\frac{1-\cos\varphi}{2}} = \sqrt{\frac{1 - \dfrac{1}{\sqrt{1+\tan^2\varphi}}}{2}}. \tag{8.25}$$

Zatem wstawiając (8.24) otrzymujemy

$$k = \sqrt{\frac{\sqrt{\alpha^2 + 4E|\beta|} - \alpha}{2\sqrt{\alpha^2 + 4E|\beta|}}} = \sqrt{\frac{2E|\beta|}{\alpha^2 + 4E|\beta| + \alpha\sqrt{\alpha^2 + 4E|\beta|}}}. \tag{8.26}$$

Ponadto, jako proste konsekwencje powyższych wzorów, możemy wypisać związki:

$$k = \sin\frac{\varphi}{2}, \quad k' = \cos\frac{\varphi}{2}, \quad \mu^2 = \frac{\alpha}{\cos\varphi}, \quad E = \frac{\alpha^2}{|\beta|}\tan^2\varphi,$$
$$\frac{k^2}{A^2} = \frac{|\beta|}{2\mu^2} = \frac{|\beta|}{2\alpha}\cos\varphi. \tag{8.27}$$

Przedstawione rozwiązanie zakłada warunek początkowy $x(0) = A$. Jeśli chcemy, aby jak poprzednio warunek początkowy miał postać $x(0) = 0$, to należy rozważać nieco bardziej ogólne rozwiązanie $x(t) = \mathrm{cn}(\mu(t - t_0))$. Wówczas $x(0) = 0$ wtedy, gdy $\mathrm{cn}(\mu t_0) = 0$, czyli na przykład $\mu t_0 = K$ (K – całka eliptyczna zupełna). Warunek ten spełniają też wartości t_0, różniące się od powyższej o wielokrotność połowy okresu. Podstawienie tego ogólniejszego rozwiązania w miejsce (8.11) nie zmienia przeprowadzonych tu rachunków.

8.2 Dyskretyzacja dokładna dla przypadku $\beta < 0$

Wykorzystując rozwiązanie dokładne (8.11), przesunięte o t_0, znajdziemy równanie spełnione przez dyskretyzację dokładną zadaną wzorem $X_n = x(t_n)$. Mamy więc

$$X_n = A\,\mathrm{cn}\,\mu(t_n - t_0), \quad X_{n\pm1} = A\,\mathrm{cn}\,\mu(t_n \pm h - t_0) \tag{8.28}$$

gdzie tym razem krok czasowy oznaczony jest przez h. Korzystamy z tożsamości (zob. [28]):

$$\mathrm{cn}(u \pm v) = \frac{\mathrm{cn}\,u\,\mathrm{cn}\,v \mp \mathrm{sn}\,u\,\mathrm{dn}\,u\,\mathrm{sn}\,v\,\mathrm{dn}\,v}{1 - k^2\,\mathrm{sn}^2\,u\,\mathrm{sn}^2\,v} = \frac{\mathrm{cn}\,u\,\mathrm{cn}\,v \mp \mathrm{sn}\,u\,\mathrm{dn}\,u\,\mathrm{sn}\,v\,\mathrm{dn}\,v}{\mathrm{dn}^2\,v + k^2\,\mathrm{cn}^2\,u\,\mathrm{sn}^2\,v} \tag{8.29}$$

oraz wzoru na pochodną (aby obliczyć pęd):

$$(\mathrm{cn}\,u)' = \mathrm{sn}\,u\,\mathrm{dn}\,u, \tag{8.30}$$

i obliczamy

$$X_{n+1} + X_{n-1} = \frac{2A\,\mathrm{cn}\,\mu(t_n - t_0)\,\mathrm{cn}\,\mu h}{\mathrm{dn}^2\,\mu h + k^2\,\mathrm{cn}^2\,\mu(t_n - t_0)\,\mathrm{sn}^2\,\mu h}, \tag{8.31}$$

$$X_{n+1} + X_{n-1} = \frac{2X_n\,\mathrm{cn}\,\mu h}{\mathrm{dn}^2\,\mu h + \dfrac{k^2}{A^2}X_n^2\,\mathrm{sn}^2\,\mu h}. \tag{8.32}$$

Następnie:

$$X_{n+1} - X_{n-1} = -\frac{2A\,\mathrm{sn}\,\mu(t_n - t_0)\,\mathrm{dn}\,\mu(t_n - t_0)\,\mathrm{sn}\,\mu h\,\mathrm{dn}\,\mu h}{\mathrm{dn}^2\,\mu h + k^2\,\mathrm{cn}^2\,\mu(t_n - t_0)\,\mathrm{sn}^2\,\mu h}, \tag{8.33}$$

$$X_{n+1} - X_{n-1} = \frac{2P_n \operatorname{sn}\mu h \operatorname{dn}\mu h}{\mu\left(\operatorname{dn}^2 \mu h + \frac{k^2}{A^2} X_n^2 \operatorname{sn}^2 \mu h\right)}. \tag{8.34}$$

Wyrażając wszystko przez parametr $\gamma := \operatorname{cn}\mu h$ otrzymujemy:

$$X_{n+1} + X_{n-1} = \frac{2\gamma X_n}{1-(1-\gamma^2)k^2 + \frac{(1-\gamma^2)k^2}{A^2} X_n^2}. \tag{8.35}$$

Wygodnie jest wprowadzić jeszcze dwa parametry (porównaj (8.27)):

$$\delta^2 := \frac{2(1-\gamma)}{\mu^2} = \frac{2(1-\gamma)\cos\varphi}{\alpha} \tag{8.36}$$

oraz

$$\theta = \frac{\operatorname{sn}\mu h}{\mu \operatorname{dn}\mu h}. \tag{8.37}$$

Wówczas (z (8.34), biorąc pod uwagę (8.27)) mamy:

$$P_n = \frac{X_{n+1} - X_{n-1}}{2\theta}\left(1 + \frac{1}{2}|\beta|\theta^2 X_n^2\right). \tag{8.38}$$

Dla małych h mamy $\delta \approx h$, $\theta \approx h$, a dokładniej

$$\delta = h\left(1 - \frac{1+4k^2}{24}\mu^2 h^2 + \ldots\right), \quad \theta = h\left(1 - \frac{1-2k^2}{6}\mu^2 h^2 + \ldots\right). \tag{8.39}$$

Wykonując elementarne przekształcenia możemy otrzymać jeszcze inną (równoważną) postać dyskretyzacji (8.35):

$$\frac{X_{n+1} - 2X_n + X_{n-1}}{\delta^2} = -\alpha\kappa_1 X_n - \frac{1}{2}|\beta|\kappa_2 X_n^2 (X_{n+1} + X_{n-1}), \tag{8.40}$$

gdzie

$$\kappa_1 = \frac{1 + \frac{(1-\cos\varphi)}{4\cos^2\varphi}\alpha\delta^2}{1 - \alpha\delta^2 \frac{(1-\cos\varphi)}{2\cos\varphi}\left(1 - \frac{\alpha\delta^2}{4\cos\varphi}\right)}, \tag{8.41}$$

$$\kappa_2 = \frac{1 - \frac{\alpha\delta^2}{4\cos\varphi}}{1 - \alpha\delta^2 \frac{(1-\cos\varphi)}{2\cos\varphi}\left(1 - \frac{\alpha\delta^2}{4\cos\varphi}\right)}. \tag{8.42}$$

Zauważmy, że parametry powyższe zależą od h i dla $h \to 0$ dążą do 1. Podobnie ze wzoru na P_n można wyeliminować X_{n-1} (dodając stronami (8.35) i (8.38)). Wówczas:

$$X_{n+1} = \frac{\kappa_3 X_n}{1+\frac{1}{2}|\beta|\delta^2\kappa_2 X_n^2} + \frac{\theta P_n}{1+\frac{1}{2}|\beta|\theta^2 X_n^2}, \tag{8.43}$$

gdzie

$$\kappa_3 = \frac{1-\frac{\alpha\delta^2}{2\cos\varphi}}{1-\alpha\delta^2\frac{(1-\cos\varphi)}{2\cos\varphi}\left(1-\frac{\alpha\delta^2}{4\cos\varphi}\right)} \tag{8.44}$$

jest trzecim parametrem o własnościach podobnych do κ_1 i κ_2.

8.3 Dyskretyzacja dokładna dla przypadku $\beta > 0$

Wykorzystujemy znajomość ścisłego rozwiązania oscylacyjnego (8.3):

$$x = A\,\mathrm{sn}(\omega t), \tag{8.45}$$

pamiętając, że zachodzą wzory (8.9). Postulujemy $X_{n\pm1} = A\,\mathrm{sn}(\omega t_n \pm \omega h)$, gdzie h jest krokiem czasowym dyskretyzacji. Wykorzystujemy teraz tożsamość

$$\mathrm{sn}(\omega t_n \pm \omega h) = \frac{\mathrm{sn}(\omega t_n)\,\mathrm{cn}(\omega h)\,\mathrm{dn}(\omega h) \pm \mathrm{cn}(\omega t_n)\,\mathrm{sn}(\omega h)\,\mathrm{dn}(\omega t_n)}{1-k^2\,\mathrm{sn}^2\,\omega t_n\,\mathrm{sn}^2\,\omega h} \tag{8.46}$$

i dodajemy X_{n+1} do X_{n-1}:

$$X_{n+1} + X_{n-1} = \frac{A\,\mathrm{sn}(\omega t_n)\,\mathrm{cn}(\omega h)\,\mathrm{dn}(\omega h)}{1-k^2\,\mathrm{sn}^2\,\omega t_n\,\mathrm{sn}^2\,\omega h} = \frac{2\,\mathrm{cn}(\omega h)\,\mathrm{dn}(\omega h)X_n}{1-\frac{k^2}{A^2}X_n^2\,\mathrm{sn}^2\,\omega h}. \tag{8.47}$$

Jest to poszukiwana dyskretyzacja dokładna. Rachunki analogiczne do przeprowadzonych dla przypadku $\beta < 0$ pozwalają zapisać ją w postaci:

$$\begin{aligned} X_{n+1} - 2X_n + X_{n-1} &= -\delta^2\alpha X_n + \frac{1}{2}\theta^2\beta X_n^2(X_{n+1} + X_{n-1}), \\ P_n &= \frac{X_{n+1} - X_{n-1}}{\theta} + \frac{\alpha\delta^2 X_n - \beta\theta^2 X_n^2 X_{n+1}}{2\theta}, \end{aligned} \tag{8.48}$$

gdzie

$$\begin{aligned} &\delta = \sqrt{\frac{2}{\alpha}}\sqrt{1-\mathrm{cn}(\Omega h;k)\,\mathrm{dn}(\Omega h;k)}, \qquad \theta = \frac{\mathrm{sn}(\Omega h;k)}{\Omega}, \\ &k = \sqrt{\frac{1-\sqrt{1-\frac{4E\beta}{\alpha^2}}}{1+\sqrt{1-\frac{4E\beta}{\alpha^2}}}}, \qquad \Omega = \sqrt{\frac{\alpha}{2}\left(1+\sqrt{1-\frac{4E\beta}{\alpha^2}}\right)}. \end{aligned} \tag{8.49}$$

Energia układu wynosi

$$E = \frac{1}{2}P_0^2 + \frac{1}{2}\alpha X_0^2 - \frac{1}{4}\beta X_0^4 \tag{8.50}$$

i spełnia w przypadku ruchu oscylacyjnego warunek

$$0 < E < \frac{\alpha^2}{4\beta}. \tag{8.51}$$

8.4 Dyskretyzacja Hiroty

Rygo Hirota podał następującą dyskretyzację oscylatora anharmonicznego [45, 72]

$$\begin{aligned} &\frac{X_{n+1} - 2X_n + X_{n-1}}{h^2} = -\alpha X_n + \frac{1}{2}\beta X_n^2 (X_{n+1} + X_{n-1}), \\ &P_n = \frac{X_{n+1} - X_{n-1}}{2h}. \end{aligned} \tag{8.52}$$

Po wyeliminowaniu X_{n-1} otrzymujemy wzór na punkt startowy tej dyskretyzacji

$$X_{n+1} = X_n \frac{1 - \frac{1}{2}\alpha h^2}{1 - \frac{1}{2}\beta h^2 X_n^2} + hP_n. \tag{8.53}$$

Porównanie tego wzoru z dyskretyzacją dokładną (8.48) w granicy $h \approx 0$ sugeruje następującą modyfikację tego wzoru:

$$X_{n+1} = \frac{X_n\left(1 - \frac{1}{2}\alpha h^2\right) + hP_n}{1 - \frac{1}{2}\beta h^2 X_n^2}. \tag{8.54}$$

Godną uwagi cechą tej dyskretyzacji jest posiadanie ścisłych rozwiązań. W pracy [72] można znaleźć ścisłe rozwiązanie równań dyskretnych (8.52) (poprzez funkcje eliptyczne). Oczywiście nie są to rozwiązania idące po trajektoriach układu ciągłego. Istnienie związku pomiędzy dyskretyzacją Hiroty a dyskretyzacją dokładną pozostaje jednak problemem otwartym.

8.5 Klasa dyskretyzacji zachowujących trajektorie

Rozpatrzmy następującą klasę schematów numerycznych

$$\begin{aligned} \frac{x_{n+1}-x_n}{\delta} &= \frac{p_n+p_{n+1}}{2}, \\ \frac{p_{n+1}-p_n}{\delta} &= -\frac{x_{n+1}-x_n}{2}\left(\alpha-\frac{1}{2}\beta(x_n^2+x_{n+1}^2)\right), \end{aligned} \tag{8.55}$$

gdzie δ jest dowolną funkcją x_n, x_{n+1}, p_n, p_{n+1}, h spełniającą warunek

$$\lim_{h\to 0}\frac{\delta(x_n, x_{n+1}, p_n, p_{n+1}, h)}{h} = 1, \tag{8.56}$$

Zapewniający, jak zwykle, zgodność każdego z tych schematów.

Twierdzenie. *Dowolna dyskretyzacja należąca do klasy (*8.55*) zachowuje dokładnie (z dokładnością do błędów zaokrągleń) wszystkie trajektorie układu (*8.1*) w przestrzeni fazowej*.

W szczególności, gdy $\delta = h$ otrzymujemy standardową metodę dyskretnego gradientu. Schemat lokalnie dokładnego dyskretnego gradientu otrzymamy wybierając

$$\delta = \frac{2}{\omega_n}\tan\frac{\omega h}{2}, \quad \omega = \sqrt{V''(\bar{x})}\,. \tag{8.57}$$

W naszym przypadku

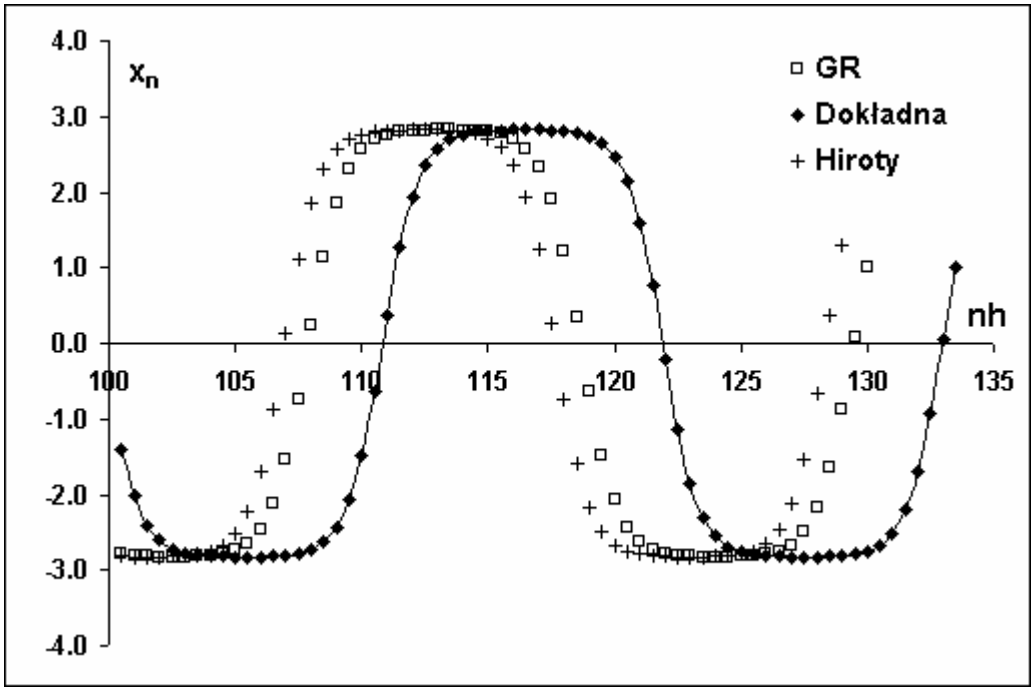
$$\omega = \sqrt{\alpha - 3\beta\bar{x}^2}\,. \tag{8.58}$$

Mamy dwa naturalne wybory dla $\bar{x}$, pierwszy to $\bar{x} = x_n$, drugi (symetryczna modyfikacja) to $\bar{x} = \frac{1}{2}(x_n + x_{n+1})$. Można również stosować wzory z rozdziału 7 otrzymując gradientowe dyskretyzacje wysokich rzędów, w tym GR-11.

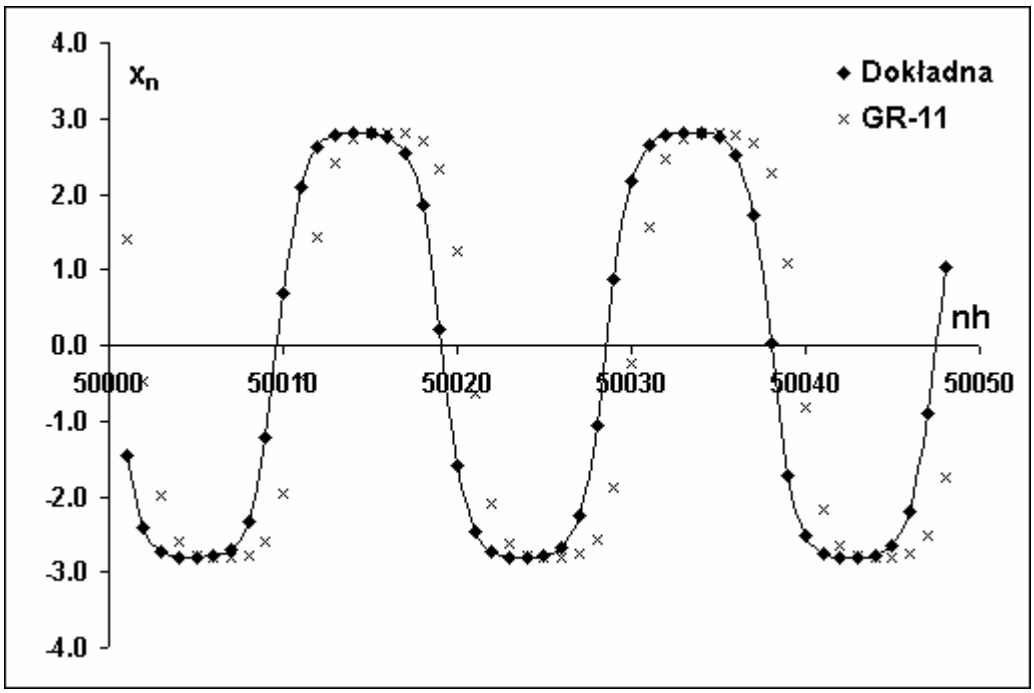
8.6 Eksperyment numeryczny

W doświadczeniu przyjęto następujące wartości stałych opisujących potencjał: $\alpha = 1$, $\beta = 0.125$. Wybór ten gwarantuje, że dla ruchu oscylacyjnego mamy $0 \le p_0 < 2$ (podobnie jak w przypadku często wykorzystywanego w tej pracy do testów potencjału wahadła matematycznego $V(x) = -\cos(x)$). W przypadku oscylatora anharmonicznego przetestowane zostały omawiane we wcześniejszych rozdziałach dyskretyzacje *leap-frog*, standardowa metoda dyskretnego gradientu (GR), lokalnie dokładne schematy dyskretnego gradientu

(GR-LEX, GR-SLEX), schematy dyskretnego gradientu wyższych rzędów (do 11 włącznie), a także dyskretyzacja Hiroty oraz wyprowadzona wyżej dyskretyzacja dokładna. Na użytek testów zaprogramowano obliczanie funkcji eliptycznych (dla rozsądnych parametrów) z dokładnością lepszą niż 10^{-14}.

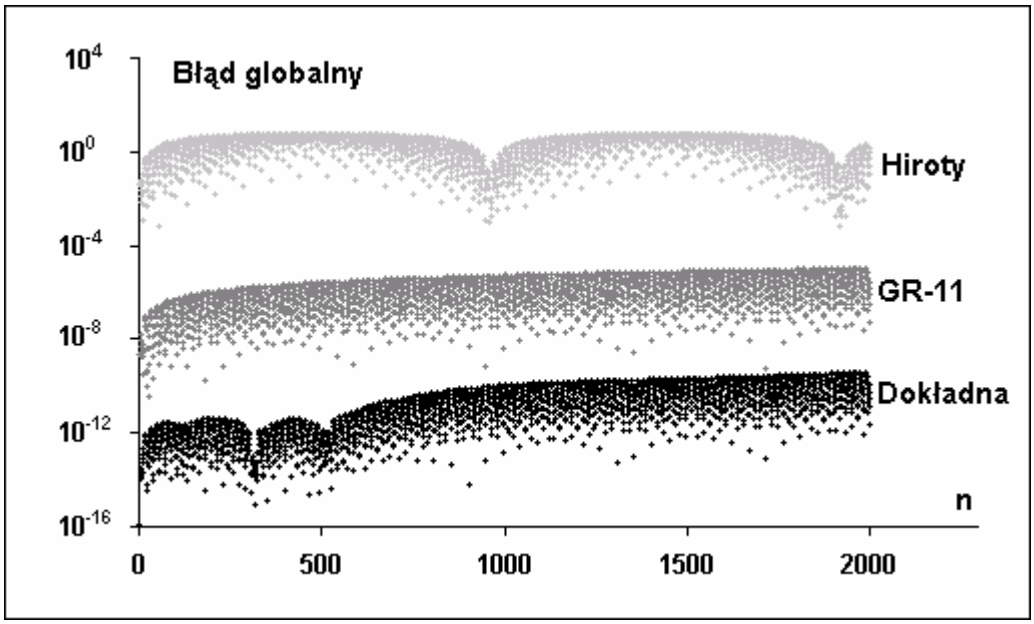


Wykres 8.1. x_n jako funkcja czasu ($t = nh$), $p_0 = 1.99999$, $T_{th} = 22.16332$, $h = 0.5$. Linia ciągła odpowiada rozwiązaniu teoretycznemu.



Wykres 8.2. x_n jako funkcja czasu ($t = nh$), $p_0 = 1.9999$, $T_{th} = 18.9072386$, $h = 1.0$. Linia ciągła odpowiada rozwiązaniu teoretycznemu.

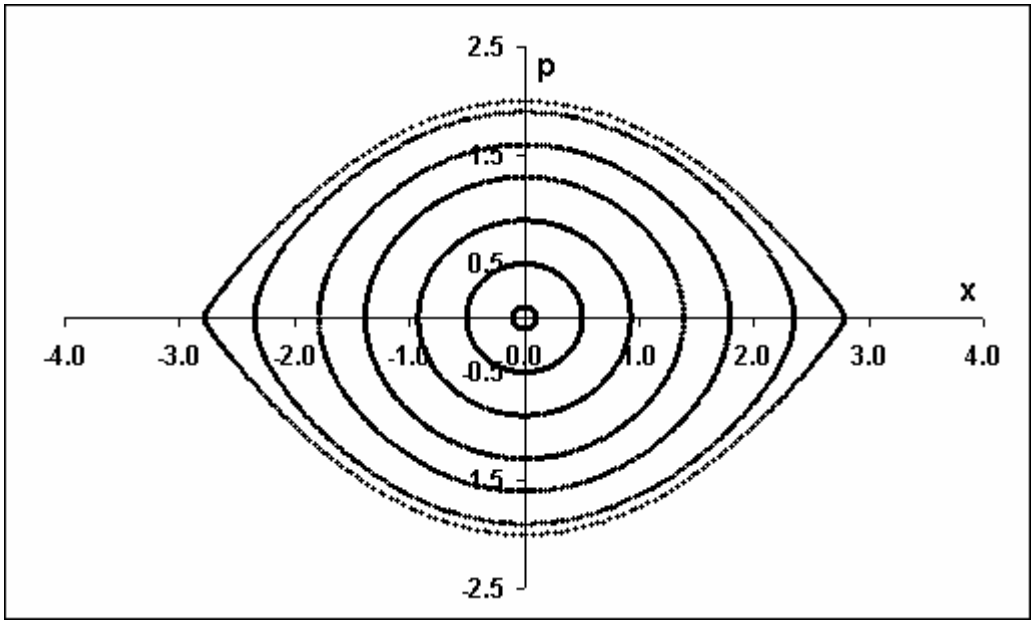
Wykresy 8.1 i 8.2 przedstawiają porównanie działania wybranych schematów numerycznych na tle dyskretyzacji dokładnej i rozwiązania ciągłego. Dyskretyzacja Hiroty, choć niezbyt dokładna, działa bardzo stabilnie w szerokim zakresie parametrów początkowych i kroków czasowych sprawując się tylko nieco gorzej niż standardowy schemat dyskretnego gradientu (wykres 8.1). Dyskretyzacja dokładna nie wydaje się być numerycznie lepsza od schematu dyskretnego gradientu rzędu 11 (GR-11, patrz rozdział 7), zwłaszcza dla małych kroków czasowych. Dopiero zwiększenie kroku czasowego (co, paradoksalnie, zmniejsza kumulację globalnego błędu w przypadku dyskretyzacji dokładnej) ujawnia lekką przewagę dyskretyzacji dokładnej (wykres 8.2). Wykres 8.3 pokazuje tempo narastania błędu globalnego dla wybranych dyskretyzacji. Skala pionowa jest logarytmiczna, zatem badane dyskretyzacje różnią się dokładnością o kilka rzędów wielkości.



Wykres 8.3. Błąd globalny w funkcji czasu ($t = nh$) dla dyskretyzacji dokładnej, Hiroty i GR-11, $p_0 = 1.999$, $T_{th} = 15.6529949$, $h = 0.5$.

Okazuje się, że dyskretyzacja dokładna pracuje na granicy stosowanej w eksperymencie dokładności obliczeniowej (10^{-14} – 10^{-15}) tylko przez pierwszych kilkanaście kroków. Dalej obserwujemy systematyczny wzrost odchylenia od teorii. Średnie tempo przyrostu błędu w wielkości $5 \cdot 10^{-15}$ na krok świadczy o tym, że natura tego przyrostu jest czysto numeryczna. Warto natomiast zwrócić uwagę na stabilność błędu globalnego dyskretyzacji Hiroty.

Rodzina schematów (8.55) pozwala na łatwe wykreślanie dokładnych trajektorii układu w przestrzeni fazowej. Wykres 8.4 prezentuje takie trajektorie wygenerowane przez schemat GR-LEX.



Wykres 8.4. Dokładne trajektorie oscylatora anharmonicznego, schemat GR-LEX, $h = 0.5$, $p_0 \in \{0.1; 0.5; 0.9; 1.3; 1.6; 1.9; 1.999\}$.

8.7 Podsumowanie

W rozdziale tym przedstawiono ścisłe rozwiązanie oscylatora anharmonicznego oraz przedstawiono sposób konstrukcji jego dyskretyzacji dokładnej. Podano też postać klasy dyskretyzacji zachowującej trajektorie układu. Testy pokazały, że dyskretyzacja dokładna charakteryzuje się dość szybkim przyrostem błędu globalnego spowodowanym niedokładnością obliczeń. Podniesienie tej dokładności zmniejszy proporcjonalnie kumulację błędu.

Dyskretyzacja zaproponowana przez Hirotę okazała się nieco mniej dokładna od standardowego dyskretnego gradientu, ale bardzo stabilna i zdolna do symulowania układu w szerokim zakresie parametrów początkowych i kroków czasowych. Jest to schemat zadany w sposób jawny, co przekłada się na szybkość działania. Potwierdziła się też stabilność i wysoka dokładność skonstruowanych w rozdziale 7 metod gradientowych wyższych rzędów.

9 Geometryczne dyskretyzacje problemu Keplera

9.1 Wprowadzenie

W rozdziale tym zajmiemy się ważnym i szeroko omawianym zagadnieniem mechaniki klasycznej jakim jest zagadnienie Keplera [2, 25, 75]. Przeanalizujemy działanie kilku bardzo dobrych, geometrycznych, dyskretyzacji, do których należą symplektyczna *metoda Eulera* [41], schemat Störmera-Verleta (*leap-frog*) [40] i zmodyfikowanej metody punktu środkowego [37, 54] (schemat dyskretnego gradientu zachowujący energię) oraz dwie dyskretyzacje zachowujące trajektorie i wszystkie całki ruchu [16, 52], z tym, że pierwsza z tych dyskretyzacji została w istotny sposób zmodyfikowana i udoskonalona w niniejszej pracy.

Całkowanie geometryczne polega na znajdywaniu numerycznych rozwiązań równań różniczkowych przy jednoczesnym zachowaniu pewnych fizycznych lub matematycznych ich własności w sposób dokładny [2]. Dla problemu Keplera znaleziono w ostatnich latach szereg dyskretyzacji zachowujących całki ruchu i trajektorie [16, 52, 73, 74, 86]. Zaprezentujemy tu schemat zaproponowany w pracy [16], który udało się zmodyfikować w taki sposób, aby odtwarzał dokładnie kształt orbit teoretycznych. Drugą tego rodzaju metodą jest dokładna dyskretyzacja znaleziona przez Kozlova [52], który zastosował transformację Kustaanheimo-Stiefela [53]. Ta zaawansowana matematycznie transformacja zostanie także omówiona w tym rozdziale, ale w sposób stosunkowo elementarny, oparty na artykule [18]. Eksperyment numeryczny objął wszystkie wymienione dyskretyzacje, w pierwszej części tego rozdziału koncentrując się na metodach standardowych, a w drugiej na porównaniu opracowanej przez nas metody ze schematem Kozlova.

9.2 Standardowe schematy geometryczne

Rozpatrujemy klasyczny problem Keplera zadany równaniami

$$\frac{d\vec{p}}{dt} = -k\frac{\vec{r}}{r^3} = \vec{f}(\vec{r}), \quad \vec{p} = m\frac{d\vec{r}}{dt}, \quad m,k = const. \tag{9.1}$$

Przypomnijmy, że moment pędu $\vec{L}$, anergia całkowita E oraz wektor Rungego-Lenza $\vec{A}$ (wskazujący na peryhelium orbity) dane wzorami

$$\vec{L} = \vec{r} \times \vec{p}, \qquad E = \frac{(\vec{p})^2}{2m} - \frac{k}{r},$$
$$\vec{A} = \frac{\vec{p} \times \vec{L}}{m} - k\frac{\vec{r}}{r}, \tag{9.2}$$

są całkami ruchu dla równań (9.1) [25, 93].

Dokładne rozwiązanie $\vec{r}(t)$ (z wyjątkiem przypadków szczególnych, takich jak orbita kołowa) nie jest znane, jednak można wyznaczyć ściśle orbity, które we współrzędnych biegunowych opisane są wzorami [93]:

$$r = \frac{p}{1 + e\cos(\varphi - \varphi_0)}, \qquad p = \frac{L^2}{km},$$
$$e = \sqrt{1 + \frac{2EL^2}{mk^2}}, \tag{9.3}$$

gdzie $L = \left|\vec{L}\right|$ i φ_0 są stałymi wyznaczonymi przez warunki początkowe.

Schemat Eulera (EU) otrzymujemy bezpośrednio z równań (9.1) zastępując pochodne odpowiednimi ilorazami różninowymi:

$$\frac{\vec{p}_{n+1} - \vec{p}_n}{h} = \frac{\vec{r}_{n+1} - 2\vec{r}_n + \vec{r}_{n-1}}{h^2} = -k\frac{\vec{r}_n}{r^3},$$
$$\vec{p}_{n+1} = m\frac{\vec{r}_{n+1} - \vec{r}_n}{h}, \tag{9.4}$$

gdzie h = const jest skończonym przyrostem siatki czasowej. Ściślej mówiąc, jest to jeden z symplektycznych schematów Eulera, porównaj wzór (4.23).

Ogólna postać metody Störmera-Verleta (*leap-frog*, w skrócie: LF) [40], zastosowanej do równania (9.1), jest następująca:

$$\vec{p}_{n+\frac{1}{2}} = \vec{p}_n + \frac{1}{2}h\vec{f}(r_n),$$
$$\vec{r}_{n+1} = \vec{r}_n + \frac{h}{m}\vec{p}_{n+\frac{1}{2}}, \tag{9.5}$$
$$\vec{p}_{n+1} = \vec{p}_{n+\frac{1}{2}} + \frac{1}{2}h\vec{f}(r_{n+1}).$$

Trzecią z omawianych tu metod numerycznych, zmodyfikowaną metodę punktu środkowego (metoda dyskretnego gradientu) [37, 54], zastosowaną do problemu Keplera, definiują wzory:

$$\vec{r}_{n+1} = \vec{r}_n + \frac{h}{m}\frac{\vec{p}_{n+1} + \vec{p}_n}{2},$$
$$\vec{p}_{n+1} = \vec{p}_n - h\frac{V(r_{n+1}) - V(r_n)}{r_{n+1} - r_n}\frac{\vec{r}_{n+1} + \vec{r}_n}{r_{n+1} + r_n}, \quad V(r) = \frac{k}{r}. \tag{9.6}$$

Równania (9.6) rozwiązujemy iteracyjnie ($(\vec{r},\vec{p}) \to (\tilde{\vec{r}},\tilde{\vec{p}})$)

$$\begin{aligned}\tilde{\vec{r}} &= \vec{r}_n + \frac{h}{m}\frac{\vec{p}+\vec{p}_n}{2},\\ \tilde{\vec{p}} &= \vec{p}_n - h\frac{V(r)-V(r_n)}{r-r_n}\frac{\vec{r}+\vec{r}_n}{r+r_n},\end{aligned} \tag{9.7}$$

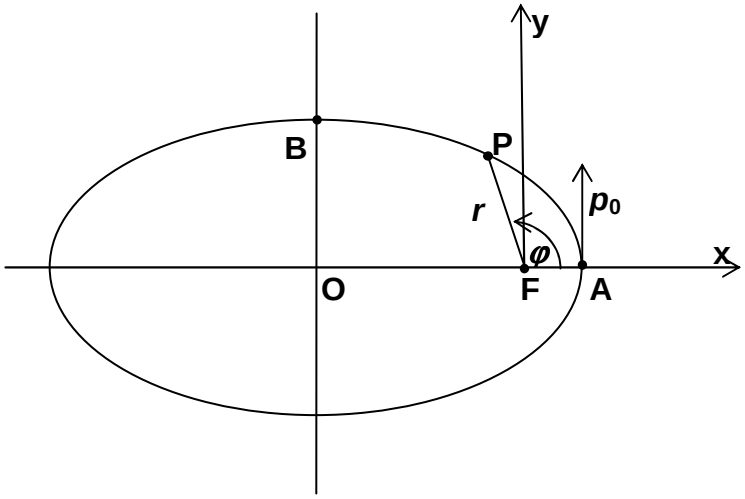
przy czym do wyznaczania punktu startowego możemy zastosować, jako predyktor, np. metodę *leap-frog*. W przypadku problemu Keplera powyższe wzory nieco się upraszczają. Równania (9.6) przyjmują postać:

$$\begin{aligned}\vec{r}_{n+1} &= \vec{r}_n + \frac{h}{m}\frac{\vec{p}_{n+1}+\vec{p}_n}{2},\\ \vec{p}_{n+1} &= \vec{p}_n - \frac{hk(\vec{r}_{n+1}+\vec{r}_n)}{r_n r_{n+1}(r_{n+1}+r_n)},\end{aligned} \tag{9.8}$$

natomiast (9.7) sprowadza się do:

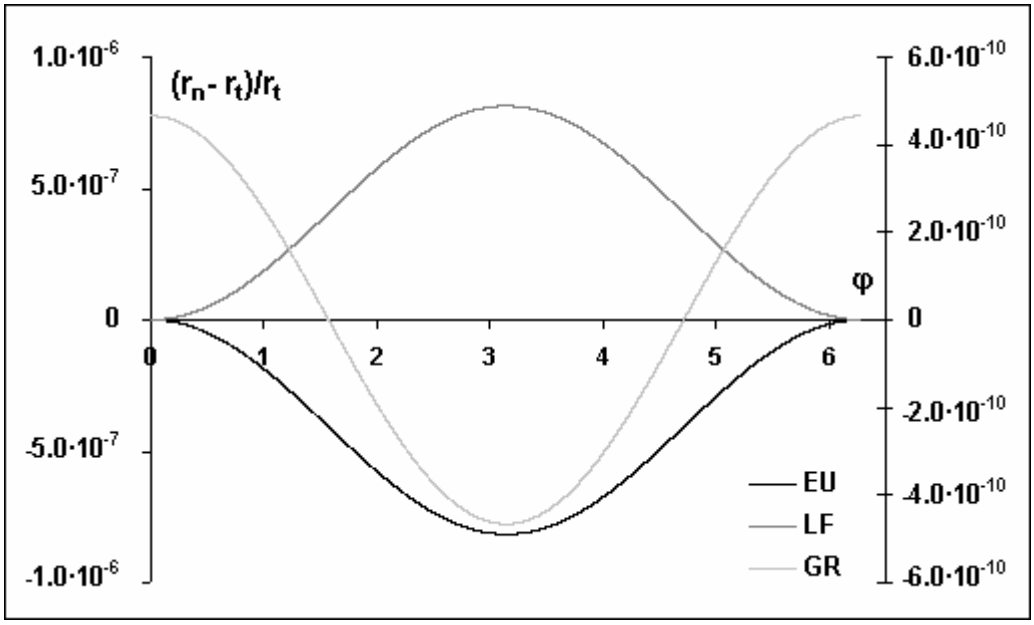
$$\begin{aligned}\tilde{\vec{r}} &= \vec{r}_n + \frac{h}{m}\frac{\vec{p}+\vec{p}_n}{2},\\ \tilde{\vec{p}} &= \vec{p}_n - \frac{hk(\vec{r}+\vec{r}_n)}{r_n r(r+r_n)}.\end{aligned} \tag{9.9}$$

Na potrzeby eksperymentu numerycznego przyjmiemy warunki początkowe ruchu takie, jak na rysunku 9.1, zaś jako przykład liczbowy rozważymy intuicyjny ruch satelity Ziemi, który w chwili $t = 0$ znajduje się na wysokości 200 km nad jej powierzchnią. Przyjmijmy, że $M = 5{,}97\cdot10^{24}$ i $G = 6{,}67\cdot10^{-11}$.

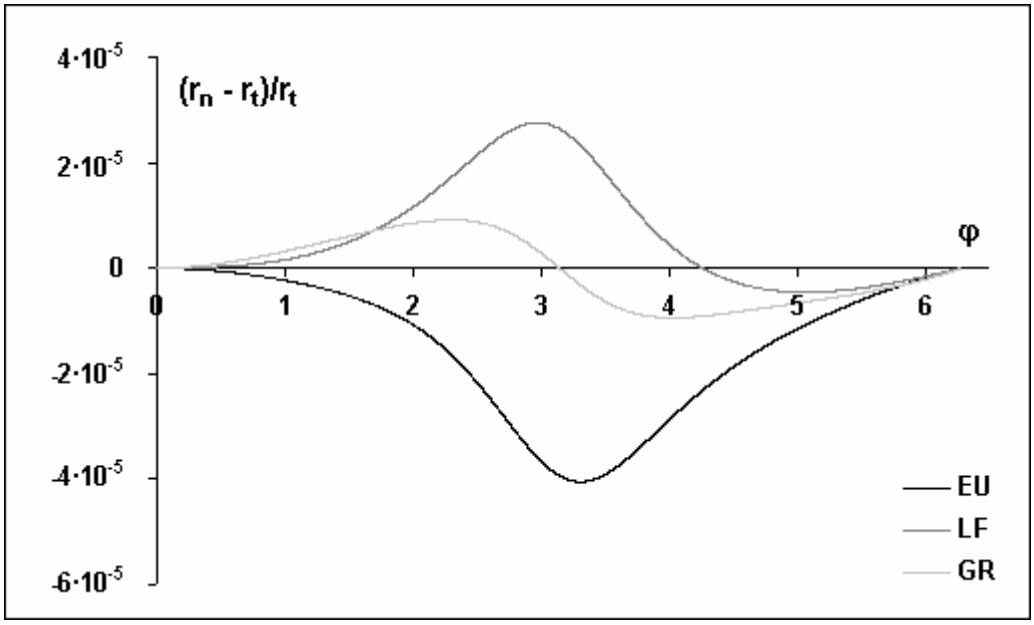


Rysunek 9.1. Warunki początkowe ruchu.

Wykresy 9.1 i 9.2 pokazują odchylenie od teorii, jakie dają trzy przedstawione w tym podrozdziale dyskretyzacje w czasie pierwszego obiegu dla orbity kołowej i eliptycznej.



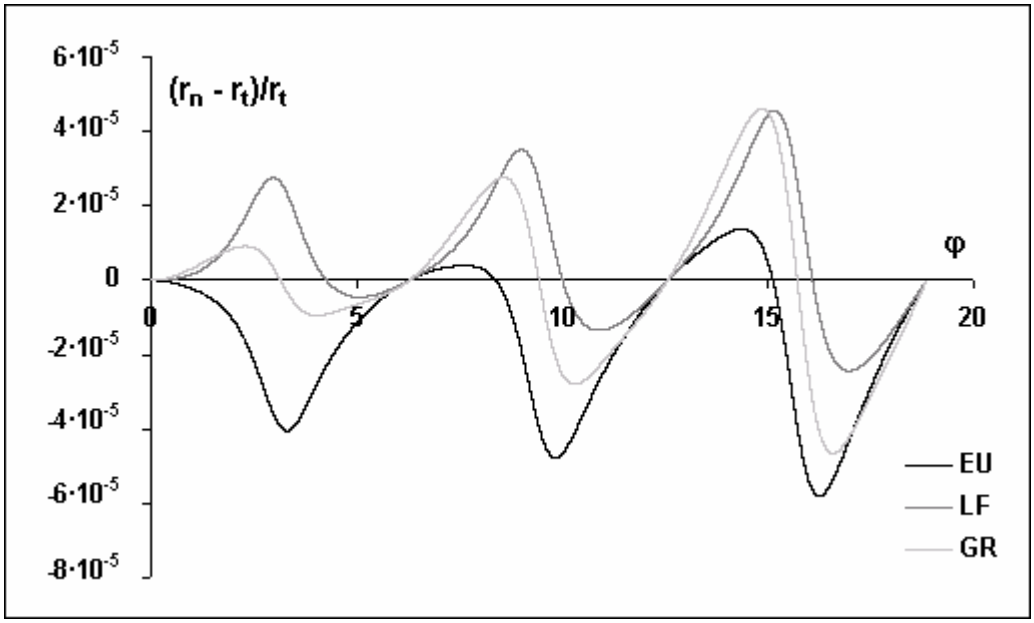
Wykres 9.1. Dyskretyzacje standardowe. Względne odchylenie od teoretycznej orbity kołowej w czasie jej pierwszego obiegu, h = 1.07, N ≈ 4924 kroków/obieg, p_0 = 7785.1577, $r_0 = 6.57 \cdot 10^6$. Lewa oś pionowa odpowiada schematom EU i LF, prawa – GR.



Wykres 9.2. Dyskretyzacje standardowe. Względne odchylenie od teoretycznej orbity eliptycznej w czasie jej pierwszego obiegu, h = 5.2, N ≈ 5000 kroków/obieg, p_0 = 10000, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$.

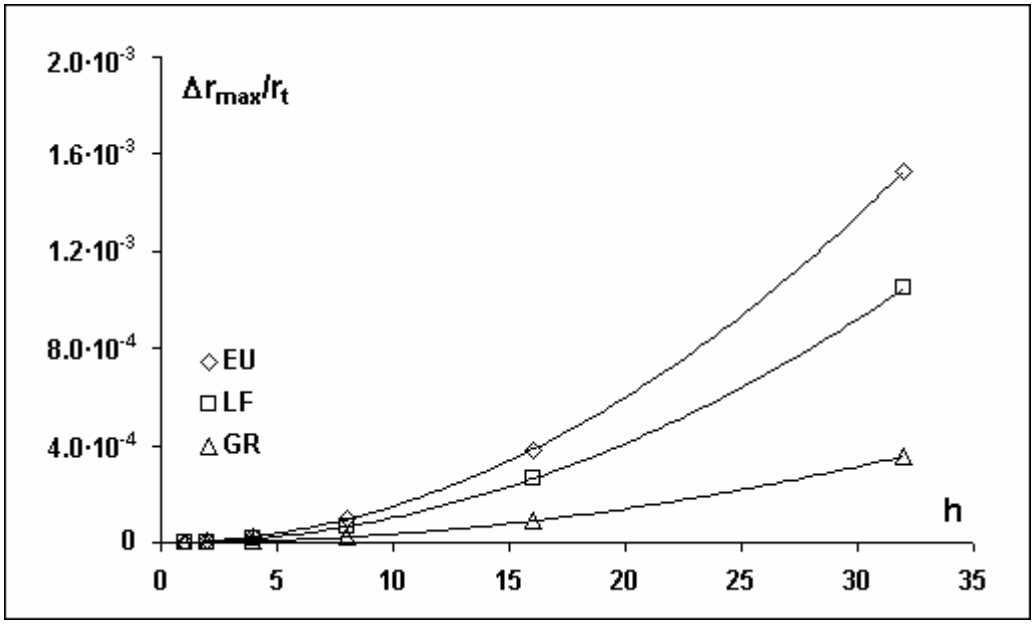
Z wykresu 9.1 widzimy, że orbitę kołową prawie 10^4 razy dokładniej od dwóch pozostałych schematów odtwarza metoda dyskretnego gradientu. W przyjętych warunkach jej najmniejsza dokładność względna wynosząca ok. $5 \cdot 10^{-10}$ (odpowiada to ok. 0.0033 w liczbach bezwzględnych) jest z grubsza równa największej dokładności konkurentek występującej na początku i końcu

pierwszego obiegu (w ich przypadku maksymalne odchylenie bezwzględne wynosi ok. 5). Stwierdzamy, że początkowe zachowania omawianych integratorów są uderzająco różne. Przewaga metody dyskretnego gradientu zdecydowanie maleje w przypadku orbity eliptycznej (wykres (9.2)). Maksymalne względne odchylenia od teorii (ok. $9 \cdot 10^{-6}$) są tu w jej przypadku tylko 3-4 razy mniejsze niż konkurentek (w liczbach bezwzględnych ok. 10^{3}), a minimalne odchylenia (ok. 10^{-10}) wszystkich trzech dyskretyzacji stają się praktycznie równe. Krzywe na tym wykresie różnią się zdecydowanie, ale nie tak drastycznie, jak na 9.1. Ciekawe jest to, że wraz z upływem czasu przebiegi odchyleń od teorii trzech omawianych dyskretyzacji upodabniają się do siebie (wykres 9.4). Jest to efekt obrotu elips dyskretnych wokół punktu leżącego w pobliżu miejsca startowego (perygeum), który to obrót maskuje subtelniejsze efekty numeryczne.



Wykres 9.3. Dyskretyzacje standardowe. Względne odchylenie od teoretycznej orbity eliptycznej w czasie trzech początkowych obiegów, h = 5.2, N ≈ 5000 kroków/obieg, p_0 = 10000, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$.

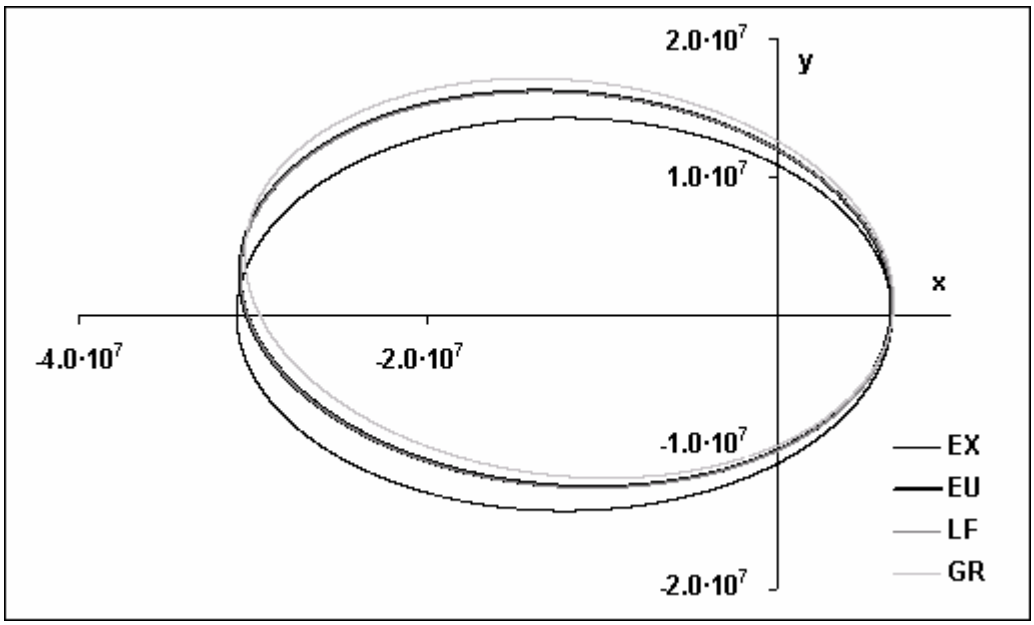
Dyskretyzacje standardowe działają często sensownie nawet przy dużych wartościach kroku czasowego, a ich wspólną cechą jest szybka poprawa dokładności towarzysząca zmniejszaniu h. Wykres 9.4 pokazuje wpływ wartości kroku czasowego na maksymalne odchylenie od teorii trzech badanych dyskretyzacji w przypadku orbity eliptycznej (tylko pierwszy obieg).



Wykres 9.4. Względne maksymalne odchylenie od orbity eliptycznej w pierwszym obiegu jako funkcja kroku czasowego, $p_0 = 10000$, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 3.232348 \cdot 10^6$.

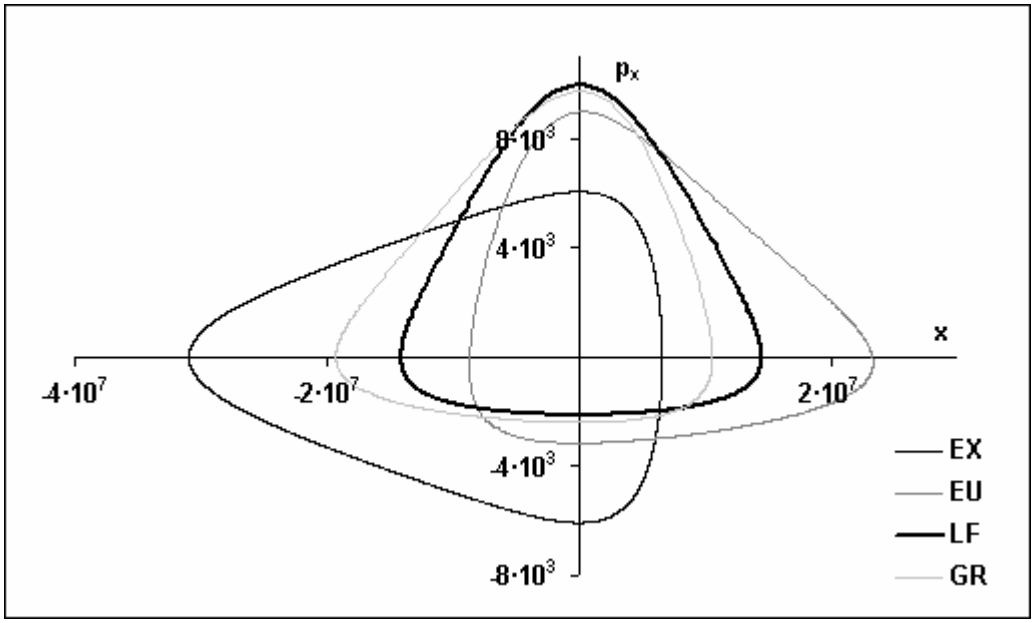
Do każdej dyskretyzacji z wykresu 9.4 bardzo dobrze pasuje wielomian stopnia drugiego, przy czym współczynniki przy x^2 mają wartości $1.49 \cdot 10^{-6}$ (EU), $1.02 \cdot 10^{-6}$ (LF) i $3.48 \cdot 10^{-7}$ (GR). Pozostałe współczynniki wielomianu są o dwa rzędy wielkości mniejsze.

Orbity badanych dyskretyzacji nie tylko odbiegają od kształtu eliptycznego, ale dość szybko zmieniają swoje położenie. Ilustruje to wykres 9.5.



Wykres 9.5. Tory ruchu dyskretyzacji standardowych po upływie n = 10000 obiegów na tle rozwiązania dokładnego (EX), $h = 5.2$, $N \approx 5000$ kroków/obieg, $p_0 = 10000$, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$.

Najszybszy obrót (mimo najmniejszych odchyleń w pierwszym obiegu) daje zmodyfikowana metoda punktu środkowego. Nieco wolniej i prawie jednakowo obracają się orbity uzyskane z pomocą metody Eulera i *leap-frog* (najwolniejszy obrót). Mimo, że metoda dyskretnego gradientu okazała się o 4 rzędy wielkości lepsza od dwóch pozostałych schematów w przypadku orbity kołowej i 3-4 razy lepsza w przypadku orbity eliptycznej. Można sądzić, że jej przewaga wynika z zachowywania energii układu. Niestety dotyczy to tylko początkowych obiegów orbity. Wraz z upływem czasu dyskretny gradient traci swoją przewagę, co ilustruje wykres 9.6 pokazujący ewolucje wybranej krzywej w przestrzeni fazowej.



Wykres 9.6. Dyskretne krzywe fazowe badanych schematów numerycznych po upływie n = 50000 obiegów na tle rozwiązania dokładnego (EX), h = 48.8, N ≈ 524 kroków/obieg, p_0 = 10000, $r_0 = 6.57 \cdot 10^6$.

Przewaga dyskretyzacji gradientowej na wykresie 9.6 jest słabo widoczna, co więcej najmniejsze zniekształcenie geometryczne wydaje się dawać metoda *leap-frog*.

9.3 Przykład zachowawczej dyskretyzacji problemu Keplera

Spośród kilku zachowawczych dyskretyzacji 3-wymiarowego problemu Keplera znalezionych w ostatnim czasie przedstawimy schemat zaproponowany przez promotora [16]. Dyskretna ewolucja $(\vec{r}_n, \vec{p}_n) \to (\vec{r}_{n+1}, \vec{p}_{n+1})$ zada jest wzorami

$$\frac{\Delta \vec{p}_n}{\Delta t_n} = -\frac{k\vec{r}_{n+1}}{\alpha r_{n+1}^2 r_n \cos\delta}, \quad \vec{p}_n = m\frac{\Delta \vec{r}_n}{\Delta t_n}, \tag{9.10}$$

gdzie *m*, *k* są stałymi, $\Delta\vec{r}_n = \vec{r}_{n+1} - \vec{r}_n$, $\Delta\vec{p}_n = \vec{p}_{n+1} - \vec{p}_n$, $\Delta t_n = t_{n+1} - t_n$, przy czym dyskretyzację charakteryzuje dodatkowo stały parametr α. Schemat pracuje ze zmiennym krokiem czasowym wybranym tak, że kąt pomiędzy $\vec{r}_{n+1}$ i $\vec{r}_n$ dany przez $2\delta_n = 2\delta = \text{const}$ nie zależy od *n*. Kąt ten wyraża się przez dane początkowe $\vec{r}_0$, $\vec{p}_0$ i Δt_0 następująco:

$$\cos 2\delta = \frac{m r_0^2 + \vec{r}_0 \cdot \vec{p}_0 \Delta t_0}{r_0 \sqrt{m^2 r_0^2 + 2m\vec{r}_0 \cdot \vec{p}_0 \Delta t_0 + p_0^2 (\Delta t_0)^2}}. \tag{9.11}$$

Granica ciągła odpowiada warunkowi $\Delta t_0 \to 0$. Wówczas $\delta \to 0$, $\vec{r}_n \to \vec{r}$, $\vec{p}_n \to \vec{p}$, a równania (9.10) dążą w granicy do (9.1) przy założeniu, że $\alpha \to 1$ (jest to więc warunek zgodności dla tego schematu numerycznego).

Dyskretyzację (9.10) można zapisać w postaci trzech jawnych równań dyskretnych:

$$\begin{aligned}
\vec{r}_{n+1} &= \vec{r}_n + \frac{\vec{p}_n \Delta t_n}{m} \\
\vec{p}_{n+1} &= \vec{p}_n - \frac{k\vec{r}_{n+1} \Delta t_n}{\alpha r_{n+1}^{\ 2} r_n \cos\delta} \\
\Delta t_{n+1} &= \frac{\Delta t_n}{2\cos(2\delta)\dfrac{r_n}{r_{n+1}} - 1 + \dfrac{k \Delta t_n^{\ 2}}{\alpha m r_{n+1}^{\ 2} r_n \cos\delta}}.
\end{aligned} \tag{9.12}$$

Bezpośredni rachunek pokazuje, że posiada ona następujące całki ruchu:

$$\begin{aligned}
&\vec{L}_n = \alpha \vec{R}_n \times \vec{p}_n, \quad E_n = \frac{(\vec{p}_n)^2}{2m} - \frac{k}{\alpha R_n}, \\
&\vec{A}_n = \frac{\vec{p}_n \times \vec{L}_n}{m} - \frac{k\vec{R}_n}{R_n},
\end{aligned} \tag{9.13}$$

gdzie

$$\begin{aligned}
\vec{R}_n &:= \frac{r_{n+1}\vec{r}_n + r_n \vec{r}_{n+1}}{r_n + r_{n+1}}, \\
R_n &:= \left|\vec{R}_n\right| = \frac{2 r_n r_{n+1} \cos\delta}{r_n + r_{n+1}},
\end{aligned} \tag{9.14}$$

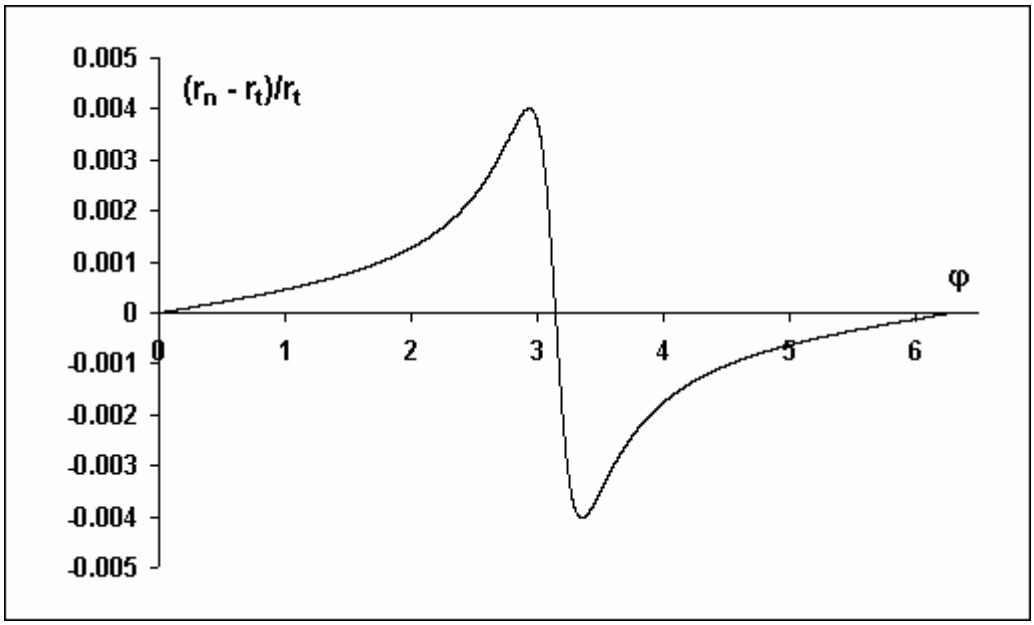
a każda dyskretna orbita sparametryzowana przez L' i E' ($L' := \left|\vec{L}_n\right|$ i $E' := E_n$) odpowiada orbicie generowanej przez ciągły problemu Keplera (9.1) określonej przez *L* i *E* spełniających warunki

$$L' = L\sqrt{\alpha \cos\delta}, \qquad E' = \frac{E\cos\delta}{\alpha} - \frac{m k^2 \sin^2\delta}{2\alpha L^2 \cos\delta}. \tag{9.15}$$

W granicy ciągłej: $L' \to L$ i $E' \to E$.

9.4 Dyskretyzacja dokładnie zachowująca orbity keplerowskie

Przestawiona w poprzednim rozdziale dyskretyzacja (9.10) zachowuje szereg wielkości będących dyskretnymi analogami całek ruchu ciągłego problemu Keplera (9.1). Jednak odpowiedniość między działaniem modelu dyskretnego i ciągłego nie idzie tu w parze z dokładnym odtworzeniem liczbowym rzeczywistych parametrów fizycznych ruchu ciągłego, przy czym nie pomaga istnienie parametru α. Na wykresie 9.7 zamieszczono przykład działania modelu (9.10) w przypadku orbity kołowej.



Wykres 9.7. Względne odchylenie modelu (9.10) od orbity teoretycznej w czasie pierwszego obiegu w funkcji φ (α = 1.0, δ = 0.000592477, $p(0)$ = 7785.1577, $r(0) = 6.57 \cdot 10^6$).

Widzimy, że rozwiązanie dyskretne różni się istotnie od teorii. Naszym celem jest taka modyfikacja dyskretyzacji (9.10), aby odtwarzała dokładnie ciągłe orbity keplerowskie. Okazuje się, że jest to możliwe.

Stałe ruchu przypadku ciągłego mogą być wyrażone przez warunki początkowe

$$L = r(0)\,p(0)\sin\theta, \qquad E = \frac{p^2(0)}{2m} - \frac{k}{r(0)}, \tag{9.16}$$

gdzie θ jest kątem pomiędzy $\vec{r}(0)$ i $\vec{p}(0)$. Jeżeli startujemy z peryhelium lub perygeum, wówczas $\sin\theta = 1$ (tak jest w naszym przypadku).

Dyskretne analogi momentu pędu i energii (L', E') są powiązane ze swoimi ciągłymi odpowiednikami wzorami (9.15), ale jednocześnie (korzystając z równań (9.13) i (9.14)) możemy napisać

$$L' = \frac{m\alpha r_0 r_1 \sin 2\delta}{\Delta t_0}, \tag{9.17}$$

$$E' = \frac{p_0^2}{2m} - \frac{k}{2\alpha\cos\delta}\left(\frac{1}{r_0} + \frac{1}{r_1}\right). \tag{9.18}$$

Ponadto

$$\vec{p}_0 = \frac{m(\vec{r}_1 - \vec{r}_0)}{\Delta t_0}, \quad p_0^2 = \left(\frac{m}{\Delta t_0}\right)^2 (r_1^2 + r_0^2 - 2r_1 r_0 \cos 2\delta). \tag{9.19}$$

Przyjmijmy $\alpha \equiv 1$ (pozostawiając ją we wzorach), $\sin\theta = 1$ oraz $r_0 = r(0)$ (chcemy odtworzyć dokładnie orbitę ciągłą). Załóżmy też, że zadana jest δ. Problem jest następujący: mając $p(0)$, $r(0)$, czyli również L, E należy obliczyć r_0, r_1, p_0 oraz Δt_0.

Korzystając z ciągłych warunków początkowych (9.16) i związków (9.15) obliczamy L', E':

$$\begin{aligned} L' &= r(0)\,p(0)\sqrt{\alpha\cos\delta}, \\ E' &= \left(\frac{p^2(0)}{2m} - \frac{k}{r(0)}\right)\frac{\cos\delta}{\alpha} - \frac{mk^2\sin^2\delta}{2\alpha(r(0)\,p(0))^2\cos\delta}. \end{aligned} \tag{9.20}$$

Zostają nam trzy równania (9.17), (9.18), (9.19) na trzy niewiadome: r_0, p_0, Δt_0. Z (9.17) obliczamy Δt_0 i wstawiamy do (9.19) otrzymując

$$p_0^2 = \left(\frac{L'}{\alpha r_0 r_1 \sin 2\delta}\right)^2 (r_1^2 + r_0^2 - 2r_1 r_0 \cos 2\delta), \tag{9.21}$$

co, po podstawieniu do (9.18), daje nam równanie kwadratowe na $\frac{1}{r_1}$:

$$\begin{aligned} &\frac{L'^2}{2m\alpha^2\sin^2 2\delta}\frac{1}{r_1^2} - \left(\frac{L'\cos 2\delta}{m\alpha^2 r(0)\sin 2\delta} + \frac{k}{2\alpha\cos\delta}\right)\frac{1}{r_1} + \\ &\frac{L'^2}{2m\alpha^2 r(0)^2\sin^2 2\delta} - \frac{k}{2\alpha r(0)\cos\delta} - E' = 0. \end{aligned} \tag{9.22}$$

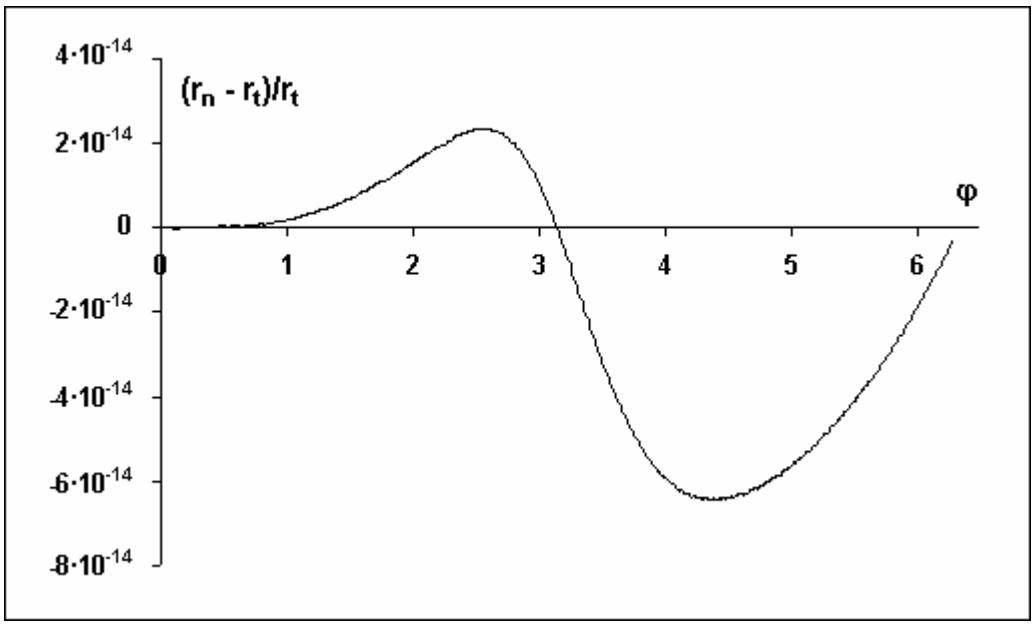
Mając r_1 z (9.22) dostajemy bezpośrednio p_0 z (9.19) i Δt_0 z (9.17).

Ostateczny wynik obliczeń (nowy warunek początkowy dla dyskretyzacji (9.12)) przedstawia się następująco:

$$
\begin{aligned}
&\mu_0 := \frac{2mk}{r(0)\,p^2(0)} \\
&\vec{r}_0 = \big(r(0),\,0\big) \\
&\vec{p}_0 = \frac{p(0)}{\sqrt{\alpha\cos\delta}}\left(-\frac{1}{2}\mu_0\sin\delta,\,\cos\delta\right) \\
&\Delta t_0 = \frac{2mr_0\sin\delta\sqrt{\alpha\cos\delta}}{p(0)(\cos(2\delta)+\mu_0\sin^2\delta)}.
\end{aligned}
\tag{9.23}
$$

Okazuje się, że dyskretyzacja (9.12) po zastosowaniu parametrów startowych (9.23) odtwarza dokładnie orbitę teoretyczną przypadku ciągłego. Przedstawia to wykres 9.8 (analogiczny do 9.7).

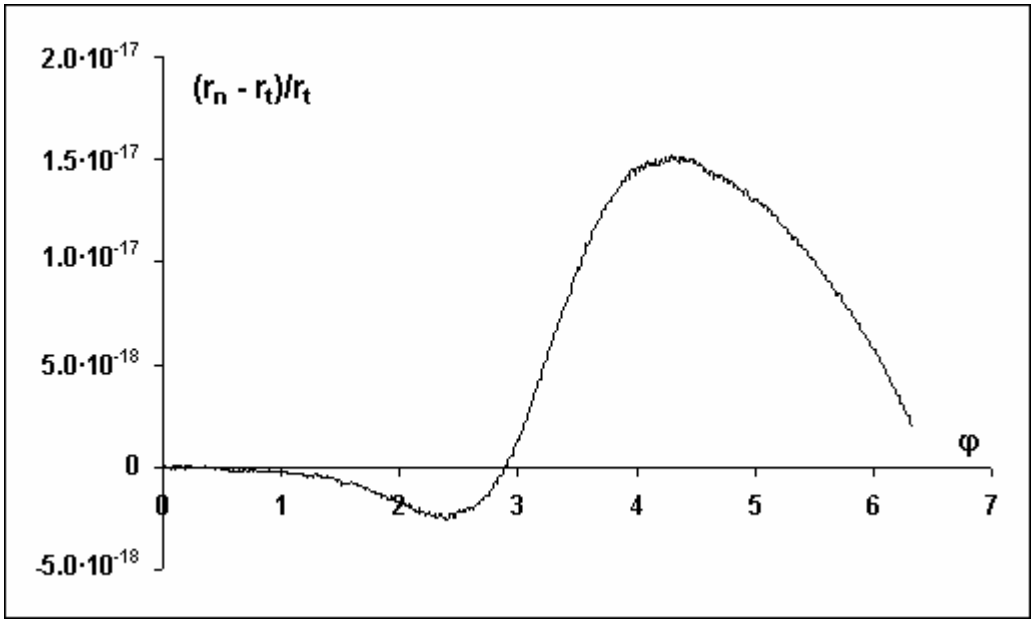
Powyższe rozumowanie jest przykładem ważnego, a chyba niezbyt docenianego problemu. Otóż dyskretyzacja (9.12) radykalnie poprawia swą dokładność przy prawidłowym (optymalnym) dopasowaniu warunków początkowych.



Wykres 9.8 Dyskretyzacja zmodyfikowana. Względne odchylenie od teoretycznej orbity eliptycznej, δ = 0.006, N ≈ 525 kroków/obieg, p_0 = 10000, $r_0 = 6.57\cdot10^6$, $r_{tmax} = 13.232348\cdot10^6$. Dokładność obliczeń – 14 ÷ 15 cyfr znaczących.

Eksperyment pokazuje, że względne odchylenie od teorii osiągnęło wartości na granicy stosowanej dokładności obliczeniowej (14 ÷ 15 cyfr znaczących). Drobne ząbki na krzywej odpowiadają skokowym zmianom cyfry na ostatnim miejscu znaczącym. Stosunkowo gładki przebieg krzywej może sugerować, że obserwowane na nim odchylenia od teorii mają charakter systematyczny związany z testowaną dyskretyzacją. Mając na uwadze te wątpliwości,

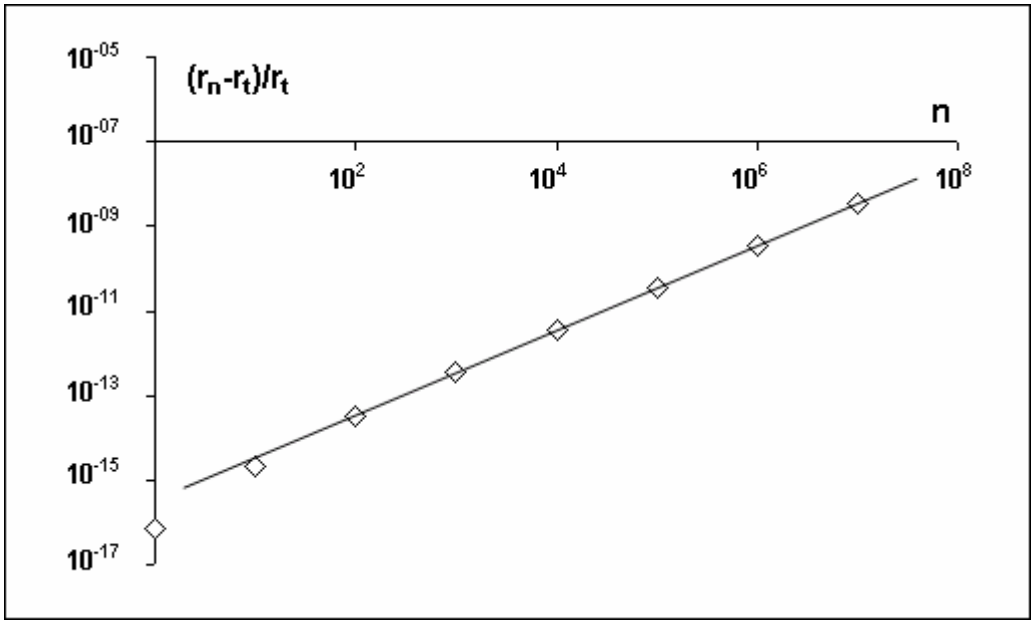
powtórzono obliczenia z wyższą dokładnością (19 ÷ 20 cyfr znaczących). Wyniki dotyczące pierwszego obiegu orbity znajdziemy na wykresie 9.9.



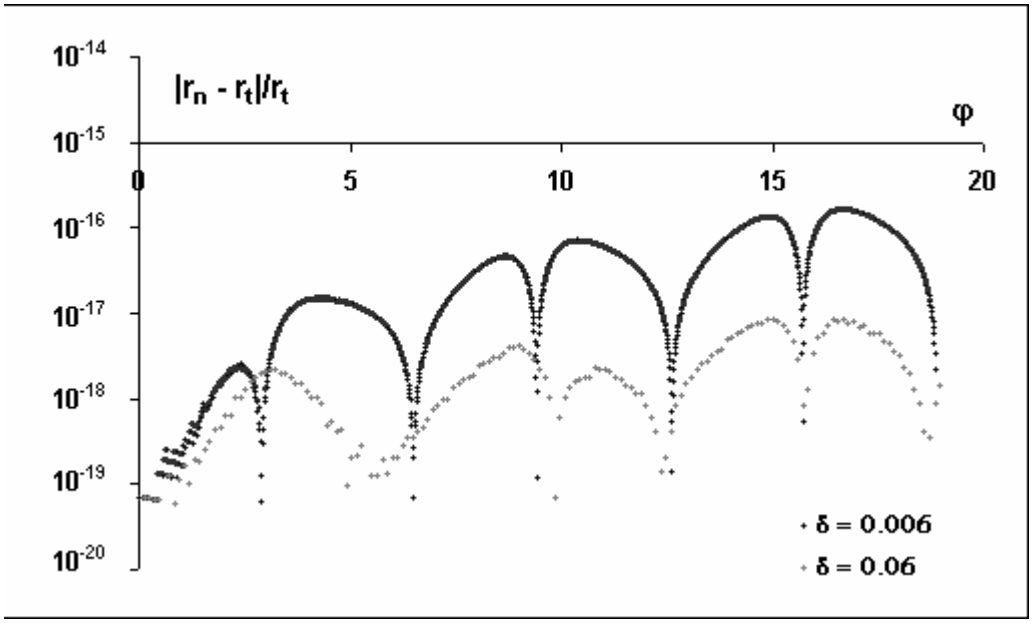
Wykres 9.9. Dyskretyzacja zmodyfikowana. Względne odchylenie od teoretycznej orbity eliptycznej, $\delta = 0.006$, $N \approx 525$ kroków/obieg, $p_0 = 10000$, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$. Dokładność obliczeń – 19 ÷ 20 cyfr znaczących.

Wykresy 9.8 i 9.9 pokazują, że dokładność zmodyfikowanej dyskretyzacji (9.12) ograniczają jedynie błędy numeryczne, a przebieg krzywej obrazującej względne odchylenie orbity dyskretnej od teoretycznej zależy od dokładności obliczeń, sposobu ich wykonywania oraz kumulacji obciążających je błędów. Kumulację tę w dłuższym okresie ilustruje wykres 9.10.

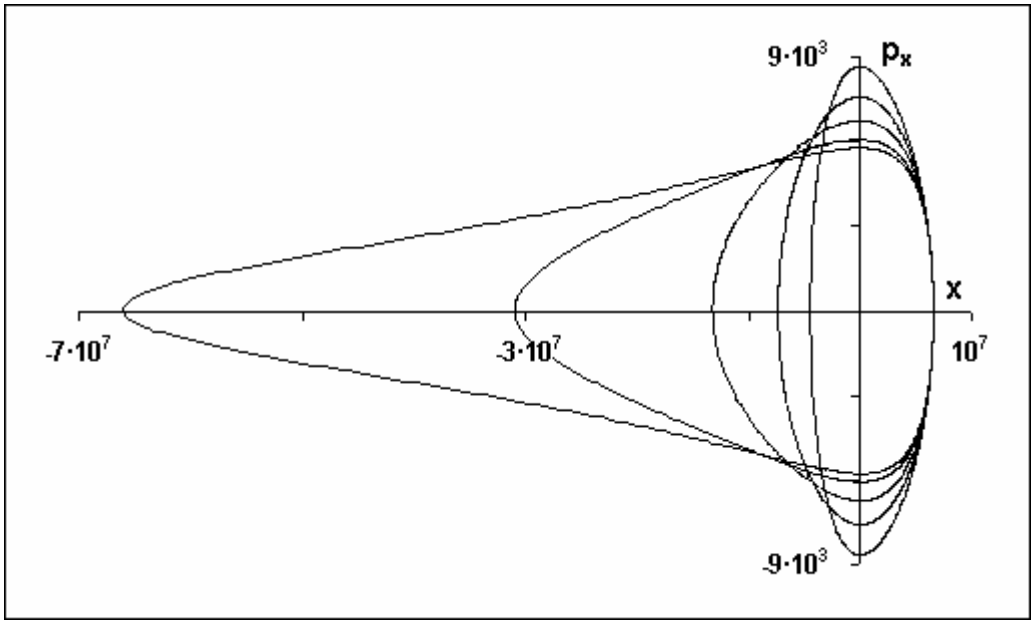
Obserwujemy tu liniowe narastanie błędu względnego w tempie $3.47 \cdot 10^{-16}$ na obieg (początkowe obiegi charakteryzujących się szybszą kumulacją). Oznacza to, że nawet po 10^9 obiegach błąd ten ciągle nie będzie przekraczał 10^{-6}. Zupełnie przeciwnie, niż to miało miejsce w dyskretyzacjach standardowych, dokładność obliczeń możemy zwiększyć zmniejszając liczbę kroków przypadających na jeden obieg orbity (wykres 9.11). Schemat (9.12) daje nam również dokładne trajektorie w przestrzeni fazowej (wykres 9.12).



Wykres 9.10. Maksymalne odchylenie od teorii zmodyfikowanego modelu (9.12) jako funkcja liczby obiegów *n*; δ = 0.006, N ≈ 500 kroków/obieg, p_0 = 10000, r_0 = $6.57 \cdot 10^6$, r_{tmax} = $13.232348 \cdot 10^6$.



Wykres 9.11. Dyskretyzacja zmodyfikowana. Względne odchylenie od teoretycznej orbity eliptycznej w funkcji φ, p_0 = 10000, r_0 = $6.57 \cdot 10^6$, r_{tmax} = $13.232348 \cdot 10^6$. Dokładność obliczeń – 19 ÷ 20 cyfr znaczących.



Wykres 9.12. Dokładne trajektorie w przestrzeni fazowej zmodyfikowanego modelu 9.12. δ = 0.006, N ≈ 525 kroków/obieg, $p_0 \in$ {7000, 8000, 9000, 10000, 10500}, $r_0 = 6.57 \cdot 10^6$.

9.5 Transformacja Kustaanheimo-Stiefela i zachowawcze dyskretyzacje ruchu keplerowskiego

9.5.1 Wprowadzenie

Dyskretyzacja omawiana w poprzednim rozdziale (patrz także [16]) i jej zmodyfikowana wersja są zaledwie rzędu pierwszego (co jest nieco zaskakujące biorąc pod uwagę dużą dokładność zmodyfikowanego schematu). W tym rozdziale dokonamy porównania tej dyskretyzacji z bardzo dobrym schematem numerycznym, który dla problemu Keplera niedawno zaproponował Kozlov [52]. Schemat ten zachowuje w ścisły sposób wszystkie trajektorie i całki ruchu. W komentarzu do tej pracy [18] zostało wskazane, że astronomowie znali już wcześniej dokładną dyskretyzację ruchu keplerowskiego [7, 8, 12, 70, 101], opartą na transformacji Kustaanheimo-Stiefela (KS) [53].

W pracy [18] zostało pokazane, iż dyskretyzację Kozlova można wyprowadzić w sposób dość elementarny. Podejście to jest zreferowane poniżej (z pewnymi uzupełnieniami). Zachowawcze dyskretyzacje 3-wymiarowego ruchu Keplera uzyskane w [52, 74] polegają na zastosowaniu metody punktu środowego (*midpoint rule*) (lub metody dyskretnego gradientu [54]) do izotropowego 4-wymiarowego równania oscylatora harmonicznego. Następnie transformacja Kustaanheimo-Stiefela [53] (w połączeniu z transformacją czasu

Levi-Civity) odwzorowuje 4-wymiarowy oscylator harmoniczny w 3-wymiarowy problem Keplera.

W celu szybkiego uzyskania wyniku Kozlova wystarczy zauważyć, że transformacja KS przez niego użyta redukuje 3-wymiarowy ruch keplerowski do 4 zwyczajnych liniowych równań różniczkowych ze stałymi współczynnikami (równań oscylatora harmonicznego), dla których istnieją jawne dokładne integratory numeryczne [83] (zobacz również [1, 17, 20, 63]).

9.5.2 Elementarna prezentacja transformacji KS

Transformacja Kustaanheimo-Stiefela (KS) jest zdefiniowana następująco

$$\vec{q} = \begin{pmatrix} 2Q_1Q_2 - 2Q_3Q_4 \\ 2Q_1Q_3 + 2Q_2Q_4 \\ Q_1^2 - Q_2^2 - Q_3^2 + Q_4^2 \end{pmatrix}, \tag{9.24}$$

$$\vec{p} = \frac{1}{2\left|\vec{Q}\right|^2} \begin{pmatrix} P_1Q_2 + P_2Q_1 - P_3Q_4 - P_4Q_3 \\ P_1Q_3 + P_2Q_4 + P_3Q_1 + P_4Q_2 \\ P_1Q_1 - P_2Q_2 - P_3Q_3 + P_4Q_4 \end{pmatrix}, \tag{9.25}$$

przy czym zmienne $(\vec{Q}, \vec{P})$ muszą spełniać równanie więzów

$$P_1Q_4 - P_2Q_3 + P_3Q_2 - P_4Q_1 = 0. \tag{9.26}$$

Oznaczmy przez $\mathcal{M}$ 7-wymiarową podprzestrzeń definiowaną przez (9.26). Transformacja KS odwzorowuje $\mathcal{M}$ w 6-wymiarową przestrzeń fazową $(\vec{q}, \vec{p})$. Bezpośrednim rachunkiem można sprawdzić użyteczne tożsamości:

$$\left|\vec{q}\right|^2 = \left|\vec{Q}\right|^4, \quad \left|\vec{p}\right|^2 = \frac{\left|\vec{P}\right|^2}{4\left|\vec{Q}\right|^2}. \tag{9.27}$$

Transformacja czasu Levi-Civity (lub Darboux-Sundmana, patrz [25, 39]) jest dana przez

$$\frac{dt}{ds} = \left|\vec{q}\right|. \tag{9.28}$$

W pracy [18] zostało wykazane następujące twierdzenie, które tu przytaczamy wraz z dowodem.

Twierdzenie. *Załóżmy, że* $(\vec{Q}, \vec{P}) \in \mathcal{M}$ *spełniają 4-wymiarowe równania izotropowego oscylatora harmonicznego*

$$\frac{d\vec{Q}}{ds} = \frac{1}{4}\vec{P}, \quad \frac{d\vec{P}}{ds} = 2E\vec{Q}, \tag{9.29}$$

gdzie E = const, czas t jest dany przez (9.28) *i* $(\vec{q}, \vec{p})$ *są zdefiniowane przez* (9.24), (9.25). *Wówczas*

$$\frac{d\vec{q}}{dt} = \vec{p}, \qquad \frac{d\vec{p}}{dt} = -\frac{k\vec{q}}{|\vec{q}|^3}, \quad \vec{q},\ \vec{p} \in \mathbf{R}^3, \tag{9.30}$$

przy czym k = const. *Ponadto*

$$\frac{1}{2}\vec{p}^{\,2} - \frac{k}{|\vec{q}|} = E, \quad \frac{1}{8}\left|\vec{P}\right|^2 - E\left|\vec{Q}\right|^2 = k. \tag{9.31}$$

Dowód: Sprawdzamy, że więzy (9.26) są zachowane przez (9.29), tj. układ (9.29) może być ograniczony do podprzestrzeni $\mathcal{M}$. Stosując wszystkie założenia otrzymujemy (po prostych, lecz żmudnych rachunkach)

$$\frac{d\vec{q}}{dt} = \vec{p}, \qquad \frac{d\vec{p}}{dt} = \frac{\vec{q}}{|\vec{q}|^2}\left(E - \frac{1}{2}\vec{p}^{\,2}\right). \tag{9.32}$$

Można łatwo sprawdzić, że układ (9.32) spełnia zasadę zachowania w postaci

$$|\vec{q}|\left(\frac{1}{2}\vec{p}^{\,2} - E\right) = k, \quad k = const \tag{9.33}$$

równoważną pierwszemu z równań (9.31). Istotnie, różniczkując lewą stronę (9.33) i stosując (9.32) otrzymujemy zero. Następnie używając (9.33) eliminujemy *E* z (9.32) w celu uzyskania drugiego z równań (9.30). Ostatecznie, podstawiając (9.27) do (9.33) dostajemy drugie z równań (9.31). Ñ

9.5.3 Dokładna dyskretyzacja równań oscylatora harmonicznego

Układ równań (9.29) równoważny równaniu 4-wymiarowego oscylatora harmonicznego, dopuszcza dokładną dyskretyzację (patrz [17, 18]):

$$\begin{aligned} \frac{\vec{Q}_{j+1} - \vec{Q}_j}{\delta(h_j)} &= \frac{1}{4}\frac{\vec{P}_{j+1} + \vec{P}_j}{2}, \\ \frac{\vec{P}_{j+1} - \vec{P}_j}{\delta(h_j)} &= 2E\frac{\vec{Q}_{j+1} + \vec{Q}_j}{2}, \end{aligned} \tag{9.34}$$

gdzie $h := s_{j+1} - s_j$ jest krokiem czasowym zmiennej *s*, $\vec{Q}_j, \vec{P}_j$ oznaczają *j*-tą iterację schematu numerycznego (nie należy mylić ze współrzędnymi Q_j, P_j) oraz

$$\delta(h_j) = \frac{2}{\omega}\tan\frac{\omega h_j}{2}, \quad \omega^2 = -\frac{1}{2}E. \tag{9.35}$$

W przypadku stałego kroku $h_j = h$, i $E < 0$, rozpoznajemy tu dokładny integrator znaleziony przez Kozlova (patrz [18, 52]). Przypadek hiperboliczny

i paraboliczny ([18], wzory (4.16) i (4.18)) wynikają bezpośrednio po przyjęciu urojonego ω (tj. $E > 0$) lub $\omega = 0$. Dokładny schemat numeryczny (9.34) zachowuje całkę energii:

$$\frac{1}{8}\left|\vec{P}_j\right|^2 - E\left|\vec{Q}_j\right|^2 = k. \tag{9.36}$$

Przedstawimy jeszcze inną, (równoważną schematowi (9.34), lecz zapisaną w sposób jawny) postać dokładnej dyskretyzacji Kozlova:

$$\begin{aligned} \vec{Q}_{j+1} &= \cos\omega h_j \vec{Q}_j + \frac{\sin\omega h_j}{4\omega}\vec{P}_j, \\ \vec{P}_{j+1} &= -4\omega\sin\omega h_j \vec{Q}_j + \cos\omega h_j \vec{P}_j. \end{aligned} \tag{9.37}$$

Układ ten jest bezpośrednią konsekwencją obliczenia dokładnego rozwiązania (9.29) w punkcie $s = s_j$ i $s = s_j + h$ (porównaj [17, 18]). Przypomnijmy w tym miejscu, że dokładna dyskretyzacja równań oscylatora harmonicznego posłużyła do skonstruowania nowych geometrycznych schematów numerycznych wysokiej dokładności (opisanych w rozdziałach 5 i 7).

9.5.4 Dokładna dyskretyzacja czasu

Równanie Levi-Civity (9.28) może być rozwiązane dokładnie na różne sposoby, porównaj [12, 52, 70]. Tu referujemy metodę zaproponowaną w pracy [18], redukującą ten problem do liniowych zwyczajnych równań różniczkowych ze stałymi współczynnikami.

Twierdzenie. *Jeżeli* $\vec{Q}$, $\vec{P}$ *spełniają* (9.29) *i t spełnia* (9.28), *wówczas*

$$\frac{d\vec{w}}{ds} = \Omega, \quad \vec{w} = \begin{pmatrix} \left|\vec{Q}\right|^2 \\ \left|\vec{P}\right|^2 \\ \vec{Q}\cdot\vec{P} \\ t \end{pmatrix}, \quad \Omega = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 4E & 0 \\ 2E & \frac{1}{4} & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{9.38}$$

Dowód: Z (9.24) obliczamy $|\vec{q}|^2 = \left|\vec{Q}\right|^2$. Stąd $dt/ds = \left|\vec{Q}\right|^2$. Inne równania wynikają bezpośrednio z (9.29), na przykład

$$\frac{d}{ds}\vec{Q}\cdot\vec{P} = \frac{1}{4}\vec{P}\cdot\vec{P} + \vec{Q}\cdot(2E\vec{Q}) = 2E\left|\vec{Q}\right|^2 + \frac{1}{4}\left|\vec{P}\right|^2. \tag{9.39}$$

Ñ

Dalej postępujemy w sposób standardowy, wykorzystując znane procedury dla równań różniczkowych zwyczajnych o stałych współczynnikach. Ogólne

rozwiązanie jest dane przez $\vec{w}(s) = \exp(s\Omega)\vec{w}(0)$, a zatem dokładna dyskretyzacja $\vec{w}_n = \vec{w}(hn)$ jest dana wzorem

$$\vec{w}_{n+1} = \exp(h\Omega)\vec{w}_n, \tag{9.40}$$

i problem redukuje się do dobrze znanej, czysto algebraicznej procedury obliczania $e^{\Omega h}$. W naszym szczególnym przypadku możemy łatwo sprawdzić, że dla Ω danej wzorem (9.38) mamy $\Omega^4 = 2E\Omega^2$. Stąd

$$\Omega = \sqrt{-2E}M, \qquad M^4 = -M^2. \tag{9.41}$$

Przy tych założeniach można obliczyć, że [18]:

$$e^{\nu M} = I + \nu M + (1-\cos\nu)M^2 + (\nu - \sin\nu)M^3. \tag{9.42}$$

W szczególności, z równania (9.40) wynika, iż [18]:

$$t_{j+1} = t_j + \frac{\sin\omega h}{4\omega}\left(|\vec{Q}_j|^2 - \frac{|\vec{P}_j|^2}{16\omega^2}\right) + \frac{h}{2}\left(|\vec{Q}_j|^2 + \frac{|\vec{P}_j|^2}{16\omega^2}\right) + \frac{\vec{Q}_j \cdot \vec{P}_j \sin^2\omega h}{4\omega^2}. \tag{9.43}$$

Ostatecznie, eliminując $|\vec{P}_j|^2$ przy pomocy wzoru (9.36), otrzymujemy

$$t_{j+1} = t_j + \frac{hk}{4\omega^2}\left(1 - \frac{\sin 2\omega h}{2\omega h}\right) + \frac{\sin 2\omega h}{2\omega}|\vec{Q}_j|^2 + \frac{\vec{Q}_j \cdot \vec{P}_j \sin^2\omega h}{4\omega^2}. \tag{9.44}$$

W pracach [12, 52, 70] sposoby dojścia do dyskretyzacji czasu były różne, ale końcowy wynik jest ten sam [18].

9.5.5 Korzyści płynące z notacji zespolonej

Godny uwagi jest ścisły związek pomiędzy słynną wiązką Hopfa a transformacją Kustaanheimo-Stiefela. Otóż wzór (9.24) jest identyczny z odwzorowaniem Hopfa (porównaj [107], Dodatek A):

$$z \mapsto (2\,\mathrm{Re}\,\bar{z}_1 z_2,\ 2\,\mathrm{Im}\,\bar{z}_1 z_2,\ |z_1|^2 - |z_2|^2), \tag{9.45}$$

jeśli przyjmiemy, że

$$z_1 = Q_1 - iQ_4, \quad z_2 = Q_2 + iQ_3. \tag{9.46}$$

W artykułach na temat geometrycznego całkowania problemu Keplera rzadko przywołuje się odwzorowanie Hopfa (wyjątkami są, np. [2, 108]). Co ciekawe, artykuł [107], którego celem było ukazanie różnorodnych zastosowań wiązki Hopfa w fizyce teoretycznej, również nie zawiera żadnej wzmianki dotyczącej mapowania Kustaanheimo-Stiefela.

Transformacja Kustaanheimo-Stiefela była dyskutowana w ramach rozmaitych podejść algebraicznych i geometrycznych włączając w to

kwaterniony [2, 25, 83, 108, 109, 110]. Tu ograniczymy się do użycia zmiennych zespolonych i zapisania analogicznej do (9.46) formuły dla $\vec{P}$:

$$w_1 = P_1 - iP_4, \quad w_2 = P_2 + iP_3. \tag{9.47}$$

Teraz wzory (9.24), (9.25) i (9.26) mogą być zapisane odpowiednio w postaci

$$q_1 + iq_2 = 2\bar{z}_1 z_2, \quad q_3 = |z_1|^2 - |z_2|^2, \tag{9.48}$$

$$p_1 + ip_2 = \frac{\overline{w}_1 z_2 + \bar{z}_1 w_2}{2\left(|z_1|^2 + |z_2|^2\right)}, \quad p_3 = \frac{w_1 \bar{z}_1 - \overline{w}_2 z_2}{2\left(|z_1|^2 + |z_2|^2\right)}, \tag{9.49}$$

$$\mathrm{Im}(w_1 \bar{z}_1 - \overline{w}_2 z_2) = 0. \tag{9.50}$$

Notacja zespolona ma wiele zalet. Po pierwsze, więzy (9.26) mają naturalną interpretację: p_3 musi być rzeczywiste. Po drugie, bezpośrednio widać, że transformacja

$$(z_1, z_2, w_1, w_2) \to e^{i\vartheta}(z_1, z_2, w_1, w_2), \quad (\vartheta \in \boldsymbol{R}), \tag{9.51}$$

jest symetrią równań (9.48), (9.49), (9.50). Innymi słowy, transformacja KS posiada jednowymiarowe jądro, którym jest okrąg parametryzowany przez ϑ.

9.5.6 Odwrotna transformacja KS

Do numerycznej symulacji ruchu Keplera przy wykorzystaniu transformacji Kustaanheimo-Stiefela potrzebna jest podana w sposób jawny transformacja odwrotna. Artykuł [52] jej nie zawiera, ale wyniki te generalnie są znane (patrz, np. [12, 70]). Dla kompletności, podamy tu jednak swoją wersję wyprowadzenia tych wzorów. Eliminując z_2 z (9.48) otrzymujemy:

$$|z_1|^4 - |z_1|^2 q_3 - \frac{1}{4}(q_1^2 + q_2^2) = 0, \tag{9.52}$$

a to równanie ma dokładnie jedno dodatnie rozwiązanie (zakładając, że $q_3 \neq -|\vec{q}|$):

$$|z_1| = \sqrt{\frac{q_3 + |\vec{q}|}{2}}, \tag{9.53}$$

(porównaj z [110]). Oznaczając fazę $\bar{z}_1$ przez α dostajemy

$$z_2 = \frac{q_1 + iq_2}{2\bar{z}_1} = \frac{(q_1 + iq_2)e^{-i\alpha}}{\sqrt{2(q_3 + |\vec{q}|)}}. \tag{9.54}$$

Zatem odwrotna transformacja dla $\vec{q}$ leżąca na zewnątrz ujemnej półosi q_3 (tj. $q_1 = q_2 = 0$, $q_3 \leq 0$) ma postać

$$\begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \end{pmatrix} = \frac{1}{\sqrt{2\left(q_3 + |\vec{q}|\right)}} \begin{pmatrix} \left(q_3 + |\vec{q}|\right)\cos\alpha \\ q_1\cos\alpha - q_2\sin\alpha \\ q_2\cos\alpha + q_1\sin\alpha \\ \left(q_3 + |\vec{q}|\right)\sin\alpha \end{pmatrix}. \tag{9.55}$$

W celu uzyskania bardziej symetrycznej postaci możemy przeprowadzić te same obliczenia dla z_2 eliminując z_1 z (9.48). Zakładając, że $q_3 \neq |\vec{q}|$ otrzymujemy:

$$|z_2| = \sqrt{\frac{|\vec{q}| - q_3}{2}} \tag{9.56}$$

Mamy zatem

$$z_1 = \sqrt{\frac{q_3 + |\vec{q}|}{2}} e^{-\alpha}, \qquad z_2 = \sqrt{\frac{|\vec{q}| - q_3}{2}} e^{\beta}, \tag{9.57}$$

przy czym, ze względu na (9.52), $\alpha + \beta = \varphi$ (φ jest fazą $q_1 + iq_2$).

Równania (9.25), (9.26) można zapisać w postaci macierzowej:

$$\frac{1}{2|\vec{Q}|^2} \begin{pmatrix} Q_2 & Q_1 & -Q_4 & -Q_3 \\ Q_3 & Q_4 & Q_1 & Q_2 \\ Q_1 & -Q_2 & -Q_3 & Q_4 \\ Q_4 & -Q_3 & Q_2 & -Q_1 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ 0 \end{pmatrix}, \tag{9.58}$$

która pozwala na łatwe otrzymanie transformacji odwrotnej (o ile znana jest odwrotna transformacja dla $\vec{q}$). Sformułujemy to w postaci twierdzenia, w którym udało się wszystkie podprzypadki ująć jednym zwięzłym wzorem.

Twierdzenie. *Odwrotna transformacja Kustaanheimo-Stiefela, parametryzowana kątem α, jest w sposób jawny zadana wzorami*

$$\vec{Q} = \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{|\vec{q}| + q_3}\cos\alpha \\ \sqrt{|\vec{q}| - q_3}\cos\beta \\ \sqrt{|\vec{q}| - q_3}\sin\beta \\ \sqrt{|\vec{q}| + q_3}\sin\alpha \end{pmatrix}, \quad \beta = \varphi - \alpha, \tag{9.59}$$

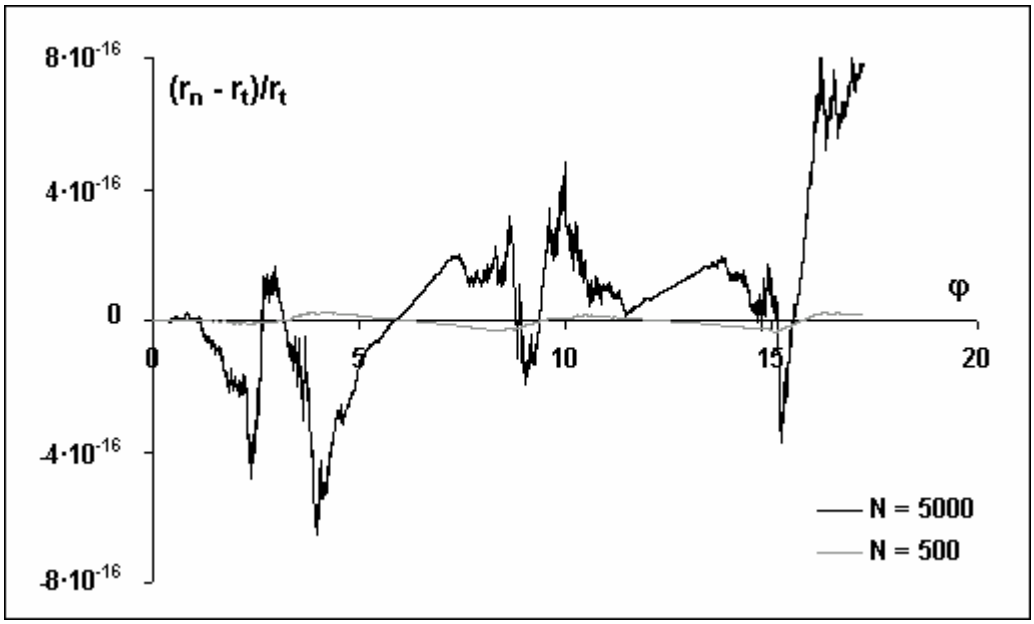
$$\vec{P} = \begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix} = 2 \begin{pmatrix} p_2Q_3 + p_1Q_2 + p_3Q_1 \\ p_2Q_4 + p_1Q_1 - p_3Q_2 \\ p_2Q_1 - p_1Q_4 - p_3Q_3 \\ p_2Q_2 - p_1Q_3 + p_3Q_4 \end{pmatrix}, \tag{9.60}$$

przy czym φ zdefiniowane jest przez $e^{i\varphi} = \dfrac{q_1 + iq_2}{\sqrt{q_1^2 + q_2^2}}$.

Łatwo sprawdzić, że równania (9.24), (9.25), (9.26) są tożsamościowo spełnione, jeżeli je podstawimy do (9.59), (9.60). Zauważmy, że transformacja KS jest zdefiniowana dla $\vec{Q} \neq 0$, a jej transformacja odwrotna dla $\vec{q} \neq 0$. Wzór (9.55) jest przypadkiem szczególnym (9.59) słusznym tylko, gdy $\vec{q} \neq (0, 0, -|\vec{q}|)$.

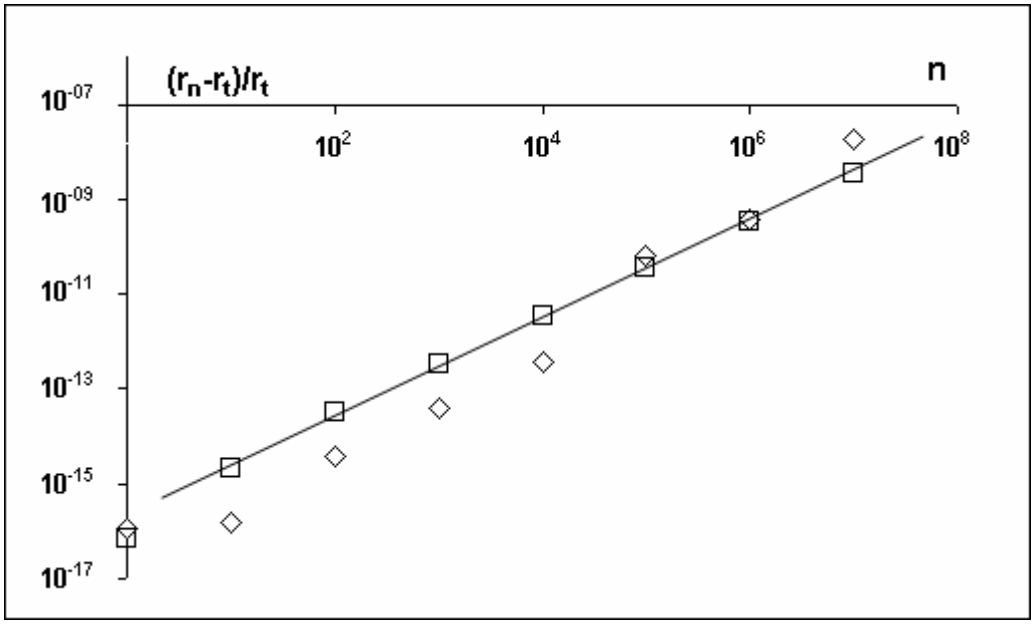
9.5.7 Testy numeryczne

Dokładność dyskretyzacji Kozlova ograniczają wyłącznie błędy numeryczne, przy czym ich wpływ można redukować zmniejszając liczbę kroków przypadających na jeden obieg orbity (analogicznie jak w zmodyfikowanej dyskretyzacji podanej w pracy [16]). Ilustruje to wykres 9.13.



Wykres 9.13. Dyskretyzacja Kozlova. Względne odchylenie od teoretycznej orbity eliptycznej w funkcji φ (parametrem jest N - liczba kroków na obieg). p_0 = 10000, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$.

Postrzępiona krzywa na wykresie 9.13 powstaje w wyniku kumulowania się przypadkowych błędów numerycznych. Dokładność pierwszych kilkudziesięciu kroków utrzymuje się na poziomie nie gorszym niż 10^{-18}, później nakładanie się błędów prowadzi do znacznie większych odchyleń. Po 3 obiegach (dla $N = 5000$) maksymalny błąd wzrasta z 10^{-20} do niemal 10^{-15}. Zwróćmy uwagę, że analogiczne krzywe (wykresy 9.8, 9.9, 9.11) dla zmodyfikowanej dyskretyzacji podanej w pracy [16] miały znacznie spokojniejszy przebieg. Kumulację błędów występującą w dyskretyzacji Kozlova w dłuższym okresie czasu w porównaniu do zmodyfikowanego modelu prezentuje wykres 9.14.



Wykres 9.14. Maksymalne odchylenie od teorii dyskretyzacji Kozlova vs zmodyfikowany model (9.12) jako funkcja liczby obiegów orbity *n*; N ≈ 500 kroków/obieg, $p_0 = 10000$, $r_0 = 6.57 \cdot 10^6$, $r_{tmax} = 13.232348 \cdot 10^6$.

W przypadku zmodyfikowanej dyskretyzacji 9.12 mamy ściśle liniową kumulację błędów numerycznych, kumulacja w dyskretyzacji Kozlova jest tylko z grubsza liniowa. Nachylenie funkcji zmienia się w dość szerokich granicach, a po przekroczeniu $n = 10^4$ obiegów tempo narastania błędów wynosi ok. $5.5 \cdot 10^{-16}$, jest wyraźnie większe niż w konkurencyjnym modelu ($3.47 \cdot 10^{-16}$). Powoduje to, że powyżej $n = 10^5$ obiegu schemat Kozlova traci swą początkową przewagę nad zmodyfikowanym integratorem 9.12.

9.6 Podsumowanie

W rozdziale tym omówiono i przetestowano szereg dyskretyzacji klasycznego problemu Keplera. Znalazły się wśród nich trzy metody standardowe (Eulera, leap-frog i dyskretny gradient) oraz dwie znalezione ostatnio dyskretyzacje zachowawcze [16, 52]. Dyskretyzację podaną w pracy [16], która zachowywała analogi całek ruchu ciągłego problemu Keplera udało się zmodyfikować w taki sposób, aby je zachowywała dokładnie.

Testy numeryczne koncentrowały się na kumulacji błędów badanych dyskretyzacji wyrażonej przez odchylenie od orbity teoretycznej wraz z upływem czasu. Sprawdzono też ewolucję trajektorii w przestrzeni fazowej.

Wśród dyskretyzacji standardowych potwierdziły się zalety cenionej przez astronomów metody Störmera-Verleta (*leap-frog*), która najlepiej spośród nich zachowuje geometrię trajektorii fazowych.

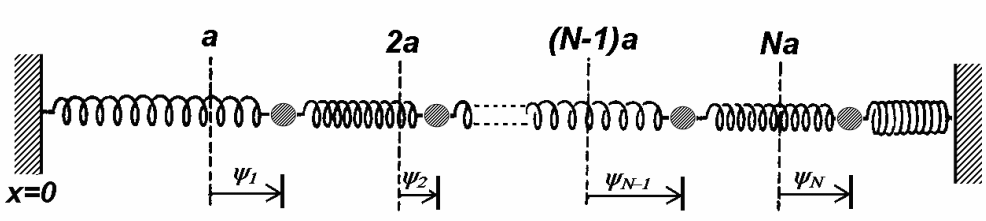
Potwierdziliśmy doświadczalnie, że odchylenia od orbity teoretycznej zmodyfikowanej dyskretyzacji (9.12) i dyskretyzacji Kozlova wynikają jedynie z numerycznych błędów zaokrągleń. Badania kumulacji błędów pokazały, że zmodyfikowana dyskretyzacja (9.12) daje ściśle liniowy przyrost błędów w czasie i przewyższa pod tym względem schemat Kozlova. Pokazano też na przykładach paradoksalną cechę dyskretyzacji dokładnych polegającą na wzroście dokładności przy zwiększeniu kroku czasowego.

10 Modele dyskretne jednowymiarowego równania falowego

W rozdziale tym poddajemy analizie możliwości badania różnych aspektów ruchu falowego przy pomocy zjawisk zachodzących w modelu zbudowanym z jednorodnej sieci sprzężonych oscylatorów harmonicznych. Pokazany zostanie też związek równania falowego z równaniem oscylatora harmonicznego (poprzez transformatę Fouriera), co prowadzi do otrzymywania ewolucji czasowej ośrodka z pomocą dyskretyzacji dokładnej równania oscylatora harmonicznego. W ostatnim podrozdziale znalazły się niektóre wyniki eksperymentów komputerowych ilustrujących działanie modelu sieci oscylatorów i dyskretnych schematów numerycznych dla rozwiązywania równania falowego.

10.1 Sprzężone oscylatory harmoniczne jako prosty model ruchu falowego

Fizycznym obrazem ośrodka dyskretnego dla ruchu falowego może być jednowymiarowa siatka o stałej sieci a, w węzłach której umieszczono N jednakowych ciężarków o masie m połączonych N+1 sprężynami o współczynniku sprężystości K i długości swobodnej a, jak to pokazano na rysunku 10.1. W razie potrzeby można go zmodyfikować usuwając pierwszą lub ostatnią sprężynę w celu zmiany warunków brzegowych na krańcach ośrodka.



Rysunek 10.1. Model dyskretny ośrodka do badania fal podłużnych

Jeśli wskaźnika i użyjemy do ponumerowania drgających mas, a przy pomocy ψ_i oznaczymy wychylenie i-tej masy z położenia równowagi (danego przez $x_i = i \cdot a$), to równania ruchu takiego układu wyglądają następująco:

$$
\begin{aligned}
m\ddot{\psi}_1 &= -K\psi_1 + K(\psi_2 - \psi_1) \\
m\ddot{\psi}_i &= -K(\psi_i - \psi_{i-1}) + K(\psi_{i+1} - \psi_i) \quad 1 < i < N \\
m\ddot{\psi}_N &= -K(\psi_N - \psi_{N-1}) - K\psi_N
\end{aligned}
\tag{10.1}
$$

Zastosowanie symetrycznej metody Eulera (rozdział 3) prowadzi do otrzymania naturalnej dyskretyzacji tego układu równań o następującej postaci (wskaźnikiem *n* posłużono się w celu rozróżnienia różnych chwil czasu, a ε oznacza krok czasowy):

$$
\begin{aligned}
\psi_{1,n+1} - 2\psi_{1,n} + \psi_{1,n-1} &= \frac{K\varepsilon^2}{m}(\psi_{2,n} - 2\psi_{1,n}), \\
\psi_{i,n+1} - 2\psi_{i,n} + \psi_{i,n-1} &= \frac{K\varepsilon^2}{m}(\psi_{i+1,n} - 2\psi_{i,n} + \psi_{i-1,n}), \\
\psi_{N,n+1} - 2\psi_{N,n} + \psi_{N,n-1} &= \frac{K\varepsilon^2}{m}(-2\psi_{N,n} + \psi_{N-1,n}).
\end{aligned}
\tag{10.2}
$$

Innymi słowy, możemy rozważać nieskończoną sieć

$$
\psi_{i,n+1} - 2\psi_{i,n} + \psi_{i,n-1} = \frac{K\varepsilon^2}{m}(\psi_{i+1,n} - 2\psi_{i,n} + \psi_{i-1,n}),
\tag{10.3}
$$

z warunkami brzegowymi

$$
\psi_{0,n} = \psi_{N+1,n} = 0.
\tag{10.4}
$$

Jest rzeczą interesującą, że zmiana kroku czasowego ε jest równoważna zmianie częstości oscylatora $\omega_0 \equiv \sqrt{K/m}$. Wyjaśnienie jest dość oczywiste: okres drgań oscylatora jest jednostką uniwersalnej skali czasowej problemu.

10.2 Związki dyspersyjne dla nieskończonego ośrodka dyskretnego

Sprawdźmy, czy nieskończony układ (10.3) (bez warunków brzegowych!) dopuszcza rozwiązanie w postaci fali harmonicznej. Załóżmy, że

$$
\psi_{i,n} = \cos(aki - \omega n\varepsilon) \equiv \cos\phi_{i,n}.
\tag{10.5}
$$

Wówczas oczywiście

$$
\begin{aligned}
\psi_{i\pm1,n} &= \cos\phi_{i,n}\cos ka \mp \sin\phi_{i,n}\sin ka, \\
\psi_{i,n\pm1} &= \cos\phi_{i,n}\cos\omega\varepsilon \pm \sin\phi_{i,n}\sin\omega\varepsilon.
\end{aligned}
\tag{10.6}
$$

Podstawiając te wyrażenia do (10.3) dostajemy

$$
2\cos\phi_{i,n}(\cos\omega\varepsilon - 1) = 2\cos\phi_{i,n}\frac{K\varepsilon^2}{m}(\cos ka - 1),
\tag{10.7}
$$

co oznacza, że $m\sin^2(\omega\varepsilon/2) = K\varepsilon^2\sin^2(ka/2)$, czyli

$$\sin\frac{\omega\varepsilon}{2} = \varepsilon\sqrt{\frac{K}{m}}\sin\frac{ka}{2}. \tag{10.8}$$

Taki warunek muszą spełniać fale harmoniczne rozchodzące się w układzie (10.3). Jest to zarazem związek dyspersyjny dla fal biegnących w tym ośrodku. Wykonując przejście graniczne $\varepsilon \to 0$ dostajemy związek dyspersyjny w przypadku *t* ciągłego:

$$\omega = 2\sqrt{\frac{K}{m}}\sin\frac{ka}{2}, \tag{10.9}$$

a przechodząc dodatkowo z $ka \to 0$ otrzymujemy zależność $\omega(k)$ w sytuacji, gdy sam ośrodek też jest ciągły:

$$\omega = \sqrt{\frac{K}{m}}ka. \tag{10.10}$$

Podstawiając do (10.8) wartości *ka* leżące w przedziale od *ka* = 0 do *ka* = π, uzyskujemy częstości leżące w przedziale od $\omega_0 = 0$ do

$$\omega_{\max} = \frac{2}{\varepsilon}\arcsin\left(\varepsilon\sqrt{\frac{K}{m}}\right). \tag{10.11}$$

Fale o częstościach z tego zakresu wnikają do ośrodka dowolnie daleko.

Sprawdźmy teraz, czy do nieskończonego ośrodka (10.3) może wniknąć fala eksponencjalna postaci

$$\psi_{i,n} = A_i\cos(\omega n\varepsilon) = Ae^{-\kappa ia}\cos(\omega n\varepsilon). \tag{10.12}$$

Wówczas

$$\begin{aligned}\psi_{i\pm1,n} &= Ae^{-\kappa(i\pm1)a}\cos(\omega n\varepsilon),\\ \psi_{i,n\pm1} &= Ae^{-\kappa ia}\cos(\omega n\varepsilon \pm \omega\varepsilon).\end{aligned} \tag{10.13}$$

Podstawiając te wyrażenia do (10.3) otrzymujemy

$$\cos\omega\varepsilon - 1 = \frac{K\varepsilon^2}{m}(\cosh\kappa a - 1), \tag{10.14}$$

co prowadzi do równania $-m\sin^2(\omega\varepsilon/2) = K\varepsilon^2\sinh^2(\kappa a/2)$, którego jedynym rozwiązaniem jest $\omega = 0$ dla $\kappa = 0$. Stwierdzamy zatem, że w układzie (10.3) nie mogą rozchodzić się fale eksponencjalne typu (10.12). Jest to zrozumiałe, gdyż ośrodek nasz posiada częstość progową $\omega_0 = 0$ i brakuje w nim dolnego obszaru reaktywnego.

Jeśli jednak wzbudzenie przekroczy częstość ω_{max}, może to skutkować wejściem w zakres reaktywny ośrodka i pojawieniem się w nim przemiennej fali eksponencjalnej postaci

$$\psi_{i,n} = A_i\cos(\omega n\varepsilon) = (-1)^i Ae^{-\kappa ia}\cos(\omega n\varepsilon). \tag{10.15}$$

Sprawdźmy to. Mamy

$$\psi_{i\pm1,n} = (-1)^{i\pm1} A e^{-\kappa(i\pm1)a} \cos(\omega n\varepsilon),$$
$$\psi_{i,n\pm1} = (-1)^{i} A e^{-\kappa i a} \cos(\omega n\varepsilon \pm \omega\varepsilon). \tag{10.16}$$

Po podstawieniu powyższych zależności do (10.3) dostajemy

$$\cos\omega\varepsilon - 1 = -\frac{K\varepsilon^2}{m}(\cosh\kappa a + 1), \tag{10.17}$$

skąd wynika, że $m\sin^2(\omega\varepsilon/2) = K\varepsilon^2\cosh^2(\kappa a/2)$, czyli

$$\sin\frac{\omega\varepsilon}{2} = \varepsilon\sqrt{\frac{K}{m}}\cosh\frac{\kappa a}{2}. \tag{10.18}$$

Jest to związek dyspersyjny dla zanikającej przemiennej fali wykładniczej o częstości $\omega > \omega_{max}$, która może wniknąć do układu (10.3) na pewną głębokość.

10.3 Związki dyspersyjne dla skończonego ośrodka dyskretnego

Ustalimy teraz związki dyspersyjne w przypadku skończonego (N punktów) ośrodka zadanego równaniami (przypadek ciągłego czasu).

$$\ddot{\psi}_1 = \frac{K}{m}(\psi_{2,n} - 2\psi_{1,n}),$$
$$\ddot{\psi}_i = \frac{K}{m}(\psi_{i+1,n} - 2\psi_{i,n} + \psi_{i-1,n}), \tag{10.19}$$
$$\ddot{\psi}_N = \frac{K}{m}(-2\psi_{N,n} + \psi_{N-1,n}),$$

które zapiszemy w formie macierzowej:

$$\frac{d^2}{dt^2}\vec{\psi} = \frac{K}{m}\hat{A}\vec{\psi}, \tag{10.20}$$

gdzie $\hat{A}$ jest macierzą $N \times N$ postaci

$$\hat{A} = \begin{pmatrix} -2 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -2 & 1 & 0 \\ 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 1 & -2 \end{pmatrix}, \qquad \vec{\psi} = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \psi_N \end{pmatrix} \tag{10.21}$$

Poszukajmy rozwiązań w postaci fali harmonicznej $\vec{\psi} = e^{i\omega t}\vec{c}$, gdzie ω i $\vec{c}$ są stałymi. Wówczas

$$\hat{A}\vec{c} = -\lambda\vec{c}, \quad \lambda := \frac{m\omega^2}{K}. \tag{10.22}$$

Rozwiązanie $\vec{c} \neq 0$ istnieje, jeśli $\det(\hat{A} + \lambda I) = 0$. Oznaczmy

$$W_N := \det(\hat{A} + \lambda I). \tag{10.23}$$

N jest ustalone. Naszym celem jest znalezienie λ takiej, że W_N = 0. Aby wyznaczyć W_N użyjemy rozwinięcia Laplace'a, które daje następujący związek rekurencyjny:

$$W_N = (\lambda - 2)W_{N-1} - W_{N-2}. \tag{10.24}$$

To równanie różnicowe może być łatwo rozwiązane:

$$W_N = a_1 r_1^N + a_2 r_2^N, \tag{10.25}$$

gdzie r_1, r_2 są pierwiastkami równania charakterystycznego $r^2 - (\lambda - 2)r + 1 = 0$, a a_1, a_2 są stałymi, które mogą być wyrażone przez W_1 i W_2. Ostatecznie dostajemy

$$W_N = \frac{r_1^{N+1} - r_2^{N+1}}{r_1 - r_2} \tag{10.26}$$

dla $r_1 \neq r_2$ oraz $W_N = (N+1)r_1^N$ w przypadku $r_1 = r_2$. Jeśli r_1, r_2 są rzeczywiste, wówczas $W_N \neq 0$.

Wynika stąd, że r_1, r_2 muszą być zespolone i takie, że $r_1 r_2 = 1$ ($r_2 = \bar{r}_1$) oraz $r_1 + r_2 = 2 - \lambda$. Zatem

$$r_1 = e^{i\alpha}, \quad r_1 = e^{-i\alpha}, \quad \lambda = 2(1 + \cos\alpha), \quad W_N = \frac{\sin(N+1)\alpha}{\sin\alpha}. \tag{10.27}$$

Warunek $W_N = 0$ oznacza, że $(N + 1)\alpha = j\pi$ (j – całkowite). Otrzymujemy w ten sposób ciąg dyskretny

$$\alpha_j = \frac{j\pi}{N+1}, \tag{10.28}$$

$$\lambda_j = 2\left(1 + \cos\frac{j\pi}{N+1}\right) = 4\cos^2\frac{j\pi}{2(N+1)} = 4\sin^2\frac{j'\pi}{2(N+1)}, \tag{10.29}$$

gdzie j = 1, 2,…, N i $j' = N + 1 - j$ (zatem również j' = 1, 2,…, N). Uwzględniając (10.22) otrzymujemy

$$\omega = 2\sqrt{\frac{K}{m}}\cos\frac{j\pi}{2(N+1)} = 2\sqrt{\frac{K}{m}}\sin\frac{j'\pi}{2(N+1)}. \tag{10.30}$$

W przypadku fal stojących w skończonym ośrodku o długości $L = (N + 1)a$ musi zachodzić $\frac{l\pi}{2(N+1)} = \frac{ka}{2}$ (k – liczba falowa, l – całkowite). Uwzględniając to otrzymujemy zależność taką samą, jak w przypadku fal biegnących w ośrodku dyskretnym nieskończonym (wzór (10.9)). Pokazaliśmy zatem, że związki dyspersyjne dla fal biegnących w ośrodku nieskończonym i dla fal stojących w takim samym ośrodku skończonym są identyczne. Zwróćmy też uwagę na to, że w przypadku fal stojących w ośrodku skończonym istnieje minimalna wartość ka wynikająca ze związku $\frac{\pi}{2(N+1)} = \frac{ka}{2}$. Zatem powinna się tu pojawić częstość progowa równa (porównaj wzór (10.8))

$$\omega_0 = \frac{2}{\varepsilon}\arcsin\left(\varepsilon\sqrt{\frac{K}{m}}\sin\left(\frac{\pi}{2(N+1)}\right)\right) \tag{10.31}$$

i reaktywność takiego ośrodka dla drgań wymuszonych o częstości niższej od ω_0.

Wróćmy jeszcze do wektora własnego $\vec{c}$. Jego składowe (związane z wartością własną λ_k) spełniają równanie różnicowe

$$c_j + 2c_{j-1}\cos\alpha_k + c_{j-2} = 0, \quad c_0 = c_{N+1} = 0. \tag{10.32}$$

Rozwiązanie jest praktycznie identyczne jak w przypadku równania (10.24). Otrzymujemy

$$c_j = C\sin j(\pi - \alpha_k) = C\sin\frac{\pi jk'}{N+1}, \tag{10.33}$$

gdzie C jest dowolną stałą. Aby porównać to rozwiązanie z przypadkiem ciągłym zauważmy, że

$$x_j = ja, \;\; L = (N+1)a. \tag{10.34}$$

Zatem

$$c_j = C\sin\frac{k'x_j}{L} \tag{10.35}$$

i oczywiście $c_j \approx \psi(x_j, t_k)$.

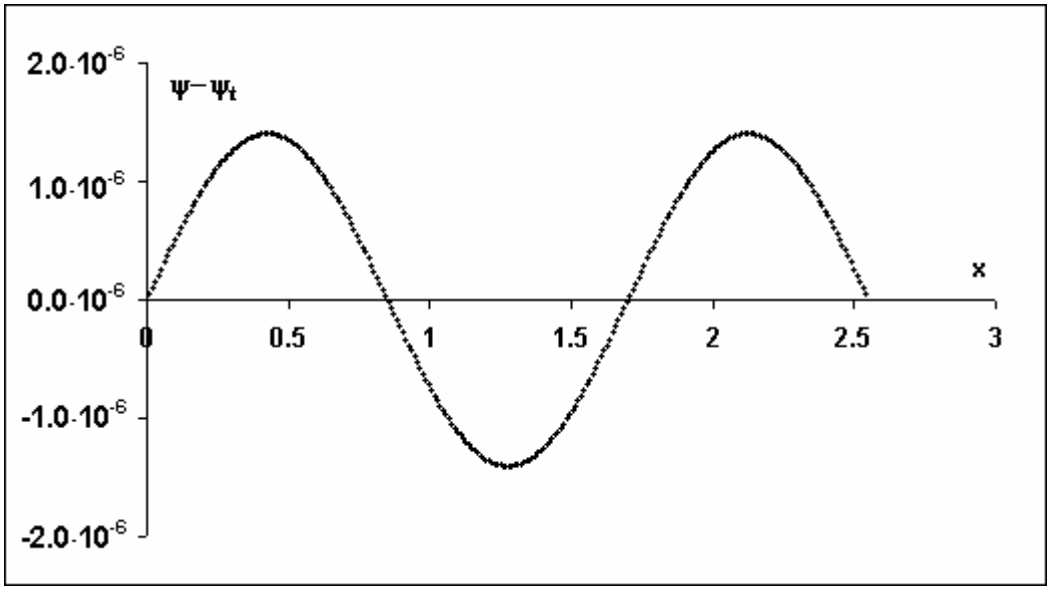
10.4 Numeryczne własności sieci sprzężonych oscylatorów

10.4.1 Drgania własne

Przydatność modelu sprzężonych oscylatorów harmonicznych i jego ograniczenia zostały sprawdzone praktycznie na przykładzie N = 250 mas m = 0.1 rozmieszczonych wzdłuż odcinka o długości L = 2.55 (a = 0.010159)

połączonych nieważkimi sprężynkami o współczynniku sprężystości k = 2.2 i długości swobodnej równej a. Amplituda fali stojącej została ustalona na A = 0.005, natomiast n = 3. Okazało się, że ten prosty model zupełnie dobrze odtwarza wszystkie zjawiska falowe, jeśli tylko krok czasowy ε nie przekracza pewnej ustalonej wartości. W większości przypadków stosowano ε = 0.1.

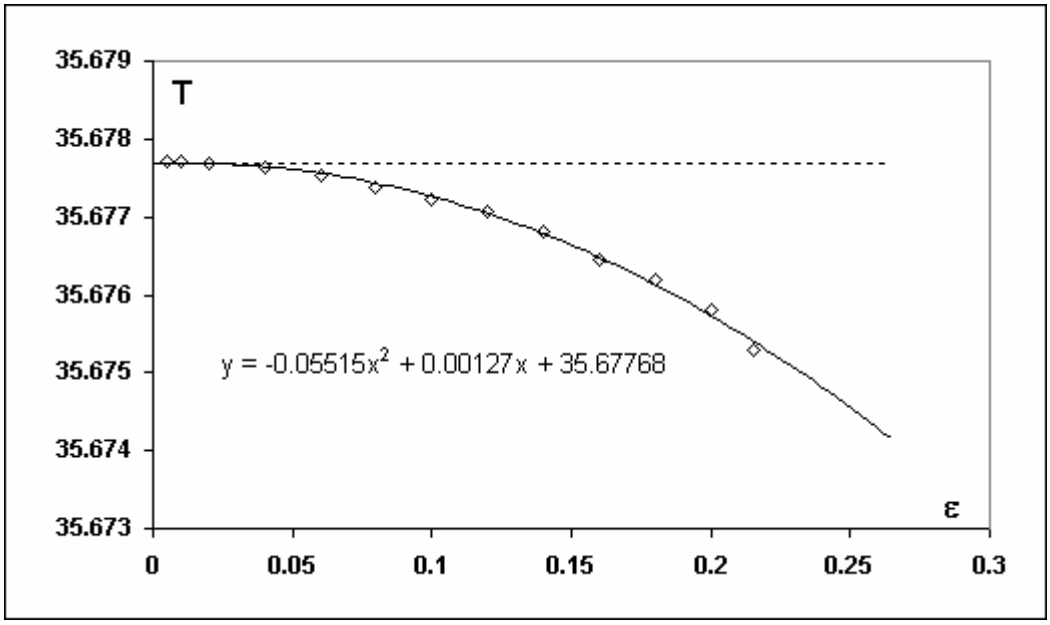
Przy tak zdefiniowanych warunkach ruchu ośrodka okazało się, że częstość drgań fali stojącej (n = 3) wynosi $\omega \cong 0.17611$ s^{-1} ($T \cong 35.67722$ s) i jest bardzo bliska częstości w przybliżeniu ciągłości ($\omega \cong 0.17612$ s^{-1}, T_{th} = 35.67768). Wykres 10.1 przedstawia odchylenie wychylenia w modelu dyskretnym od teorii po stosunkowo krótkim czasie działania (2000 kroków).



Wykres 10.1. Odchylenie punktów ośrodka od teorii po upływie t = 200 (h = 0.1).

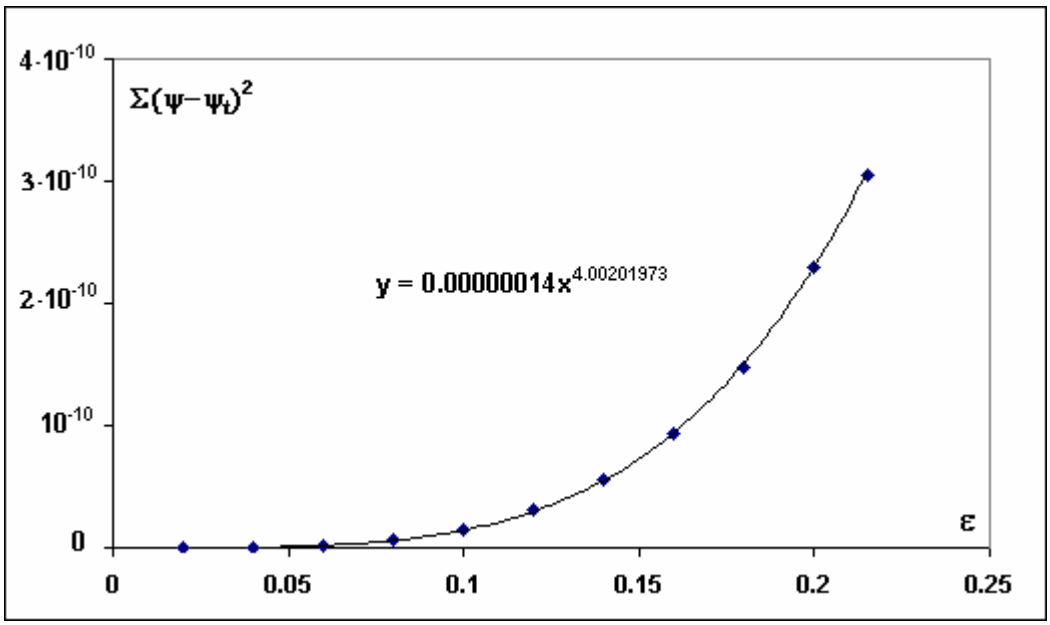
Widzimy, że ośrodek ma nieco za duże wychylenie w kierunku dodatnim i jednocześnie za małe w kierunku ujemnym, a maksymalne odchylenie od teorii po upływie 6 okresów przekracza 10^{-6}. Drgania obserwowane w ośrodku dyskretnym są stabilne przez długi czas i zachowują kształt sinusoidalny.

Działanie badanego modelu zależy od wyboru przyrostu kroku czasowego ε. Jego zwiększanie pogarsza zgodność drgań dyskretnych z teorią, jednak ich zachowanie jest zupełnie dobre dla wszystkich $\varepsilon \in (0, \varepsilon_{kr})$. Pokazuje to wykres 10.2 zależności okresu drgań fali stojącej (n = 3) od ε, z którego wynika, że do opisu tej zależności wystarcza rozwinięcie wielomianowe uwzględniające wyrazy stopnia drugiego. Widzimy tu prawidłowe asympotyczne zachowanie się modelu w granicy $\varepsilon \to 0$, gdyż wyraz wolny rozwinięcia (w przybliżeniu 35.67768249) jest równy z dokładnością lepszą niż 10^{-7} teoretycznej wartości okresu drgań wynoszącej ok. 35.67768245.



Wykres 10.2. Zależność okresu drgań modelu dyskretnego od ε (T_{th} = 35.67768).

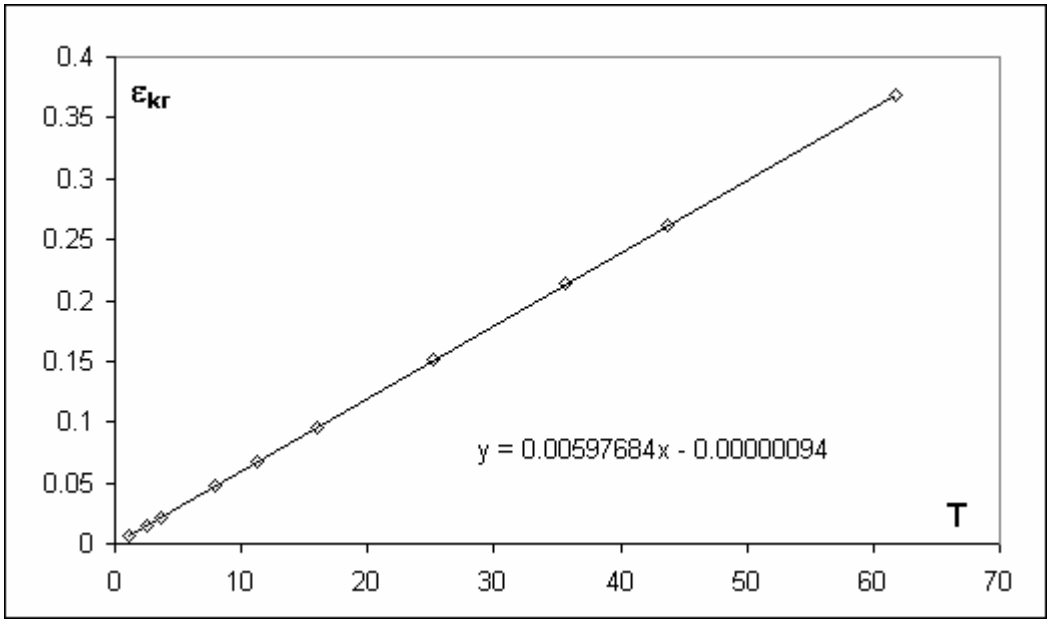
Inną naturalną miarą globalnego odchylenia od teorii w przestrzennym modelu dyskretnym jest suma kwadratów odchyleń dla wszystkich punktów ośrodka. Jej zależność od wartości kroku czasowego prezentuje wykres 10.3. Okazuje się, że z dużą dokładnością opisuje ją wielomian stopnia czwartego.



Wykres 10.3. Zależność błędu globalnego w modelu dyskretnym od ε (t = 20).

Zbliżanie się do krytycznej wartości przyrostu czasu ε_{kr} powoduje, że model dyskretny gwałtownie traci stabilność. Wykres 10.4 przedstawia liniową zależność między krytyczną wartością kroku czasowego a okresem drgań

układu. Postać krzywej na tym wykresie staje się oczywista, gdy przypomnimy sobie, że jedynym parametrem modelu (10.1) jest $\frac{K\varepsilon^2}{m} = \omega^2\varepsilon^2$, co powoduje, że zmiana kroku czasowego jest równoważna zmianie częstości drgań układu.



Wykres 10.4. Zależność pomiędzy ε_{kr} i okresem drgań fali stojącej (n = 3)

10.4.2 Drgania wymuszone

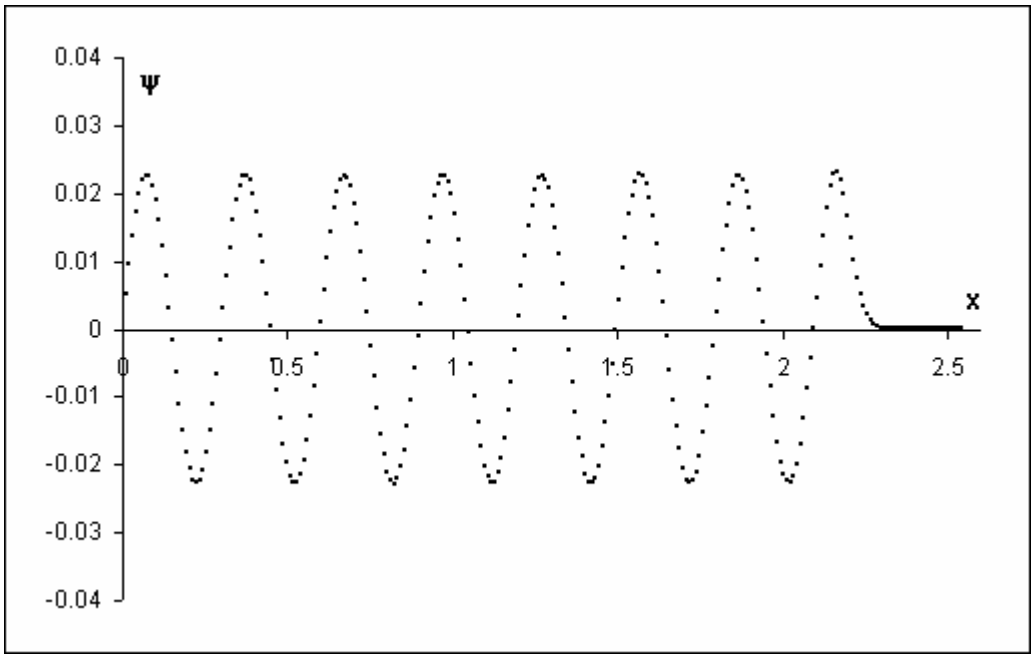
Aby uzyskać możliwość obserwacji drgań wymuszonych wystarczy zmienić pierwsze równanie w modelu (10.2) na

$$\psi_{1,n+1} - 2\psi_{1,n} + \psi_{1,n-1} = \frac{K\varepsilon^2}{m}(\psi_{2,n} - 2\psi_{1,n}) + \frac{F_0}{m}\cos(\omega n\varepsilon)\,. \qquad (10.36)$$

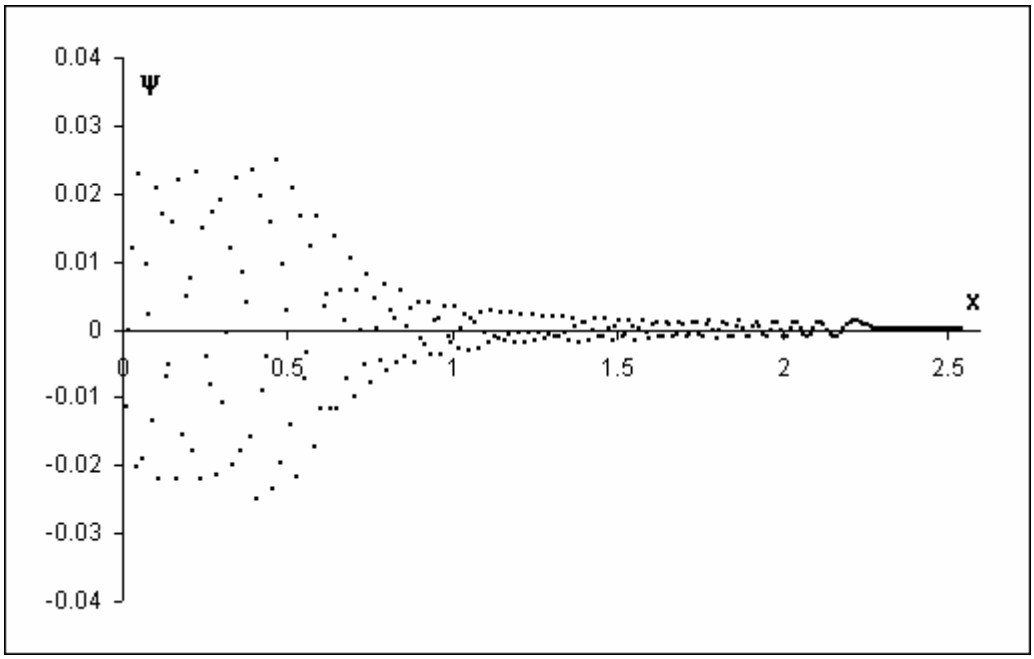
W przypadku ośrodka nieskończonego, o czym już wspominano, zakres dyspersyjny określony jest przedziałem od $\omega_{min}^2 = 0$ do $\omega_{max}^2 = 4K/m$ odpowiadającym iloczynom ka należącym do przedziału $\langle 0, \pi\rangle$. Fakt, że badany ośrodek jest skończony skutkuje pewnym zawężeniem zakresu dyspersyjnego. Mianowicie $\omega_{\min} = \sqrt{\frac{4K}{m}}\sin\frac{\pi a}{2l}$ oraz $\omega_{\max} = \sqrt{\frac{4K}{m}}\sin\frac{N\pi a}{2l}$. W tym przedziale częstości możliwych jest N postaci drgań własnych. Jeśli do modelu dyskretnego przyłożymy harmoniczną siłę wymuszającą o częstości z przedziału $[\omega_{min}; \omega_{max}]$ zawsze będziemy w pobliżu jednej z N częstości rezonansowych.

W analizowanym przykładzie numerycznym $\omega_{min} \cong 0.0415116$, $\omega_{max} \cong 9.3806478$ (dla ośrodka nieskończonego o tych samych parametrach $\omega_{max} = 44$). Odległości pomiędzy częstościami rezonansowymi są rzędu

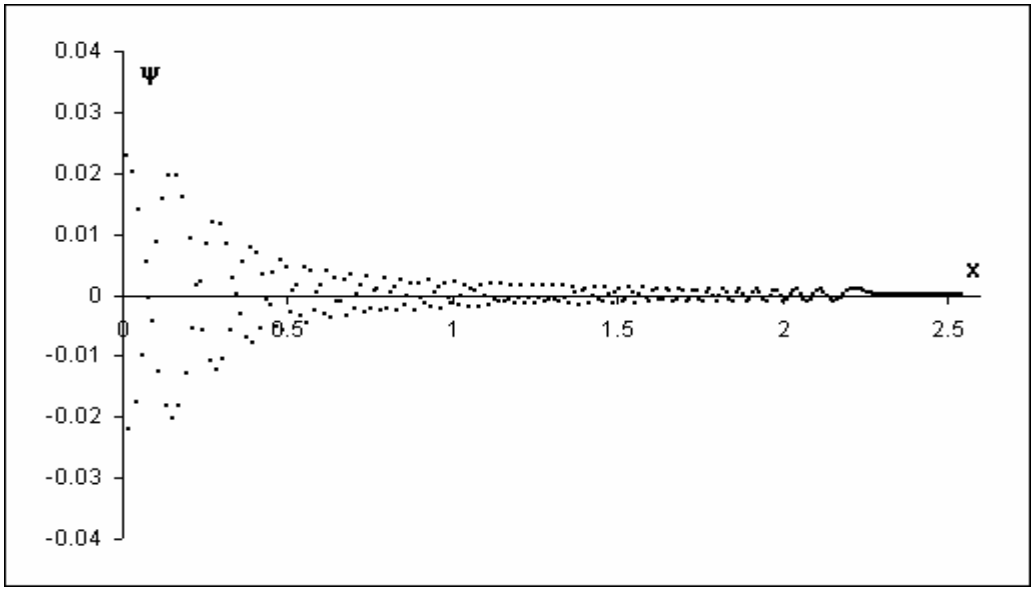
$\Delta\omega \cong 0.05$. Poza wskazanym wyżej przedziałem ośrodek powinien być reaktywny. Eksperyment numeryczny potwierdza omówione wyżej własności fizyczne badanego ośrodka. Obraz drgań wymuszonych przy różnych wartościach ω przedstawiają wykresy 10.5, 10.6 i 10.7.



Wykres 10.5. Obraz drgań ośrodka dla częstości siły wymuszającej ω = 1 po upływie *t* = 47 od jej włączenia (ε = 0.1)

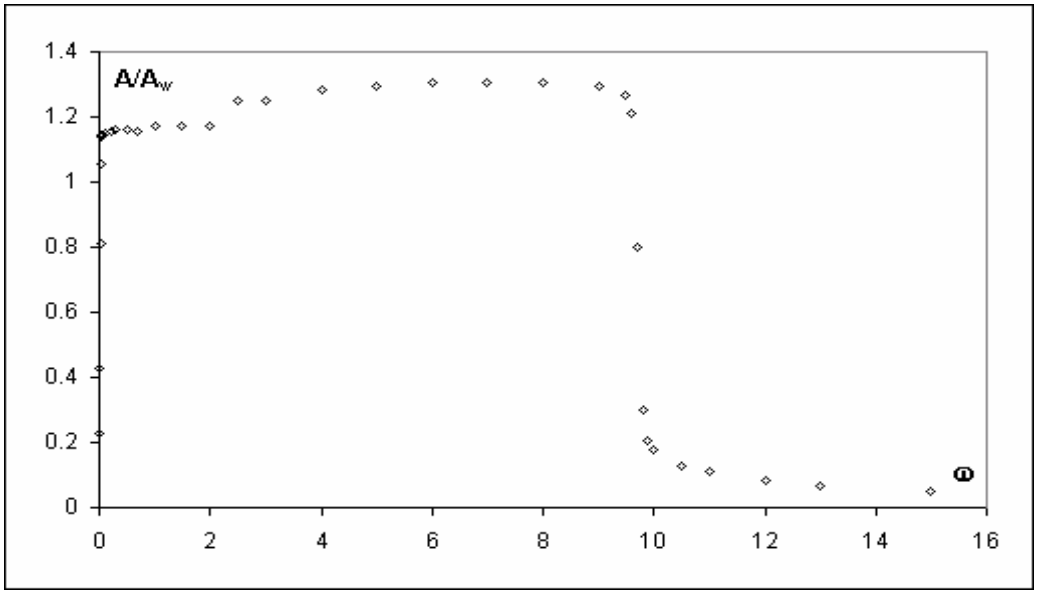


Wykres 10.6. Obraz drgań ośrodka dla częstości siły wymuszającej ω = 9.38 (granica reaktywności) po upływie *t* = 47 od jej włączenia (ε = 0.1)



Wykres 10.7. Obraz drgań ośrodka dla częstości siły wymuszającej ω = 9.7 (obszar reaktywny) po upływie t = 47 od jej włączenia (ε = 0.1)

Przekroczenie ω_{max} powoduje wejście w obszar, w którym w ośrodku powinna powstać przemienna fala eksponencjalna postaci (10.15). Reakcję ośrodka na wzbudzenie w funkcji jego częstości przedstawia wykres 10.8. Widzimy tu dwa obszary reaktywne, bardzo wąski na lewo od ω_{min} i szeroki na prawo od ω_{max}, którego położenie (gwałtowny spadek amplitudy drgań rozpoczyna się od częstości $\omega \cong 9.5$) dobrze zgadza się z podaną wyżej wartością teoretyczną.



Wykres 10.8. Zależność względnej maksymalnej amplitudy drgań ośrodka dyskretnego od częstości siły wymuszającej (t = 24, ε = 0.1).

10.5 Równanie falowe a dokładna dyskretyzacja oscylatora harmonicznego

W podrozdziale tym pokazane zostanie, że dokładna dyskretyzacja oscylatora harmonicznego może być zastosowana w przypadku równań różniczkowych cząstkowych. Rozpatrzmy równanie falowe postaci

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}. \tag{10.37}$$

Załóżmy, że $u(x,t) = \hat{u}(k,t)e^{ikx}$, co jest równoznaczne z rozpatrywaniem ewolucji czasowej modu odpowiadającego liczbie falowej *k*, która parametryzuje równanie. Po podstawieniu do (10.37) otrzymujemy

$$\frac{1}{c^2}\frac{\partial^2 \hat{u}(k,t)}{\partial t^2}e^{ikx} = -k^2\hat{u}(k,t)e^{ikx} \tag{10.38}$$

czyli

$$\frac{d^2\hat{u}}{dt^2} = -k^2c^2\hat{u} = -\omega^2\hat{u}. \tag{10.39}$$

Okazuje się zatem, że ewolucja czasowa każdego z modów opisana jest równaniem oscylatora harmonicznego. Wszystkie jego dyskretyzacje są więc de facto dyskretyzacjami równania falowego (10.37). Do równania (10.39) można zastosować (porównaj [14]) standardową (symplektyczną) metodę niejawną punktu środkowego (implicit midpoint):

$$\hat{u}^{n+1} - 2\hat{u}^n + \hat{u}^{n-1} = -\varepsilon^2\omega^2\hat{u}^n - \frac{1}{4}\varepsilon^2\omega^2(\hat{u}^{n+1} - 2\hat{u}^n + \hat{u}^{n-1}) \tag{10.40}$$

z drugim punktem dyskretyzacji danym przez

$$\hat{u}^1 = \frac{\left(2 - \frac{1}{2}\varepsilon^2\omega^2\right)}{\left(2 + \frac{1}{2}\varepsilon^2\omega^2\right)}\hat{u}^0 + \frac{2\varepsilon}{\left(2 + \frac{1}{2}\varepsilon^2\omega^2\right)}\dot{\hat{u}}^0 \tag{10.41}$$

jednak nie ma żadnych przeszkód, aby wykorzystać w tym miejscu zalety dokładnej dyskretyzacji (rozdział 3) równania oscylatora harmonicznego:

$$\hat{u}^{n+1} - 2(\cos\omega\varepsilon)\hat{u}^n + \hat{u}^{n-1} = 0. \tag{10.42}$$

Dzięki jej zastosowaniu odnosimy wiele korzyści. Jedna z nich jest taka, że prędkość grupowa pakietu falowego staje się dokładna, co sprawia, że schemat (10.42) odtwarza transport energii w ruchu falowym.

Otrzymaliśmy zatem narzędzie, które można wykorzystać choćby do badania ewolucji czasowej dowolnego odkształcenia ośrodka zadanego funkcjami $u(x,0)$

oraz $\dot{u}(x,0)$. Wykonując transformatę Fouriera otrzymujemy warunki początkowe dla schematów numerycznych (10.40) i (10.42):

$$\hat{u}(k,0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x,0) e^{-ikx} dx \tag{10.43}$$

$$\dot{\hat{u}}(k,0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \dot{u}(x,0) e^{-ikx} dx \tag{10.44}$$

Po wyznaczeniu dokładnej ewolucji czasowej funkcji $\hat{u}$ za pomocą dyskretyzacji (10.42) musimy wykonać jeszcze transformatę odwrotną, aby odtworzyć przebieg $u(x,t)$:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{u}(k,t) e^{ikx} dx\,. \tag{10.45}$$

Transformaty (10.43) i (10.44) należy obliczyć możliwie dokładnie, gdyż każde odchylenie będzie poddane propagacji w czasie niezależnie od dokładności zastosowanej dyskretyzacji. Na szczęście dla zadanego czasu *t* transformaty trzeba liczyć tylko dwukrotnie, a istniejące procedury są szybkie i dokładne (np. szybka transformata Fouriera FFT) [5, 84].

10.6 Dyskretyzacje równania falowego – eksperyment numeryczny

W celu sprawdzenia działania schematu numerycznego opisanego w poprzednim rozdziale podjęto próbę odtworzenia ewolucji czasowej wygięcia gaussowskiego danego wzorem

$$u_{th}(x,0) = Ae^{\frac{-(x-x_0)^2}{d^2}}\,. \tag{10.46}$$

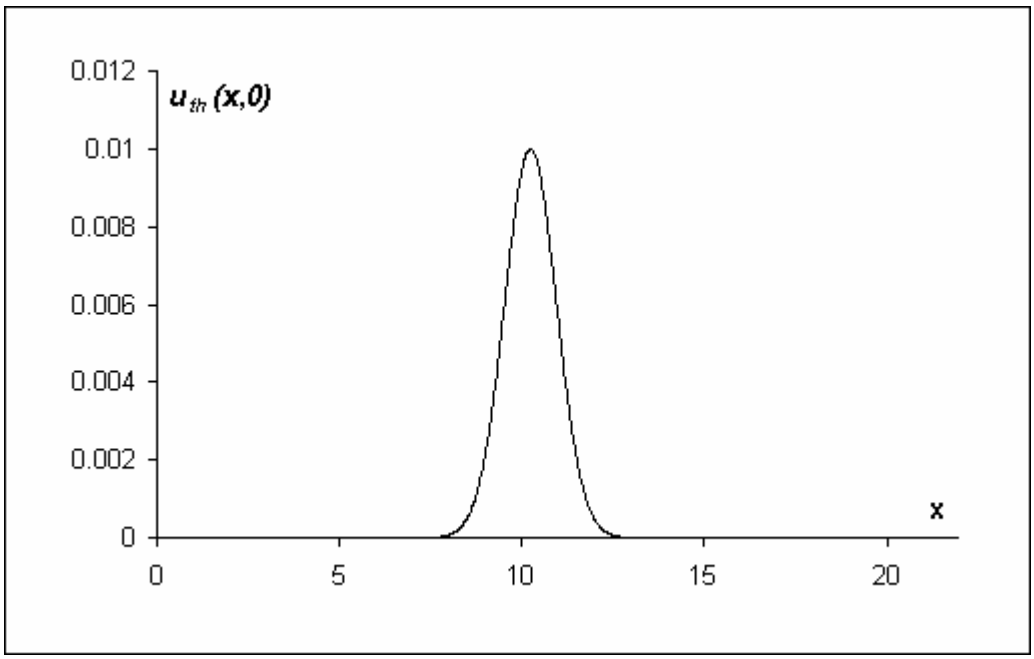
W przypadku ośrodka ciągłego związek dyspersyjny dany jest zależnością (10.10), prędkości fazowa i grupowa są sobie równe, a teoretyczny przebieg ewolucji odkształcenia (10.46) opisany jest przez

$$u_{th}(x,t) = \frac{1}{2} Ae^{\frac{-(x-x_0+ct)^2}{d^2}} + \frac{1}{2} Ae^{\frac{-(x-x_0-ct)^2}{d^2}}, \tag{10.47}$$

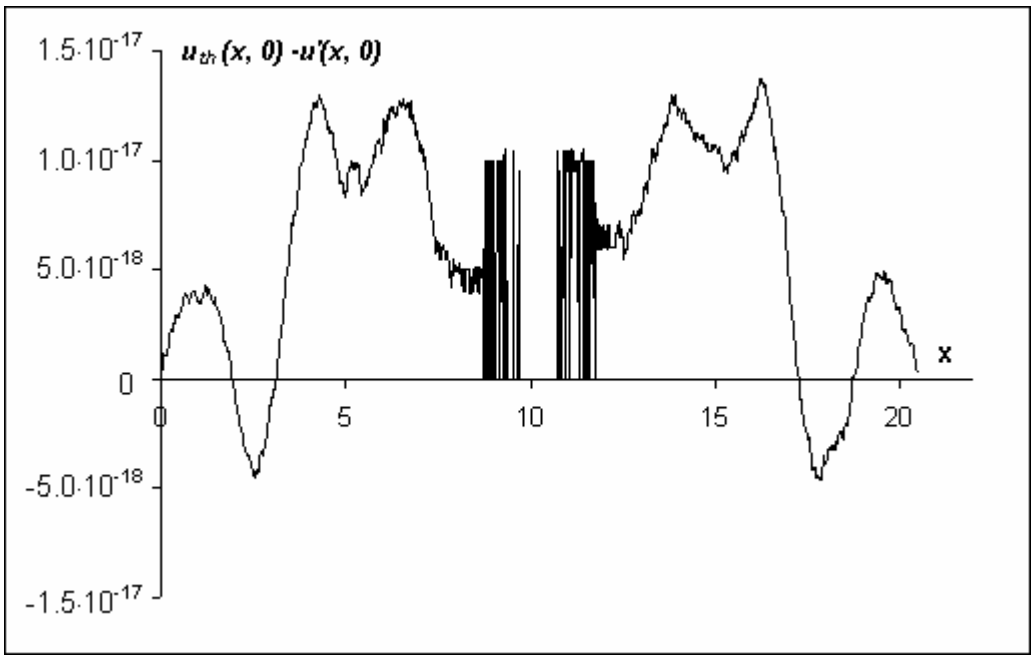
gdzie *c* oznacza prędkość grupową. Warto w tym miejscu zaznaczyć, że badany schemat może działać w przypadku dowolnego ośrodka. Jest to tylko kwestia innego wyboru związku dyspersyjnego.

W eksperymencie numerycznym nieskończony ośrodek ciągły modelowany był z pomocą skończonego odcinka o długości L = 20.48. Współczynniki wzoru (10.10) dobrano tak, że prędkości grupowa i fazowa wyniosły c = 0.077459667.

Parametry impulsu: A = 0.01, x_0 = L/2, d = 1. Dzięki temu w chwili t = 0 w pobliżu końców ośrodka wychylenie początkowe miało wartość $4.35 \cdot 10^{-48}$. Widzimy zatem, że ośrodek skończony może przez pewien czas bardzo dokładnie symulować zachowanie ośrodka nieskończonego. Wykresy 10.9 i 10.10 przedstawiają $u_{th}(x,0)$ oraz $u_{th}(x,0) - u'(x,0)$ przy czym $u'(x,0)$ oznacza przebieg warunku początkowego po wykonaniu transformaty odwrotnej ze zmiennej $\hat{u}$.

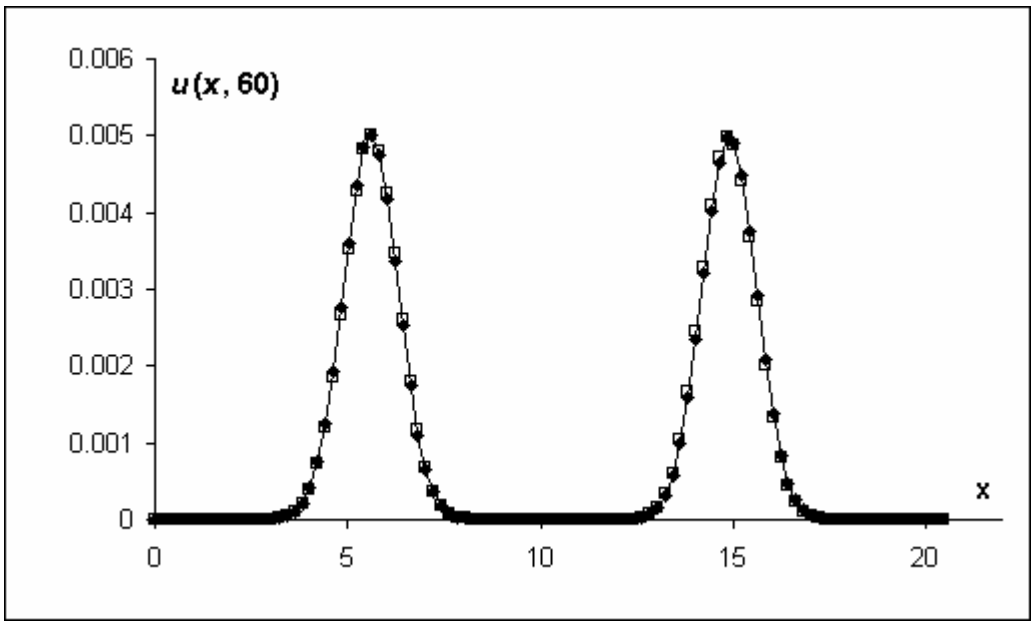


Wykres 10.9. Początkowe wychylenie ośrodka $u_{th}(x, 0)$.

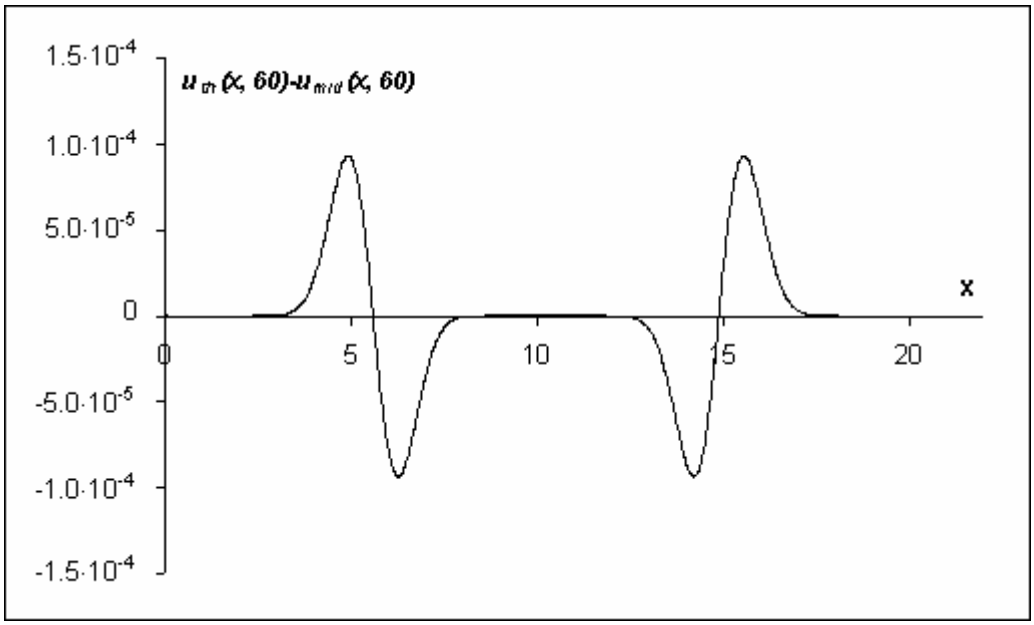


Wykres 10.10. Błąd wprowadzany przez transformatę Fouriera warunku początkowego do zmiennej $\hat{u}$ i na odwrót.

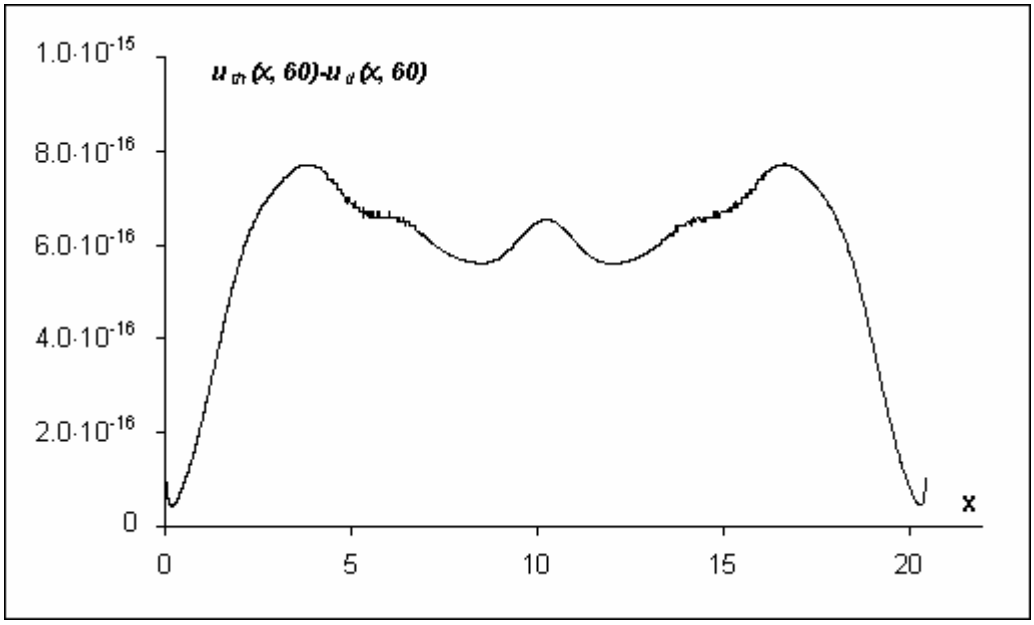
Wykres 10.10 pokazuje, że błąd względny odtworzenia przebiegu początkowego jest na poziomie 10^{-15} w szerokim przedziale wokół maksimum. W centralnej części wykresu odchylenie jest mniejsze niż poziom zarejestrowanego tam szumu numerycznego. Tak przetransformowany warunek początkowy poddano ewolucji czasowej z pomocą metody niejawnej punktu środkowego (10.40) i dyskretyzacji dokładnej (10.42). Wykres 10.11 przedstawia przykładowy obraz wychylenia ośrodka po ewolucji dla obu metod, natomiast 10.12 i 10.13 ich odchylenie od teorii.



Wykres 10.11. Wychylenie ośrodka po ewolucji czasowej dla dyskretyzacji implicit midpoint (białe kwadraty) i dokładnej (czarne romby); t = 60, ε = 0.5.



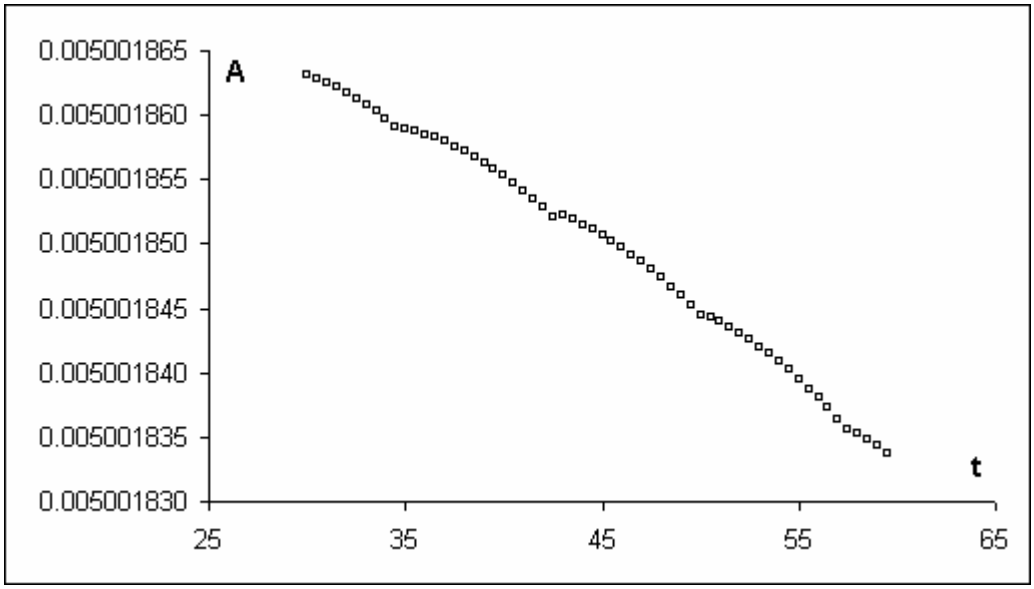
Wykres 10.12. Odchylenie przebiegu $u\ (t, x)$ od teorii dla dyskretyzacji niejawnej punktu środkowego (implicit midpoint); t = 60, ε = 0.5.



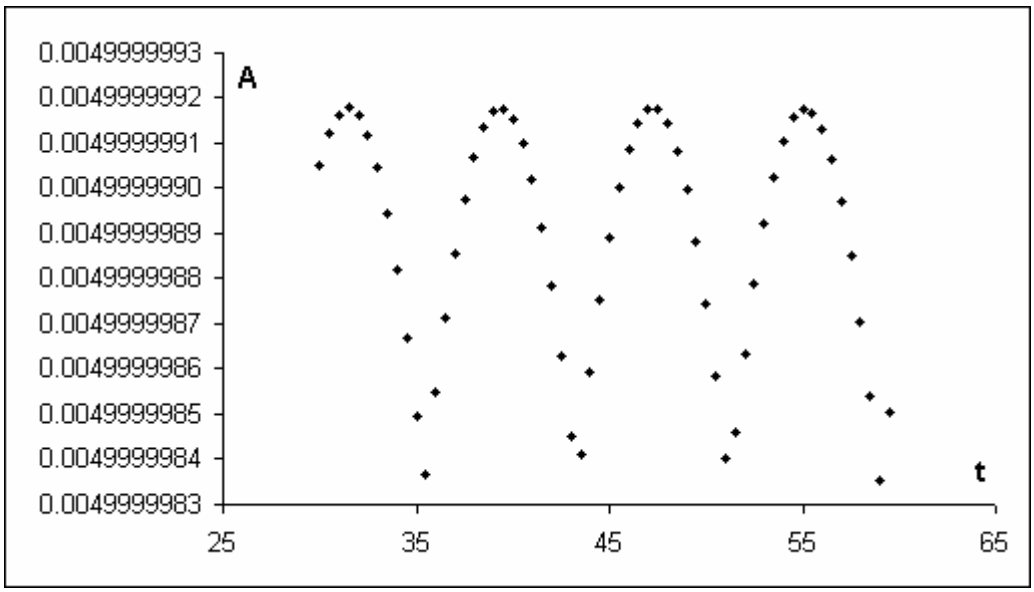
Wykres 10.13. Odchylenie *u(t, x)* od teorii dla dyskretyzacji dokładnej; *t* = 60, ε = 0.5.

Obydwie dyskretyzacje bardzo dobrze jakościowo odtwarzają przebieg teoretyczny *u(t, x)*, jednak odkształcenie ośrodka w przypadku metody niejawnej punktu środkowego porusza się z inną prędkością grupową i wraz z upływem czasu coraz bardziej rozmija się z przebiegiem teoretycznym. Dyskretyzacja dokładna zachowuje prędkość grupową z dokładnością do błędów obliczeń i odchylenie od teorii na wykresie 10.13 jest ponad 10 rzędów wielkości mniejsze niż w przypadku zastosowania metody punktu środkowego (*implicit midpoint*), choć także i w tym przypadku obserwujemy pewną kumulację rozmaitych błędów numerycznych.

W celu zbadania ruchu wygięć gaussowskich wyznaczono ich położenia i amplitudy. Wystarczającą dokładność uzyskano dopasowując wielomian stopnia drugiego do pięciu punktów ośrodka w wierzchołku maksimum. Wykresy 10.14 i 10.15 przedstawiają czasową zależność amplitudy wygięcia w przypadku dyskretyzacji niejawnej punktu środkowego oraz dokładnej. Na obu wykresach ujawniają się omawiane w rozdziale 4 efekty numeryczne związane ze zmienną w czasie konfiguracją punktów w wierzchołku wygięcia, jednak w przypadku dyskretyzacji *implicit midpoint* są one słabo widoczne, gdyż amplituda wygięcia systematycznie maleje. Zmiana amplitudy w dyskretyzacji dokładnej jest na tyle słaba, że dominującą rolę na jej wykresie odgrywają omawiane wcześniej dudnienia geometryczne. Na wykresie 10.15 można dostrzec bardzo mały dryf średniej amplitudy pochodzenia czysto numerycznego, ale z całą pewnością błąd numeryczny współczynnika kierunkowego przewyższy znacznie jego wartość.

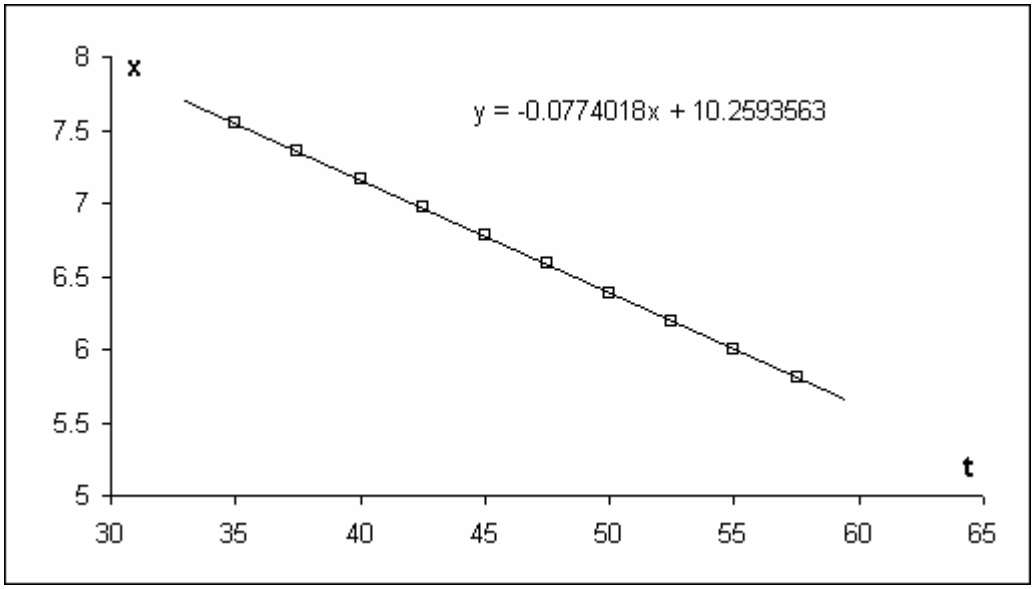


Wykres 10.14. Zmiana amplitudy wygięcia w czasie w przypadku dyskretyzacji niejawnej punktu środkowego; $\varepsilon = 0.5$.

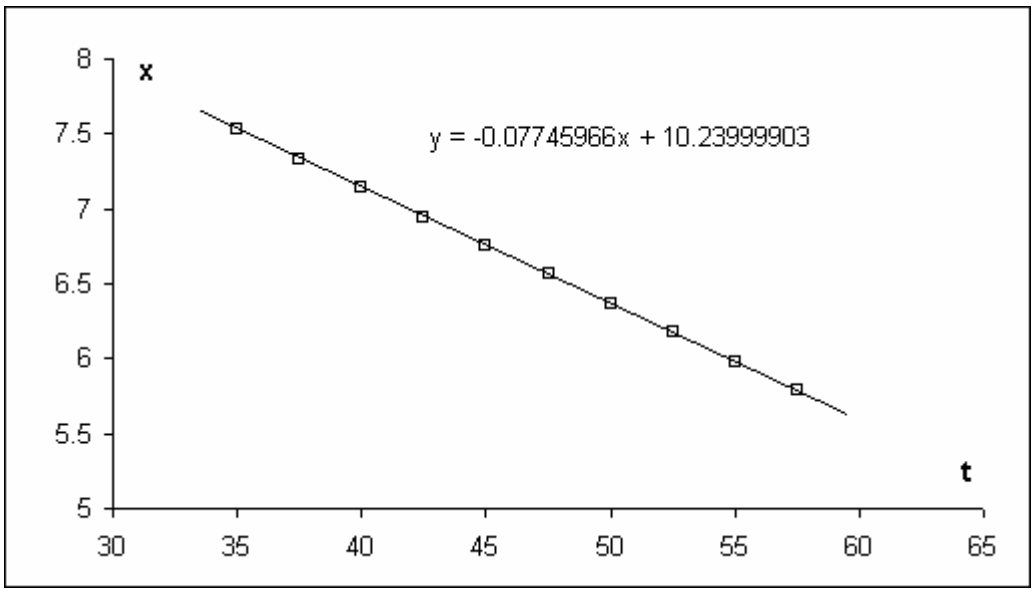


Wykres 10.15. Zmiana amplitudy wygięcia w czasie w przypadku dyskretyzacji dokładnej; $\varepsilon = 0.5$.

Wykresy 10.16 i 10.17 pozwalają porównać prędkości grupowe przemieszczania się odkształcenia w ośrodku dla obu dyskretyzacji.



Wykres 10.16. Położenie maksimum przemieszczającego się w lewo w funkcji czasu dla dyskretyzacji niejawnej punktu środkowego ($\varepsilon = 0.5$, $c = 0.077459667$).



Wykres 10.17. Położenie maksimum przemieszczającego się w lewo w funkcji czasu dla dyskretyzacji dokładnej ($\varepsilon = 0.5$, $c = 0.077459667$).

Współczynnik kierunkowy w obu przypadkach jest obarczony błędem rzędu $1.0 \cdot 10^{-7}$ natomiast wyraz wolny błędem rzędu $4.8 \cdot 10^{-6}$. Oznacza to, że dyskretyzacja dokładna w granicach błędów numerycznych daje prędkość grupową równą wartości teoretycznej. Równie dokładnie na podstawie wykresu 10.17 możemy ekstrapolować położenie wygięcia gaussowskiego w chwili $t = 0$.

10.7 Podsumowanie

W pierwszej części rozdziału poddano analizie teoretycznej i eksperymentalnej prosty model ruchu falowego w postaci sprzężonych oscylatorów harmonicznych, który zupełnie poprawnie symuluje wiele zjawisk falowych.

W drugiej jego części pokazano, w jaki sposób dokładna dyskretyzacja równania harmonicznego pozwala na otrzymanie dokładnej dyskretyzacji jednowymiarowego równania falowego. Pomysł polega na rozpatrywaniu ewolucji czasowej modu odpowiadającego liczbie falowej k, opisanej równaniem oscylatora harmonicznego. Dowolny warunek początkowy $u(x,0)$ i $\dot{u}(x,0)$ należy poddać transformacji Fouriera w celu uzyskania warunku początkowego $\hat{u}(k,0)$, $\dot{\hat{u}}(k,0)$ dla równania oscylatora. Po wyznaczeniu ewolucji czasowej funkcji $\hat{u}$ za pomocą dokładnej dyskretyzacji tego równania (dla wszystkich wartości k), pozostaje wykonać odwrotną transformatę Fouriera w celu powrotu do funkcji $u(x,t)$. Testy numeryczne pokazały, że algorytm ten pozwala obliczać ewolucję czasową fal z dokładnością ograniczoną jedynie błędami zaokrągleń numerycznych.

11 Dodatki

11.1 Liniowe równania różnicowe ze stałymi współczynnikami

W podrozdziale tym, opartym na pracy [20], przypomniana zostanie metoda rozwiązywania równań różnicowych ze stałymi współczynnikami. Polega ona na przedstawieniu równania w postaci równania macierzowego pierwszego rzędu. Ogólna postać dyskretnego równania liniowego drugiego rzędu jest następująca:

$$x_{n+2} = 2Ax_{n+1} + Bx_n, \tag{11.1}$$

i może być zapisana w formie macierzowej

$$y_{n+1} = My_n, \tag{11.2}$$

gdzie

$$y_n = \begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix}, \quad M = \begin{pmatrix} 2A & B \\ 1 & 0 \end{pmatrix}. \tag{11.3}$$

Ogólne rozwiązanie równania (11.2) ma oczywiście postać:

$$y_n = M^n y_0, \tag{11.4}$$

co redukuje rozwiązywanie równania dyskretnego do czysto algebraicznego problemu potęgowania zadanej macierzy.

Ta sama procedura może być zastosowana dla dowolnego liniowego równania różnicowego ze stałymi współczynnikami. Jeśli równanie różnicowe jest rzędu *m*, wówczas w celu otrzymania (11.2) definiujemy

$$y_n \coloneqq (x_{n+m}, x_{n+m-1}, ..., x_{n+1}, x_n)^T, \tag{11.5}$$

gdzie symbol *T* oznacza transpozycję.

Potęga M^n może być łatwo wyznaczona w przypadku, gdy macierz *M* może zostać zdiagonalizowana, czyli przestawiona w formie

$$M = NDN^{-1}, \tag{11.6}$$

gdzie *D* jest macierzą diagonalną. Wówczas oczywiście $M^n = ND^nN^{-1}$. Diagonalizacja jest możliwa, jeśli tylko macierz *M* posiada dokładnie *m* liniowo niezależnych wektorów własnych (w szczególności, jeśli równanie charakterystyczne (11.7) posiada *m* parami różnych pierwiastków). Wówczas kolumny macierzy *N* są właśnie wektorami własnymi *M*, a współczynniki diagonalne *D* jej wartościami własnymi.

Równanie charakterystyczne (det(M-λI) = 0) dla m = 2 (czyli dla (11.1)) ma postać

$$\Lambda^2 - 2A\lambda - B = 0\,. \tag{11.7}$$

Oznaczmy jego pierwiastki przez Λ_1, Λ_2 (patrz (3.39)). Jeżeli $\Lambda_1 \neq \Lambda_2$, wówczas procedura diagonalizacji daje

$$M = N\begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} N^{-1}, \tag{11.8}$$

przy czym kolumny N są wektorami własnymi M, to znaczy

$$N = \begin{pmatrix} \Lambda_1 & \Lambda_2 \\ 1 & 1 \end{pmatrix}. \tag{11.9}$$

Dlatego

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = N\begin{pmatrix} \Lambda_1{}^n & 0 \\ 0 & \Lambda_2{}^n \end{pmatrix} N^{-1}\begin{pmatrix} x_1 \\ x_0 \end{pmatrix}, \tag{11.10}$$

a po wykonaniu mnożenia otrzymujemy (3.38).

Przypadek wielokrotnych wartości własnych M jest technicznie bardziej skomplikowany. Aby obliczyć M^n możemy na przykład przetransformować M do postaci kanonicznej Jordana (patrz np. [56]). Tutaj zasugerujemy metodę, która jest bardzo efektywna w przypadku macierzy 2 × 2. Otóż na mocy twierdzenia Cayley'a–Hamiltona [56] każda macierz spełnia jej równanie charakterystyczne. W przypadku (11.1) oznacza to, że $M^2 = 2A\lambda M + B$. Jeśli pierwiastek jest podwójny ($B = -A^2$) można łatwo udowodnić przez indukcję, że

$$M^n = (1-n)A^n M + nA^{n-1}. \tag{11.11}$$

Wstawiając to do (11.2) otrzymujemy bezpośrednio (3.40).

11.2 Metody numeryczne dla równań różniczkowych zwyczajnych

Przedstawione tu zostaną, na podstawie pracy [20], informacje na temat niektórych metod numerycznych dla równań różniczkowych zwyczajnych i przykłady ich zastosowania do równania oscylatora harmonicznego (3.8).

Układ liniowych równań różniczkowych zwyczajnych (dowolnego rzędu) może zawsze być przedstawiony w postaci pojedynczego równania macierzowego pierwszego rzędu:

$$\dot{y} = Sy\,, \tag{11.12}$$

gdzie y jest wektorem a S zadaną macierzą (w ogólności zależną od czasu). Metody numeryczne są prawie zawsze (patrz [55]) konstruowane dla szerokiej klasy równań różniczkowych zwyczajnych (włączając nieliniowe):

$$\dot{y} = f(t, y). \tag{11.13}$$

Oznaczmy przez y_n numeryczną aproksymację dokładnego rozwiązania $y(t_n)$.
Metoda Eulera.

$$y_{k+1} = y_k + \varepsilon f(t_k, y_k). \tag{11.14}$$

W tym przypadku dyskretyzacja $\ddot{x} + x = 0$ dana jest przez (3.5).
Zmodyfikowana metoda Eulera.

$$\begin{aligned} y_{k+1} &= y_k + \varepsilon f\left(t_k + \frac{1}{2}\varepsilon, y_k + \frac{1}{2}\varepsilon f(t_k, y_k)\right), \\ y_{k+1} &= y_k + \frac{1}{2}\varepsilon(f(t_k, y_k) + f(t_k + \varepsilon, y_k + \varepsilon f(t_k, y_k))). \end{aligned} \tag{11.15}$$

Obydwie metody prowadzą do następującej dyskretyzacji $\ddot{x} + x = 0$:

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + x_n + \frac{1}{4}\varepsilon^2 x_{n-1} = 0. \tag{11.16}$$

Pierwiastki równania charakterystycznego są urojone i wynoszą

$$|\Lambda_1| = |\Lambda_2| = \sqrt{1 + \frac{\varepsilon^2}{4}}. \tag{11.17}$$

Metoda Gaussa-Legendre'a-Runge-Kutty pierwszego rzędu..

$$y_{k+1} = y_k + \varepsilon f\left(t_k + \frac{1}{2}\varepsilon, \frac{y_k + y_{k-1}}{2}\right). \tag{11.18}$$

Zastosowanie tego schematu numerycznego daje następującą dyskretyzację równania tłumionego oscylatora harmonicznego:

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{\varepsilon^2} + 2\gamma\frac{x_{n+1} - x_{n-1}}{2\varepsilon} + \omega_0{}^2\,\frac{x_{n+1} + 2x_n + x_{n-1}}{4} = 0. \tag{11.19}$$

W przypadku gdy $\gamma = 0$, $\omega_0 = 1$ (tj. $\ddot{x} + x = 0$) mamy

$$\Lambda_1 = \frac{2 + i\varepsilon}{2 - i\varepsilon}, \quad \Lambda_2 = \frac{2 - i\varepsilon}{2 + i\varepsilon}. \tag{11.20}$$

Formuła ekstrapolacyjna Adamsa-Bashforth'a.

$$y_{k+1} = y_k + \varepsilon\sum_{j=0}^{k} b_{kj} f(t_{n-j}, y_{n-j}), \tag{11.21}$$

gdzie b_{kj} są specjalnie wybranymi liczbami rzeczywistymi. W szczególności $b_{10} = 3/2$, $b_{11} = -1/2$, $b_{20} = 23/12$, $b_{21} = -4/3$, $b_{22} = 5/12$.
W przypadku $k = 1$, dla $\ddot{x} + x = 0$ otrzymamy dyskretyzację

$$\frac{x_{n+1}-2x_n+x_{n-1}}{\varepsilon^2}+\frac{9x_{n-1}-6x_{n-2}+x_{n-3}}{4}=0 \tag{11.22}$$

oraz równanie charakterystyczne postaci

$$\Lambda^4-2\Lambda^3+\left(1+\frac{9}{4}\varepsilon^2\right)\Lambda^2-\frac{3}{2}\varepsilon^2\Lambda+\frac{1}{4}\varepsilon^2=0\,. \tag{11.23}$$

Jest to równanie czwartego rzędu (nie posiadające rzeczywistych pierwiastków dla $\varepsilon \neq 0$).

11.3 Uwagi na temat eksperymentów numerycznych

W trakcie pracy nad niniejszą rozprawą kluczowe znaczenie miało praktyczne stosowanie komputerowych obliczeń numerycznych. W tym rozdziale omówione zostały metody iteracyjne niezbędne w przypadku stosowania niejawnych (zamkniętych) schematów numerycznych, w szczególności kosztochłonność (czas obliczeniowy) stosowanych metod.

11.3.1 Wprowadzenie

Jednym z ważniejszych nowych wyników prezentowanych w tej rozprawie jest przedstawiona w rozdziale 6 dyskretyzacja określana jako metoda lokalnie dokładnego dyskretnego gradientu. Mimo, że nie posiada ona tak pożądanych cech jak odwracalność w czasie i zachowywanie objętości przestrzeni fazowej, to jednak wykazuje się dużą stabilnością czasową i będąc metodą rzędu trzeciego niezwykle dokładnie symuluje rozwiązania równań różniczkowych. Precyzja działania tej metody, (odwzorowanie okresu oscylacji nieliniowych układów drgających dla małych kroków czasowych i pędów początkowych ze względnym odchyleniem na poziomie 10^{-14}) będąca jej największą zaletą, jest jednocześnie wyzwaniem dla eksperymentatora numerycznego pragnącego takie rezultaty liczbowe zaprezentować. Bardzo istotne stają się wówczas takie parametry obliczeń jak ilość miejsc dziesiętnych w używanych liczbach, błędy zaokrągleń, promienie zbieżności iteracji czy schematy całkowania numerycznego.

Istnieje wiele znakomitych książek z metod numerycznych, dlatego celem tego rozdziału nie będzie (może z małymi wyjątkami) powtarzanie zawartych w nich rozumowań. Chodzi raczej o to, aby czytelnik, zapoznając się z wynikami pewnej liczby eksperymentów numerycznych, zobaczył jak pracują wybrane metody i mógł ewentualnie wykorzystać te doświadczenia w swojej pracy. Dane liczbowe cytowane w tym rozdziale odnoszą do wahadła matematycznego ($V(x)=-\cos x$).

Wszyscy słyszeli o błędach zaokrągleń, ale niezbyt wiele osób myśli o nich na co dzień. Wiadomo, że zdarzają się w arytmetyce zmiennoprzecinkowej na odległych miejscach dziesiętnych i przeszkadzają raczej niewielkiej grupie ludzi zmuszonej do przeprowadzania bardzo precyzyjnych obliczeń. Zilustrujemy to przykładem. Weźmy ciąg geometryczny

$$\begin{aligned} a_0 &= 1 \\ a_{i+1} &= a_i q \end{aligned}. \tag{11.24}$$

Wzór na sumę *n* elementów tego ciągu jest powszechnie znany, obliczmy ją jednak bezpośrednio z pomocą komputera dla $q = 0.999$ posługując się standardowym typem rzeczywistym z 15 cyframi znaczącymi. Wyniki obliczeń dla różnych wartości *n* znajdziemy w tabeli 11.1.

Tabela 11.1. Wyniki sumowania ciągu geometrycznego dla $a_0 = 1$, $q = 0.999$ w zależności od liczby wyrazów i kolejności dodawania.

n	$a_0 + a_1 + a_2 + \ldots + a_n$	$a_n + a_{n-1} + a_{n-2} + \ldots + a_0$
5000	993.278888040130	993.278888040133
30000	999.999999999896	999.999999999907

Przyznajmy, że przykład jest dobrany dość tendencyjnie, jednak faktem pozostaje, że różne wartości w drugiej i trzeciej kolumnie tabeli 11.1 spowodowane są wyłącznie inną kolejnością dodawania tych samych liczb. Okazuje się, że przy większej liczbie składników można otrzymać w ten sposób różnicę w sumie już na 3 miejscu znaczącym od końca. Widzimy też przy okazji jak funkcjonują w naszych komputerach podstawowe aksjomaty matematyki. Na szczęście istnieje wiele algorytmów numerycznych dających akceptowalne wyniki, pomimo nieuchronnego wpływu błędów zaokrągleń na końcowy wynik.

11.3.2 Przegląd wybranych metod numerycznego rozwiązywania równań nieliniowych

Najlepsze dyskretyzacje prezentowane w tej pracy podane są często w postaci uwikłanej wymuszającej numeryczne rozwiązywanie równań. Dotyczy to w szczególności metody dyskretnego gradientu oraz jej ulepszonych wersji.

Problem generalnie polega na znalezieniu rozwiązań równania wektorowego

$$\vec{f}(\vec{x}) = 0, \tag{11.25}$$

lub, inaczej rzecz ujmując, układu *n* równań skalarnych na *n* niewiadomych

$$f_j(x_1,\ldots,x_n) = 0 \quad (j = 1,\ldots,n). \tag{11.26}$$

Służą do tego różne metody iteracyjne [26, 88]. Jeśli przez $\vec{x}^{(k)}$ oznaczymy kolejną aproksymację pierwiastka, to dowolną metodę iteracyjną możemy zapisać w postaci

$$\vec{x}^{(k+1)} = \vec{\varphi}(\vec{x}^{(k)}, \vec{x}^{(k-1)}, ..., \vec{x}^{(k-m+1)}) \tag{11.27}$$

przy czym funkcję $\vec{\varphi}$ nazywamy funkcją iteracyjną. Zakładamy, że zależy ona od wartości funkcji $\vec{f}$ i jej pochodnych w *m* punktach $\vec{x}^{(k)}, \vec{x}^{(k-1)}, ..., \vec{x}^{(k-m+1)}$. Sytuacja jest najprostsza, gdy *m* = 1:

$$\vec{x}^{(k+1)} = \vec{\varphi}(\vec{x}^{(k)}) \tag{11.28}$$

mamy wówczas do czynienia z metodą iteracyjną jednopunktową. Iteracja jest zbieżna, o ile $\vec{\varphi}$ jest odwzorowaniem zwężającym.

Metoda punktu stałego

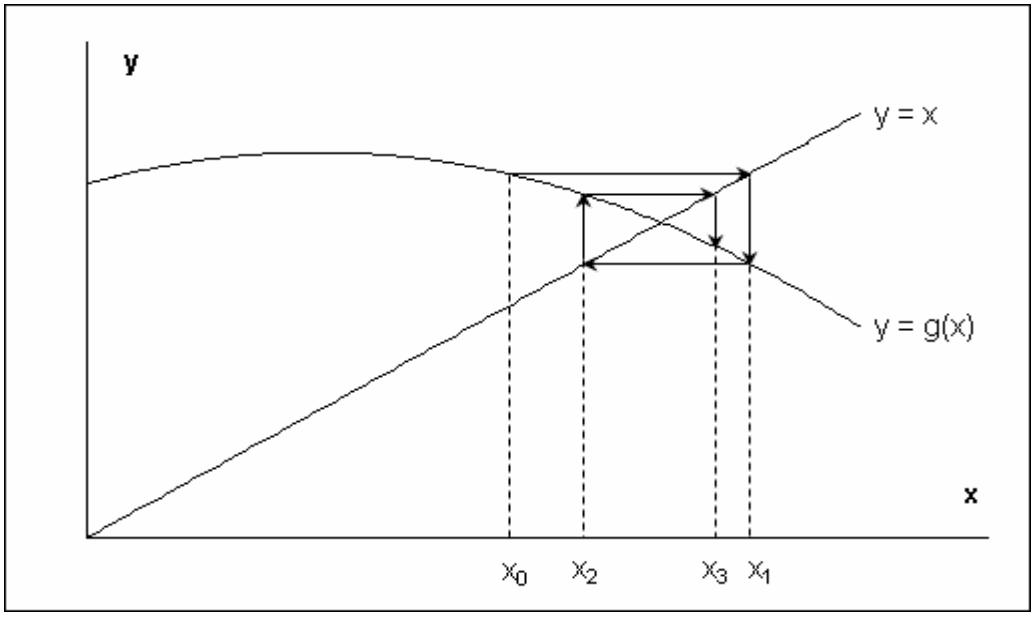
Jest to prosta metoda iteracyjna, którą możemy zastosować, o ile równanie wyjściowe (11.25) zapiszemy w postaci

$$\vec{x} = \vec{g}(\vec{x}) . \tag{11.29}$$

Wykorzystując twierdzenie Banacha o punkcie stałym [61] możemy wówczas zastosować iterację (metoda jednopunktowa)

$$\vec{x}^{(k+1)} = \vec{g}(\vec{x}^{(k)}) , \tag{11.30}$$

która jest zbieżna, jeżeli odwzorowanie $\vec{g}$ jest zwężające. W jednym wymiarze oznacza to, że $|g'(x)| < 1$ dla każdego *x* z otoczenia pierwiastka, które zawiera x_0 i x_1 (odpowiednio: punkt startowy i pierwszy punkt iteracji). Wykres 11.1 pokazuje działanie tej metody w przypadku malejącej funkcji jednej zmiennej.



Wykres 11.1. Ilustracja graficzna metody punktu stałego (funkcja jednej zmiennej).

W przypadku schematu lokalnie dokładnego dyskretnego gradientu $\vec{g}$ ma postać:

$$\vec{g}(\vec{x}) = \vec{g}(x, p) = \begin{pmatrix} x_n + \frac{1}{2}\delta_n (p + p_n) \\ p_n - \delta_n \frac{V(x) - V(x_n)}{x - x_n} \end{pmatrix}. \tag{11.31}$$

Jest ona jednocześnie funkcją iteracyjną i określa bezpośrednio kolejne kroki iteracji.

Metoda Newtona

Jest to bardzo ważna i szybkozbieżna metoda, z którą blisko związanych jest szereg metod pokrewnych jak metoda siecznych, *regula falsi*, metoda Steffensena i inne [26, 88]. Pomysł jest następujący. Otóż rozwijając $\vec{f}$ (równanie 11.25) w szereg wokół punktu $\vec{x}^{(i)}$ dostajemy

$$f_j(\vec{x}) \approx f_j(\vec{x}^{(i)}) + \sum_{k=1}^{n} (x_k - x_k{}^{(i)}) \left. \frac{\partial f_j{}^{(i)}}{\partial x_k} \right|_{\vec{x}=\vec{x}^{(i)}}. \tag{11.32}$$

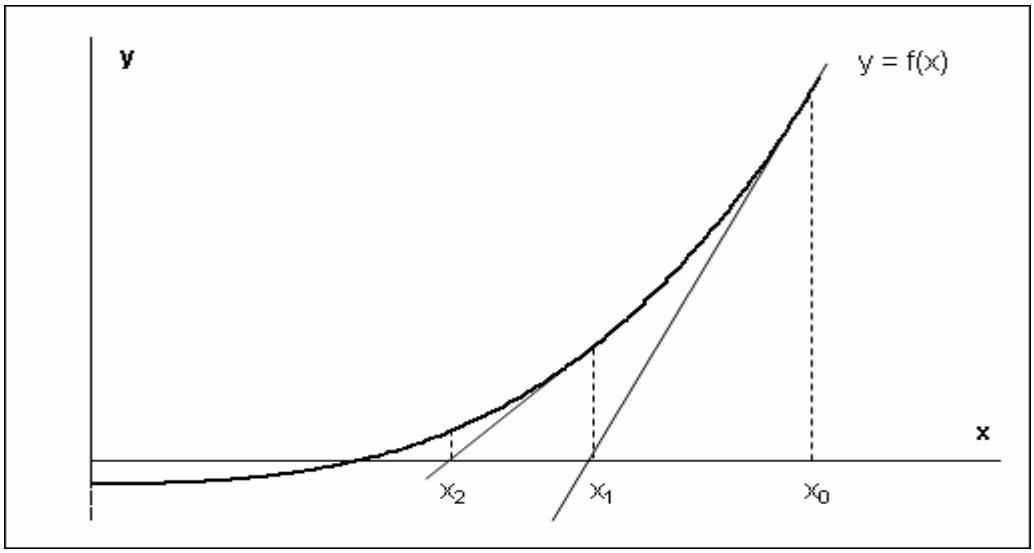
Następną aproksymację pierwiastka, $\vec{x}^{(i+1)}$ otrzymujemy kładąc $\vec{f}(\vec{x}) = 0$ i rozwiązując równanie liniowe względem $\vec{x}$. Otrzymujemy

$$\vec{x}^{(i+1)} = \vec{x}^{(i)} - (\nabla \vec{f}^{(i)})^{-1} \vec{f}^{(i)}(\vec{x}^{(i)}) \tag{11.33}$$

przy czym

$$\vec{f}^{(i)} = \vec{f}(\vec{x}^{(i)}), \quad (\nabla \vec{f}^{(i)})_{jk} = \frac{\partial f_j(\vec{x})}{\partial x_k}. \tag{11.34}$$

Wykres 11.2 ilustruje powyższą ideę na przykładzie funkcji jednej zmiennej.



Wykres 11.2. Ilustracja graficzna metody Newtona (funkcja jednej zmiennej).

W przypadku schematu lokalnie dokładnego dyskretnego gradientu mamy

$$\vec{f}(\vec{x}) = \vec{f}(x,p) = \begin{pmatrix} \dfrac{x-x_n}{\delta_n} - \dfrac{1}{2}(p+p_n) \\ \dfrac{p-p_n}{\delta_n} + \dfrac{V(x)-V(x_n)}{x-x_n} \end{pmatrix} \tag{11.35}$$

$$\nabla\vec{f}(\vec{x}) \equiv \nabla\vec{f}(x,p) = \left(\frac{\partial\vec{f}}{\partial x}, \frac{\partial\vec{f}}{\partial p}\right) = \begin{pmatrix} \dfrac{1}{\delta_n} & -\dfrac{1}{2} \\ \dfrac{V'(x)(x-x_n)+V(x_n)-V(x)}{(x-x_n)^2} & \dfrac{1}{\delta_n} \end{pmatrix} \tag{11.36}$$

$$(\nabla\vec{f}(\vec{x}))^{-1} = \frac{1}{\dfrac{1}{\delta_n^{\,2}} + \dfrac{V'(x)(x-x_n)+V(x_n)-V(x)}{2(x-x_n)^2}} \begin{pmatrix} \dfrac{1}{\delta_n} & -\dfrac{1}{2} \\ \dfrac{V(x)-V(x_n)-V'(x)(x-x_n)}{(x-x_n)^2} & \dfrac{1}{\delta_n} \end{pmatrix} \tag{11.37}$$

Oczywiście

$$\vec{x}^{(i)} = (x_{n+1}^{(i)}, p_{n+1}^{(i)})^T \tag{11.38}$$

lecz wygodnie jest przyjąć $\vec{x}^{(i)} = (x,p)^T$ oraz $\vec{x}^{(i+1)} = (\tilde{x},\tilde{p})^T$, co pozwala napisać równanie iteracyjne w postaci

$$\begin{pmatrix} \tilde{x} \\ \tilde{p} \end{pmatrix} = \begin{pmatrix} x \\ p \end{pmatrix} - (\nabla\vec{f}(x,p))^{-1}\,\vec{f}(x,p)\,. \tag{11.39}$$

Osiągalna dokładność metod iteracyjnych

Niech $x^{(1)}$, $x^{(2)}$, … będzie ciągiem przybliżeń pierwiastka α uzyskanym przy pomocy metody iteracyjnej $x^{(n+1)} = \varphi(x^{(n)})$. Niech wartość $\varphi(x^{(n)})$ będzie obarczona błędem $\delta^{(n)}$. Wtedy

$$x^{(n+1)} = \varphi(x^{(n)}) + \delta^{(n)} \quad (n = 0, 1, ...)\,. \tag{11.40}$$

Odejmując od tego związku równość $\alpha = \varphi(\alpha)$ i stosując twierdzenie o wartości średniej otrzymujemy

$$x^{(n+1)} - \alpha = \varphi'(\xi_n)(x^{(n)} - \alpha) + \delta^{(n)}, \text{ gdzie } \xi_n \in \text{int}(x^{(n)}, \alpha)\,. \tag{11.41}$$

Wynika stąd równość

$$(1-\varphi'(\xi_n))(x^{(n+1)} - \alpha) = \varphi'(\xi_n)(x^{(n)} - x^{(n+1)}) + \delta^{(n)}\,. \tag{11.42}$$

Zakładając, że

$$|\varphi'(\xi_n)| \le m < 1 \text{ i } |\delta^{(n)}| < \delta \tag{11.43}$$

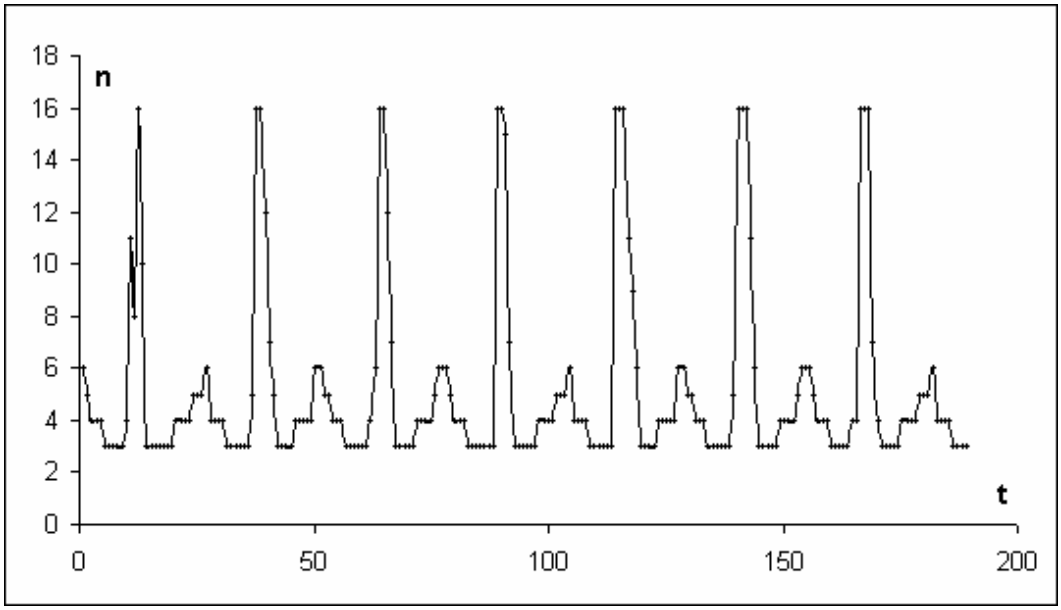
otrzymujemy nierówność

$$|x^{(n+1)} - \alpha| < \frac{m}{1-m}|x^{(n)} - x^{(n+1)}| + \frac{1}{1-m}\delta\,. \tag{11.44}$$

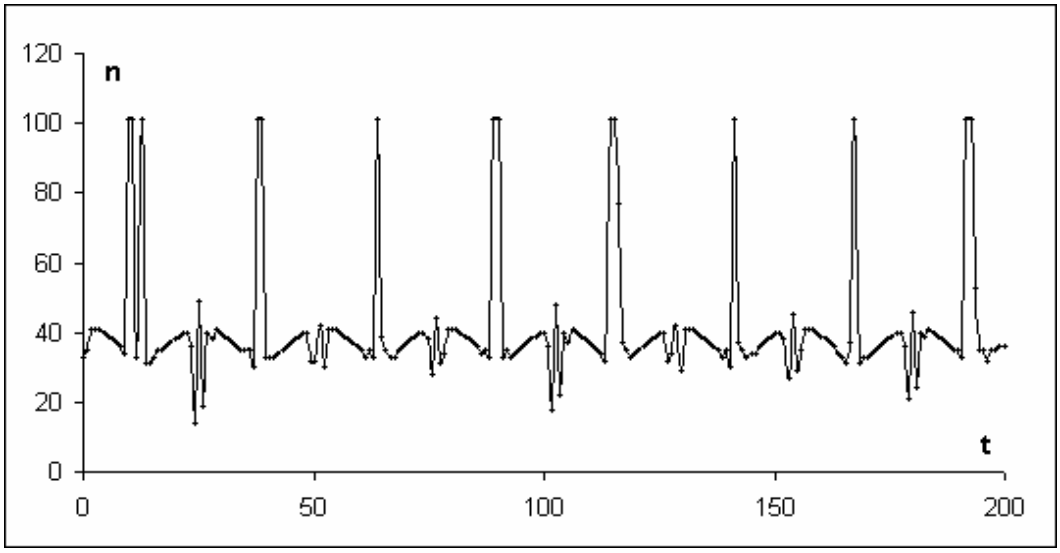
Prawa strona tej nierówności składa się z dwóch składników. Pierwszy szacuje błąd odcięcia, a drugi – błąd obliczeń. Dla dostatecznie dużych n błąd obliczeń staje się dominującą częścią składową błędu przybliżenia $x^{(n)}$ i dalsze iteracje nie poprawią już dokładności. Błąd powinien zachowywać się nieregularnie, pozostając w granicach $\pm\delta(1\text{-}m)$ [26].

11.3.3 Porównanie działania metod punktu stałego, Newtona i połowienia przedziału

Iteracje wykonywano przy założeniu dokładności bezwzględnej (błąd odcięcia) 10^{-16} i ustaleniu limitu kroków w liczbie 15 dla iteracji Newtona i 100 dla metody punktu stałego. Limity te są z grubsza dwukrotnie większe od obserwowanej maksymalnej średniej liczby kroków. Konieczność ich stosowania wynika z wzoru (11.44) - zarówno metoda punktu stałego, jak i Newtona dochodziły w niektórych punktach do teoretycznej granicy osiągalnej dokładności. Błąd obliczeń stawał się wówczas większy od błędu odcięcia, co przejawiało się przekraczaniem limitu kroków iteracji. Wykresy 11.3 i 11.4 pozwalają wskazać miejsca, w których błąd wynikający z możliwej do osiągnięcia dokładności jest większy niż 10^{-16}.

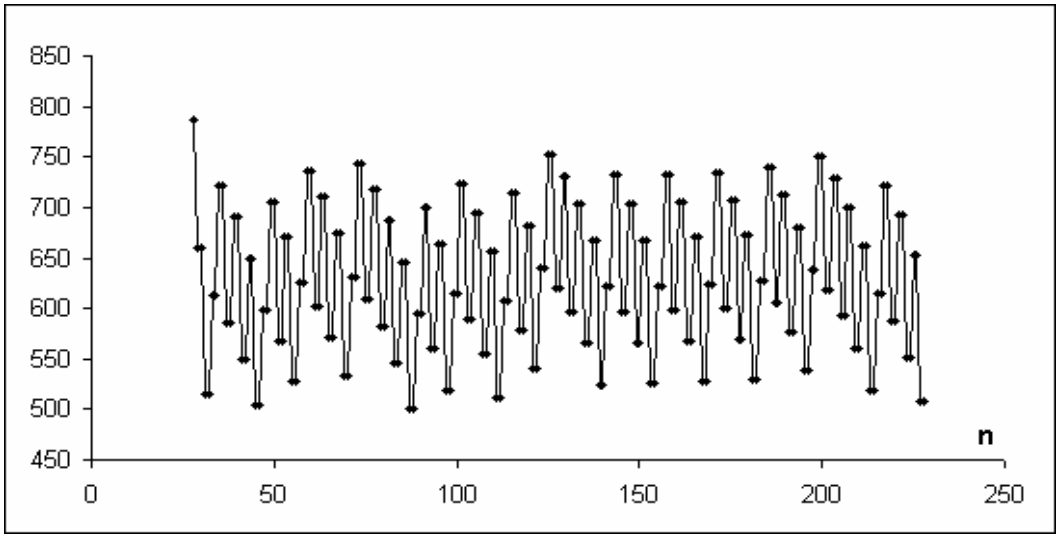


Wykres 11.3. Liczba iteracji w kolejnych punktach dyskretyzacji (p_0 = 1.999 999 999 9, ε = 0.9, T_{th} = 51.59687914). Metoda Newtona.

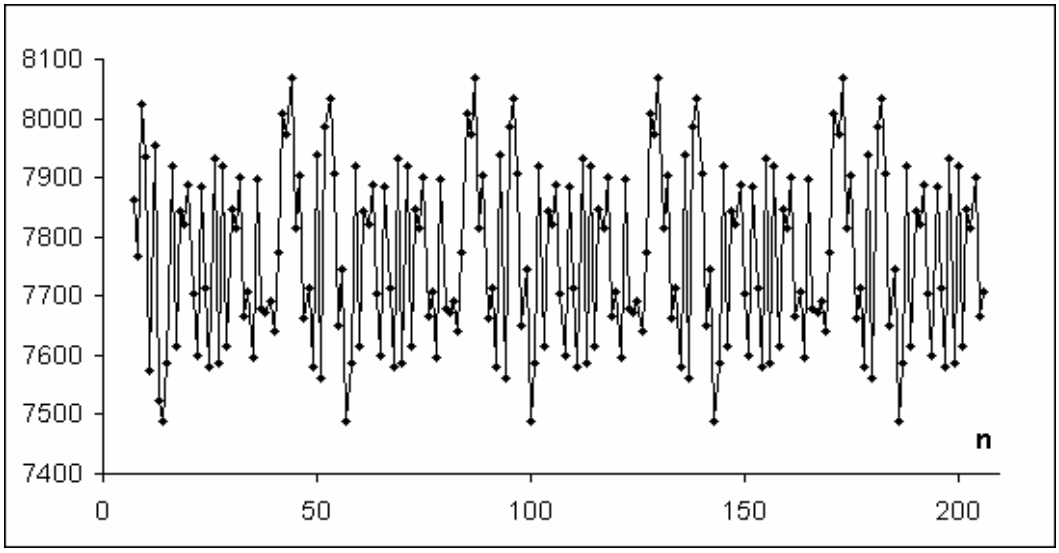


Wykres 11.4. Liczba iteracji w kolejnych punktach dyskretyzacji (p_0 = 1.999 999 999 9, ε = 0.9, T_{th} = 51.59687914). Metoda punktu stałego.

Liczby iteracji w obydwu metodach oscylują w pierwszym rzędzie z okresem, który możemy przypisać dyskretyzacji. Nakładają się na to subtelniejsze efekty powtarzające się w innych odstępach czasu. Ich natura została omówiona w rozdziale 4. Widzimy, że najtrudniejszą pracę obydwie metody iteracyjne mają w tych samych punktach – przy przechodzeniu dyskretyzacji przez położenie równowagi (tu iteracje zawsze kończyły się przed osiągnięciem limitu) i w pobliżu maksymalnych wychyleń. Bardzo płaski przebieg krzywej wychylenia w punktach zwrotnych (p_0 = 2 odpowiada separatrysie) ogranicza tu możliwą do osiągnięcia dokładność poniżej założonego błędu odcięcia dla każdej z metod.

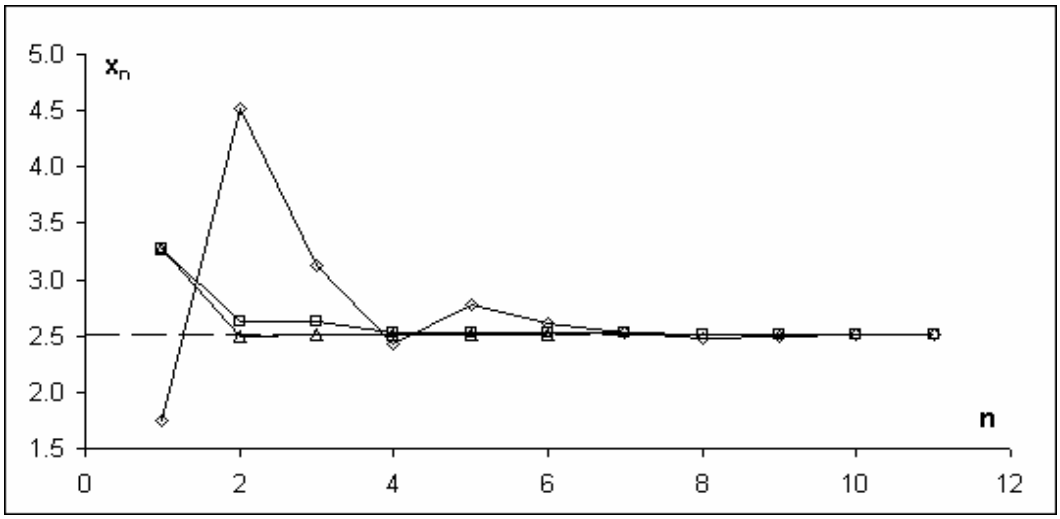


Wykres 11.5. Oscylacje na 3 ostatnich cyfrach po przecinku (15-17) w metodzie punktu stałego (poczynając od 28 kroku iteracji), p0 = 1.999 999 999 9, ε = 0.9, t = 12.6.



Wykres 11.6. Oscylacje na 4 ostatnich cyfrach po przecinku (14-17) w metodzie Newtona (poczynając od 7 kroku iteracji). p0 = 1.999 999 999 9, ε = 0.9, t = 12.6.

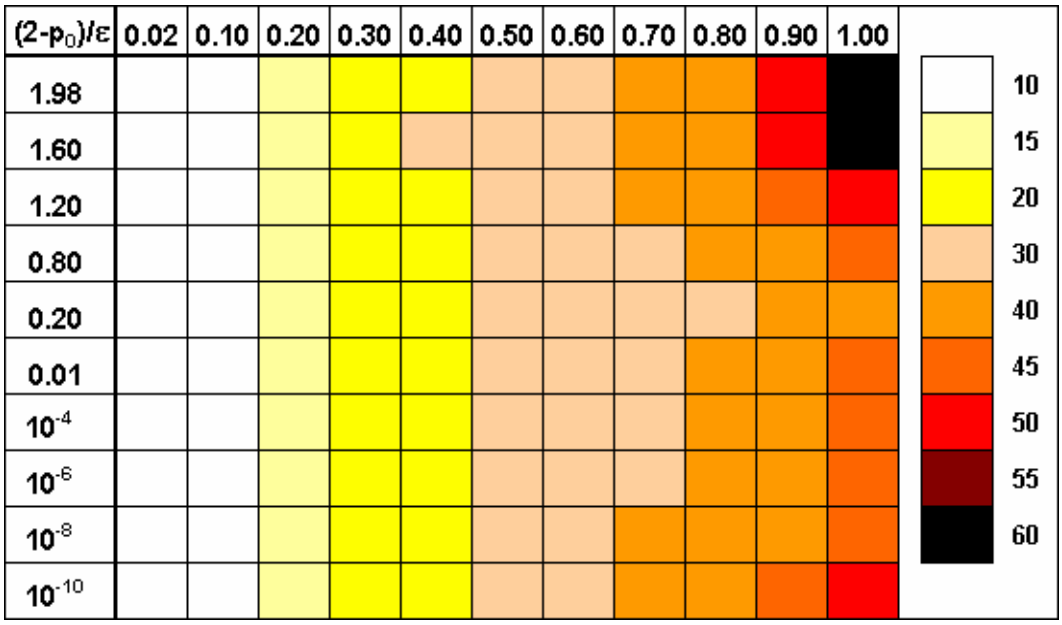
Na wykresach 11.5 i 11.6 przedstawiono oscylacje pojawiające się w metodzie punktu stałego i Newtona po przekroczeniu przez nie granicy osiągalnej dokładności. Dotyczą one tego samego punktu dyskretyzacji leżącego na pierwszym maksimum (wychylenie początkowe $x_0 = 0$). Iteracje pracowały na liczbach mających 18 znaczących cyfr dziesiętnych. Poszczególnym punktom przypisano liczby całkowite powstałe z 3 (metoda Banacha) lub 4 (metoda Newtona) ostatnich cyfr, które podlegały zmianom wskutek oscylacji. Ich promień okazał się dwukrotnie większy w metodzie Newtona. Poprawienie dokładności możemy uzyskać uśredniając wynik pewnej liczby oscylacji lub wykonując dalsze obliczenia za pomocą wolnozbieżnej, ale dokładniejszej metody bisekcji.



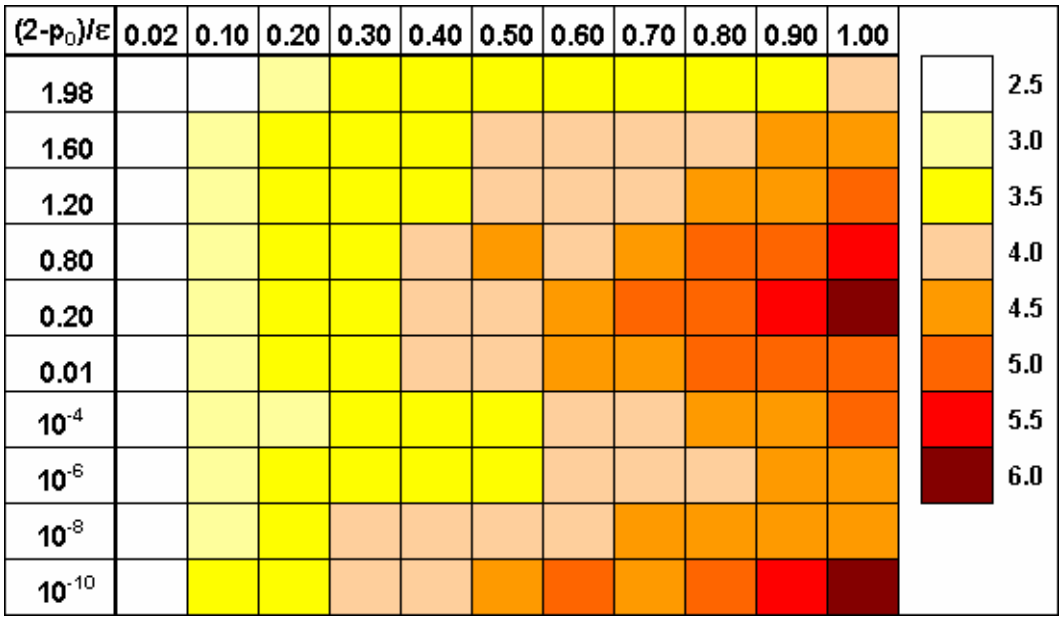
Wykres 11.7. Porównanie zbieżności metody bisekcji (romby), metody Banacha (kwadraty) i Newtona (trójkąty). p0 = 1.999 999 999 9, ε = 0.9, t = 0.9.

Wykres 11.7 przedstawia sposób, w jaki zbiegają do rozwiązania trzy prezentowane dyskretyzacje. Do osiągnięcia dokładności 10^{-14} metoda Newtona potrzebowała 5 kroków, Banacha – 26, natomiast bisekcji – 50.

Wykresy 11.8 i 11.9 pozwalają prześledzić zależność średniej liczby kroków metody Banacha i Newtona niezbędnych do osiągnięcia dokładności 10^{-16} w zależności od p_0 i ε dla dyskretyzacji lokalnie dokładnego dyskretnego gradientu symulującej wahadło matematyczne.



Wykres 11.8. Średnia liczba iteracji (z pierwszych 200 kroków) jako funkcja pędu początkowego i ε dla metody Banacha.



Wykres 11.9. Średnia liczba iteracji (z pierwszych 200 kroków) jako funkcja pędu początkowego i ε dla metody Newtona.

Na osi poziomej tych wykresów odłożono ε, zaś na osi pionowej dla skrócenia zapisu liczbę 2 - p_0 (p_0 = 2 odpowiada separatrysie). Liczby odpowiadające kolorom wyznaczają przedziały, w których zawiera się średnia liczba kroków wymaganych w danej iteracji. Metoda bisekcji ma stałą liczbę kroków (57 przy dokładności 10^{-16}) zależną wyłącznie od założonej dokładności obliczeń.

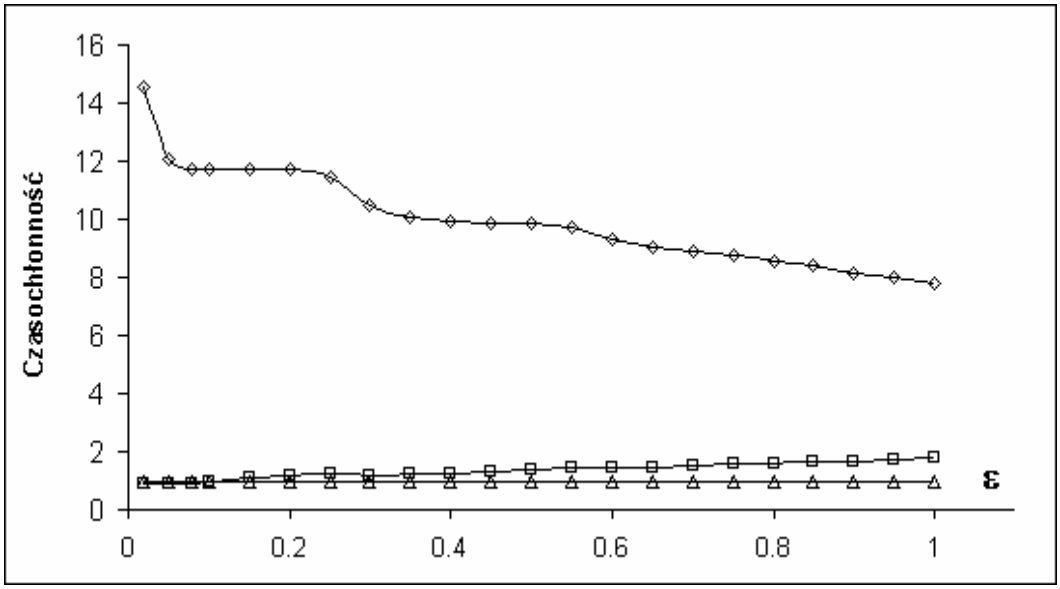
Średnia liczba kroków w metodzie Banacha jest zawsze zdecydowanie większa niż w metodzie Newtona (przy mniejszym koszcie wykonania jednego kroku), jednak ich mierzona liczbą kroków względna efektywność zależy od kombinacji parametrów wyjściowych. Przy małych pędach i dużych ε metoda Newtona jest 15-krotnie lepsza, przy dużych pędach i dużych ε jest lepsza 9-krotnie i tylko 5-krotnie przy p_0 = 1.8 i dużym kroku czasowym.

Rzeczywista przewaga metody Newtona jest mniejsza ze względu na różne koszty wykonania jednego kroku. Tabela 11.2 zawiera oszacowanie względnych czasów wykonania niektórych obliczeń dla różnych typów iteracji w metodzie lokalnie dokładnego dyskretnego gradientu zastosowanej do wahadła matematycznego w odniesieniu do schematu numerycznego *leap-frog*.

Tabela 11.2. Szacunkowe względne czasy wykonania wybranych obliczeń (wahadło matematyczne, metoda *leap-frog* i lokalnie dokładny dyskretny gradient).

1 krok *leap-frog*	1.00
1 iteracja Newtona	2.21
1 iteracja punktu stałego	0.67
1 iteracja bisekcji	0.95
predyktor (metoda Newtona i punktu stałego)	2.30
przedziały dla metody bisekcji	0.48

Okazuje się, że koszt jednego kroku w metodzie Banacha jest ponad trzykrotnie niższy niż w metodzie Newtona. Wykres 11.10 przedstawia porównanie względnej efektywności trzech badanych iteracji uwzględniające zarówno liczbę kroków, jak i czas wykonania jednego kroku przy niekorzystnym z punktu widzenia metody Newtona pędzie początkowym p_0 = 1.8. Dla kroków czasowych nie przekraczających 0.2 efektywności metod Banacha i Newtona są bardzo zbliżone (również dla innych pędów początkowych). Przy większych wartościach ε metoda Newtona staje się prawie 2-krotnie szybsza. Testując inne warunki początkowe możemy stwierdzić, że przewaga jest maksymalnie 4-krotna (dla p_0 = 0.4). Metoda bisekcji jest o rząd wielkości gorsza od pozostałych metod, jednak jej atrakcyjność rośnie w obszarach sprawiających trudność iteracjom Banacha i Newtona. Gdy kluczowa jest dokładność obliczeń (10^{-15} i większa), metoda bisekcji nie ma konkurencji – ma w porównaniu z nimi znacznie większą osiągalną dokładność.



Wykres 11.10. Porównanie czasochłonności różnych metod iteracyjnych (wahadło matematyczne, $p_0 = 1.8$). Romby – metoda bisekcji, kwadraty – metoda Banacha, trójkąty – metoda Newtona.

11.3.4 Podsumowanie

Rozdział ten prezentując szereg praktycznych aspektów testów numerycznych przeprowadzonych dla schematu numerycznego lokalnie dokładnego dyskretnego gradientu (rozdział 5) był okazją do krótkiego przeglądu metod stosowanych do numerycznego rozwiązywania równań (metoda punktu stałego, Newtona oraz bisekcji). Zwrócono uwagę na czynniki mające nieuchronny wpływ na dokładność obliczeń, takie jak błędy zaokrągleń i promień zbieżności iteracji. Dokonano też porównania czasochłonności wybranych metod standardowych i iteracyjnych.

12 Bibliografia

[1] R.P.Agarwal: *Difference equations and inequalities* (Chapter 3), Marcel Dekker, New York 2000.

[2] D.V.Anosov: "A note on the Kepler problem", *J. Dyn. Control Sys.*. **8** (2002) 413-442.

[3] C.Anteneodo, C.Tsallis: "Breakdown of exponential sensitivity to initial conditions: role of the range of interactions", *Phys. Rev. Lett.* **102** (1998) 5313-6.

[4] U.Ascher, S.Reich: "On some difficulties in integrating highly oscillatory Hamiltonian systems", *Computational Molecular Dynamics (Lect. Notes Comput. Sci. Eng.)*, Springer, Berlin 1999, 281-296.

[5] B.Baron, A.Marcol, S.Pawlikowski: *Metody numeryczne w Delphi 4*, Helion, Gliwice 1999.

[6] G.Benettin, A.Giorgilli: "On the Hamiltonian interpolation of near to the identity symplectic mappings with applications to symplectic integration algorithms", *J. Stat. Phys.* **74** (1994) 1117-43.

[7] D.G.Bettis: "Numerical integration of products of Fourier and ordinary polynomials", *Numer. Math.* **14** (1970) 421-434.

[8] D.G.Bettis: "Stabilization of finite difference methods of numerical integration", *Celestial Mech.* **2** (1970) 282-295.

[9] S.Blanes: "High order numerical integrators for differential equations using composition and processing of low order methods", *Appl. Numer. Math.* **37** (2001) 289-306.

[10] M.Błaszak: *Multi-Hamiltonian theory of dynamical systems*, Springer, Berlin-Heidelberg 1998.

[11] A.I.Bobenko, Yu.B.Suris: "Discrete time Lagrangian mechanics on Lie groups, with an application to the Lagrange top" *Commun. Math. Phys.* **204** 147-88.

[12] S.Breiter: "Explicit symplectic integrator for highly eccentric orbits", *Celestial Mech. Dyn. Astron.* **71** (1999) 229-241.

[13] S.Breiter, M.Fouchard, R.Ratajczak, W.Borczyk: "Two fast integrators for the galactic tide effects in the Oort cloud", *Mon. Not. R. Astron. Soc.* **377** (2007) 1151-62.

[14] T.Bridges, S.Reich: "Numerical methods for Hamiltonian PDE's", *J. Phys. A: Math. Gen.* **39** (2006) 5287-5320.

[15] A.J.Brizard: "A primer on elliptic functions with applications in classical mechanics", *Eur. J. Phys.* 30 (2009) 729-750.

[16] J.L.Cieśliński: "An orbit-preserving discretization of the classical Kepler problem", *Phys. Lett. A* **370** (2007) 8-12.

[17] J.L.Cieśliński: "On the exact discretization of the classical harmonic oscillator equation", *preprint arXiv:* 0911.3672v1 [math-ph] (2009).

[18] J.L.Cieśliński: "Comment on `Conservative discretizations of the Kepler motion'", *J. Phys. A: Math.Theor.* **43** (2010) 228001 (4pp).

[19] J.L.Cieśliński: “Locally exact modifications of numerical integrators”, w przygotowaniu.
[20] J.L.Cieśliński, B.Ratkiewicz: “On simulations of the classical harmonic oscillator equation by difference equations”, *Adv. Difference Eqs.* **2006** (2006) 40171.
[21] J.L.Cieśliński, B.Ratkiewicz: “Long-time behaviour of discretizations of the simple pendulum equation”, *J. Phys. A: Math. Theor.* **42** (2009) 105204.
[22] J.L.Cieśliński, B.Ratkiewicz: “Improving the accuracy of the discrete gradient method in the one-dimensional case”, *Phys. Rev. E*, **81** (2010) 016704.
[23] J.L.Cieśliński, B.Ratkiewicz: “Discrete gradient algorithms of high order for one-dimensional systems”, *preprint arXiv:* 1008.3895 [physics.comp-ph]
[24] J.L.Cieśliński, B.Ratkiewicz: “Energy-preserving numerical schemes of high accuracy for one-dimensional Hamiltonian systems”, *preprint arXiv:* 1009.2738 [cs.NA].
[25] B.Cordani: *The Kepler problem*. Birkhäuser, Basel 2003.
[26] G.Dahlquist, Å.Björck: *Metody numeryczne*, PWN, Warszawa, 1983.
[27] A.Deprit, A.Elipe, S.Ferrer: “Linearization: Laplace vs. Stiefel”, *Celestial Mech. Dyn. Astron.* **58** (1994) 151-201.
[28] H.B.Dwight: *Tables of integrals and other mathematical data*, 4th edition, Macmillan, New York 1961.
[29] K.Feng, M.-z.Qin: “Hamiltonian algorithms for Hamiltonian systems and a comparative numerical study”, *Comput. Phys. Commun.* **65** (1991) 173-187.
[30] É.Forest: “Geometric integration for particle integrators”, *J. Phys. A: Math. Gen.* **39** (2006) 5321-5377.
[31] A.Friedman, S.P.Auerbach: “Numerically induced stochasticity”, *J. Comput. Phys.* **93** (1991) 171-88.
[32] I.R.Gatland: “Numerical integration of Newton’s equations including velocity dependent forces”, *Am. J. Phys.* **62** (1994) 259-265.
[33] W.Gautschi: “Numerical integration of ordinary differential equations based on trigonometric polynomials”, *Numer. Math.* **3** (1961) 381-397.
[34] Z.Ge, J.Marsden: “Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators” *Phys. Lett . A* **133** (1998) 135-9.
[35] O.Gonzalez: “Time integration and discrete Hamiltonian systems”, *J. Nonl. Sci.* **6** (1996) 449-467.
[36] K.Goździewski, S.Breiter, W.Borczyk: “The long-term stability of extrasolar system HD 37124. Numerical study of resonance effects”, *Mon. Not. R. Astron. Soc.* **383** (2008) 989-99.
[37] D.Greenspan: “An algebraic, energy conserving formulation of classical molecular and Newtonian *n*-body interaction”, *Bull. Amer. Math. Soc.* **79** (1973) 432-427.
[38] E.Hairer. “Symmetric projection methods for differential equations on manifolds”, *BIT* **40** (2000) 726-34.
[39] E.Hairer: “Variable time step integration with symplectic methods”, *Appl. Num. Math.* **25** (1997) 219-227.

[40] E.Hairer, C.Lubich, G.Wanner: “Geometric numerical integration illustrated by the Störmer/Verlet method”, *Acta Numerica* **12** (2003) 399-450.
[41] E.Hairer, C.Lubich, G.Wanner: *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, wyd. 2., Springer, Berlin 2006.
[42] B.M.Herbst, M.J.Ablowitz, “Numerically induced chaos in the nonlinear Schrödinger equation, *Phys. Rev. Lett.* **62** (1989), no. 18, 2065–2068.
[43] J.Hietarinta, B.Grammaticos, B.Dorizzi, A.Ramani: “Coupling-constant metamorphosis and duality between integrable Hamiltonian systems”, *Phys. Rev. Lett.* **53** (1984) 1707-1710.
[44] F.B.Hildebrand, *Finite-Difference Equations and Simulations*, Prentice-Hall, New Jersey, 1968.
[45] R.Hirota: *Tech. Rep.* No. A-12, Hiroshima Univ., 1982 (cytowane za pracą [72]).
[46] A.Iserles: “Insight, not just numbers”, *Proc. 15th IMACS World Congress*, vol 2 ed. A.Sydow (Berlin: Wissenshaft & Technik Verlag) (1997) 589-94.
[47] A.Iserles: “Multistep methods on manifolds”, *IMA J. Numer. Anal.* **21** (2001) 407-19.
[48] A.Iserles, A.Zanna: “Qualitativ numerical analysis of ordinary differential equations”, w: *The Mathematics of Numerical Analysis (Lect. Appl. Math.)*, red. J.Renegar; American Mathematical Society, Providence RI 1996.
[49] T.Itoh, K.Abe: “Hamiltonian conserving discrete cannonical equations based on variational difference quotients”, *J. Comput. Phys.* **77** (1988) 85-102.
[50] Z.Jia, Leimkuhler: “Geometric integrators for multiple time-scale simulation” *J. Phys. A: Math. Gen.* **39** (2006) 5379-403.
[51] R.L.Kautz: “Chaos in a computer animated pendulum”, *Am. J. Phys.* **61** (1993) 407-415.
[52] R.Kozlov: “Conservative discretizations of the Kepler motions”, *J. Phys. A: Math. Theor.* **40** (2007) 4529-4539.
[53] P.Kustaanheimo, E.Stiefel: “Perturbation theory of Kepler motion based on spinor regularization”, *J. reine angew. Math.* **218** (1965) 204-219.
[54] R.A.LaBudde, D.Greenspan: “Discrete mechanics – a general treatment”, *J. Comput. Phys.* **15** (1974) 134-167.
[55] J.D.Lambert: *Numerical Methods for Ordinary Differential Systems*, John Wiley & Sons, Chichester, 1991.
[56] S.Lang: *Algebra*, Addison-Wesley, Massachusetts, 1965.
[57] J.Laskar, P.Robutel: “High order symplectic integrators for perturbed Hamiltonian systems” *Celst. Mech.* **80** (2001) 39-62.
[58] J.Laskar, P.Robutel, F.Joutel, A.C.M.Gastineau Correira, B.Levrand: “A long-term numerical solution for the insolation quantities of the earth”, *Astron. Astrophys.* **428** (2004) 261-85.
[59] H.Laue: “Elementary numerical integration methods”, *Am. J. Phys.* **56** (1988) 849-850.
[60] B.Leimkuhler, S.Reich: *Simulating Hamiltonian Dynamics*, Cambridge Univ. Press, Cambridge 2000.

[61] K.Maurin: *Analiza*, Cz. 1, PWN, Warszawa, (1977).
[62] R.I.McLachlan, M.Perlmutter, G.R.W.Quispel: "On the nonlinear stability of symplectic integrators", *BIT* **44** (2004) 99-117.
[63] R.I.McLachlan, G.R.W.Quispel: "Splitting methods", *Acta Numer*. **11** (2002) 341-434.
[64] R.I.McLachlan, G.R.W.Quispel: "Geometric integrators for ODE's", *J. Phys. A: Math. Gen.* **39** (2006) 5251-85.
[65] R.I.McLachlan, G.R.W.Quispel, N.Robidoux: "A unified approach to Hamiltonian systems, Poisson systems, gradient systems and systems with Lyapunov functions and/or first integrals", *Phys. Rev. Lett.* **81** (1998) 2399-403.
[66] R.I.McLachlan, G.R.W.Quispel, N.Robidoux: "Geometric integration using discrete gradients", *Phil. Trans. R. Soc.* A **357** (1999) 1021-45.
[67] R.E.Mickens: "Stable explicit schemes for equations of Schrödinger type", *Phys. Rev. A* **39** (1989) 5508-5511.
[68] R.E.Mickens: *Nonstandard finite difference models of differential equations*, World Scientific, Singapore 1994.
[69] R.E.Mickens: "A nonstandard finite-difference scheme for the Lotka-Volterra system", *Appl. Numer. Math.* **45** (2003) 309-314.
[70] S.Mikkola: "Practical symplectic methods with time transformations for the few-body problem", *Celestial Mech. Dyn. Astron.* **67** (1997) 145-165.
[71] B.V.Minchev, W.M.Wright: "A review of exponential integrators for first order semi-linear problems", *preprint NTNU/Numerics/N2/2005*, Trondheim 2005.
[72] Y.Minesaki, Y.Nakamura: "On an integrable discretization of integrable systems with a separatrix", *Phys. Lett. A* **250** (1998) 300-310.
[73] Y.Minesaki, Y.Nakamura: "A new discretization of the Kepler motion which conserves the Runge-Lenz vector", *Phys. Lett. A* **306** (2002) 127-133.
[74] Y.Minesaki, Y.Nakamura: "A new conservative numerical integration algorithm for three-dimensional Kepler motion based on the Kustaanheimo-Stiefel regularization theory", *Phys. Lett. A* **324** (2004) 282-292.
[75] J.Moser: "Regularization of Kepler's problem and the averaging method on a manifold", *Commun. Pure Appl. Math.* **23** (1970) 609-636.
[76] A.S.Mounim, B. de Dormale: "A note on Micken's finite-difference scheme for the Lotka-Volterra system", *Appl. Numer. Math.* **51** (2004) 341-344.
[77] W.Oevel: "Symplectic Runge-Kutta schemes", w: *Symmetries and Integrability of Difference Equations*, red. P. A. Clarkson, F. W. Nijhoff, *London Math. Soc. Lecture Note Ser.*, tom 255, Cambridge University Press, Cambridge, 1999, ss. 299–310.
[78] I.P.Omelyan, I.M.Mryglod, R.Folk: "Molecular dynamics simulations of spin and pure liquids with preservation of all the conservation laws", *Phys. Rev. E* **64** (2001) 016105.
[79] I.P.Omelyan, I.M.Mryglod, R.Folk: "Construction of high-order force-gradient algorithms for integration of motion in classical and quantum systems", *Phys. Rev. E* **66** (2002) 026701.
[80] A.Palczewski: *Równania różniczkowe zwyczajne*, WNT, Warszawa 1999.

[81] D.A.Pope: “An exponential method of numerical integration of ordinary differential equations”, *Commun. ACM* **6** (1963) 491-493.
[82] D. Potter: *Computational Physics*, JohnWiley & Sons, New York, 1973.
[83] R.B.Potts: “Differential and difference equations”, *Am. Math. Monthly* **89** (1982) 402-407.
[84] W.H.Press, S.A.Teukolsky, W.T.Vetterling, B.P.Flannery: *Numerical Recipes in C: the art of scientific computing*, Cambridge University Press 2002.
[85] G.R.W.Quispel, H.W.Capell: “Solving ODE’s numerically while preserving a first integral”, *Phys. Lett.* A **(218)** (1996) 223-8.
[86] G.R.W.Quispel, C.Dyt: “Solving ODE’s numerically while preserving symmetries, hamiltonian structure, phase space volume, or first integrals”, *Proc. 15th IMACS World Congress*, tom II, red. A.Sydow, ss. 601-607; Wissenschaft & Technik Verlag, Berlin 1997.
[87] G.R.W.Quispel, G.S.Turner: “Discrete gradient methods for solving ODE’s numerically while preserving a first integral”, *J. Phys. A: Math. Gen.* **(29)** (1996) L341-9.
[88] A.Ralston: *Wstęp do analizy numerycznej*, wyd. 3, PWN, Warszawa 1983.
[89] J.G.Reid: *Linear System Fundamentals, Continuous and Discrete, Classic and Modern* (New York: McGraw-Hill) (1983).
[90] L.I.W.Roeger: “Nonstandard discretization methods on Lotka-Volterra differential equations”, [in:] *Advances in the Applications of Nonstandard Finite Difference Schemes*, pp 615-650, edited by R.E.Mickens, World Scientific 2005.
[91] L.I.W.Roeger: “Nonstandard finite difference schemes for the Lotka-Volterra systems: generalization of Mickens’s method”, *J. Difference Equ. Appl.* **12** (2006) 937-948.
[92] L.I.W.Roeger: “Periodic solutions preserved by nonstandard finite-difference schemes for the Lotka-Volterra system: a different approach”, *J. Difference Equ. Appl.* **14** (2008) 481-493.
[93] W.Rubinowicz, W.Królikowski: *Mechanika teoretyczna*, PWN, Warszawa, (1978).
[94] R.D.Ruth: “A canonical integration technique”, *IEEE Trans. Nuclear Science* NS-30 (1983) 2669-2671.
[95] I.M.Ryzhik, I.S.Gradshteyn: *Tablice całek, sum, szeregów i iloczynów*, wyd. 3, GITTL, Moskwa 1951 [po rosyjsku].
[96] J.M.Sanz-Serna: “An unconventional symplectic integrator of W. Kahan”, *Appl. Numer. Math.* **16** (1994) 245-250.
[97] A.Sergyeyev, M.Błaszak: “Generalized Stäckel transform and reciprocal transformation for finite-dimensional integrable systems”, *J. Phys. A: Math. Theor.* **41** (2008) 105205.
[98] J.C.Simo, N.Tarnow, K.K.Wong: “Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics", *Comput. Methods Appl. Mech. Eng.* **100** (1992) 63-116.

[99] M.M. de Souza: "Discrete-to-continuum transitions and mathematical generalizations in the classical harmonic oscillator", *preprint* hep-th/0305114v5 (2003).
[100] R.W.Stanley: "Numerical methods in mechanics", *Am. J. Phys.* **52** (1984) 499-507.
[101] E.Stiefel, D.G.Bettis: „Stabilization of Cowell's method", *Numer. Math.* **13** (1969) 154-175.
[102] A.S.Stuart: "Numerical analysis of dynamical systems", *Acta Numerica* **3** (1994), 467–572.
[103] Yu.B.Suris: "On integrable standard-like mappings", *Funct. Anal. Appl.* **23** (1989) 74-6.
[104] Yu.B.Suris: *The Problem of Integrable Discretization: Hamiltonian Approach*, (Basel: Birkhäuser) (2003) chapter 20.
[105] G.J.Sussman, J.Wisdom: "Chaotic evolution of the solar system", *Science* **257** (1992) 56-62.
[106] M.Suzuki: "Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations", *Phys. Lett. A* **146** (1990) 319-323.
[107] H.K.Urbantke: "The Hopf fibration – seven times in physics", *J. Geom. Phys.* 46 (2003) 125-150.
[108] M.D.Vivarelli: "The KS transformation in hypercomplex form", *Celestial Mech.* **29** (1983) 45-50.
[109] J.Vrbik: "Perturbed Kepler problem in quaternionic form", *J. Phys. A: Math. Gen.* **28** (1995) 193-198.
[110] J.Waldvogel: "Quaternions and the perturbed Kepler problem", *Celestial Mech. Dyn. Astron.* **95** (2006) 201-212.
[111] A.R.Walton, D.E.Manolopoulos: "A new semiclassical initial value method for Franck-Condon spectra", *Mol. Phys.* **87** (1996) 961-78.
[112] J.Wisdom, M.Holman: "Symplectic maps for *N*-body problem", *Astron. J.* **102** (1991) 1528-38.
[113] H.Yoshida: "Construction of higher order symplectic integrators", *Phys. Lett. A* **150** (1990) 262-268.
[114] H.Yoshida: "Recent progress in the theory and application of symplectic integrators", *Celest. Mech. Dynam. Astron.* **56** (1993) 27-43.