Discrete gradient algorithms of high order for one-dimensional systems

Jan L. Cieśliński†, Bogusław Ratkiewicz‡‡
Uniwersytet w Białymstoku, Wydział Fizyki, ul. Lipowa 41, 15-424 Białystok, Poland

Abstract
We show how to increase the order of one-dimensional discrete gradient numerical integrator without losing its advantages, such as exceptional stability, exact conservation of the energy integral and exact preservation of the trajectories in the phase space. The accuracy of our integrators is higher by several orders of magnitude as compared with the standard discrete gradient scheme (modified midpoint rule) and, what is more, our schemes have very high accuracy even for large time steps.

PACS Numbers: 45.10.-b; 02.60.Cb; 02.70.-c; 02.70.Bf
Key words and phrases: geometric numerical integration, long time numerical evolution, energy integral, discrete gradient method

†e-mail: janek@alpha.uwb.edu.pl
‡e-mail: bogr@poczta.onet.pl
‡‡Permanent address: I LO, ul. Śródmieście 31, 16-300 Augustów, Poland
1 Introduction

In this paper we introduce and develop energy-preserving discrete gradient schemes of high order. Discrete gradient schemes are useful tools for numerical integration of many-body dynamical systems [1, 2, 3, 4, 5]. They preserve exactly (up to round-off errors) both the total energy and angular momentum. More recently discrete gradient methods have been extended and developed in the context of geometric numerical integration [6]. Quispel and his coworkers constructed numerical integrators preserving all integrals of motion of a given system of ordinary differential equations [7, 8, 9, 10].

In general, geometric numerical integrators are very good in preserving qualitative features of simulated differential equations. It is important to enhance their accuracy without losing their geometric properties. Symplectic algorithms can be improved using appropriate splittings and composition of low order methods [11, 12, 13, 14, 15, 16]. Some other methods, like exponential/trigonometric fitting [17, 18, 19] and linearization-preserving preprocessing (locally exact modification) [20], are available as well. Energy-preserving methods include discrete gradient schemes [1, 3, 10], projection techniques [5, 21, 22], Hamiltonian boundary value methods [23, 24, 25] and collocation methods [26, 27].

Our research is concentrated on improving the efficiency of the discrete gradient method (which is energy-preserving but is not symplectic) without loosing its outstanding qualitative advantages. Results reported in earlier papers are very promising [28, 29]. In this paper we present a further essential improvement of this approach by constructing energy-preserving discrete gradient schemes of any prescribed order $N$ for one-dimensional Hamiltonian systems of the form:

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

where $V(x)$ is a potential, and the dot and the prime denote differentiation with respect to $t$ and $x$, respectively. In this case the discrete gradient method reduces to the so called modified midpoint rule:

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

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

$$2$$
where \( h \) is the time step. Our modification consists in replacing \( h \) by an appropriate function \( \delta_n \).

## 2 Explicit Taylor schemes of \( N \)th order

In this section we consider explicit numerical schemes of any order, derived by standard Taylor expansions (compare, e.g., [30], p. 18). There are at least two reasons for considering here the Taylor schemes. First, the expansion coefficients will be useful in the construction of discrete gradient schemes of high order. Then, the Taylor schemes can be used for comparison with discrete gradient methods of similar order. Moreover, they are good candidates for predictors when modified gradient methods are used as correctors.

In order to obtain explicit schemes of arbitrary high order we expand \( x(t+h) \) and \( p(t+h) \) in Taylor series:

\[
x(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \frac{d^k x(t)}{dt^k}, \quad p(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \frac{d^k p(t)}{dt^k},
\]

(3)

where all derivatives can be replaced by functions of \( x, p \) using (1) and its differential consequences (e.g., \( \dddot{p} = -V''(x) \dot{x} = -V''(x)p \)). Thus we get

\[
x(t+h) = x + ph - \frac{1}{2}V'h^2 - \frac{1}{6}pV''h^3 + \frac{1}{24} (V'V'' - V'''p^2) h^4 + O(h^5)
\]

\[
p(t+h) = p - V'h - \frac{1}{2}pV''h^2 + \frac{1}{6} (V'V'' - V'''p^2) h^3
\]

\[
+ \frac{1}{24} (3pV'V''' + p(V'')^2 - p^3V^{(4)}) h^4 + O(h^5).
\]

The Taylor expansion can be represented in the form

\[
x(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} b_k(x, p), \quad p(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} c_k(x, p),
\]

(5)

where \( b_k = \frac{d^k}{dx^k}x, c_k = \frac{d^k}{dx^k}p \) and we compute these derivative using (1). For instance, \( b_0 = x, b_1 = \dot{x} = p \) and \( b_2 = \ddot{x} = \dot{p} = -V'(x). \) In general,

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

(6)

Then, \( p = \dot{x} \) implies

\[
c_k = \frac{d}{dt} b_k = b_{k+1}.
\]

(7)
The coefficients $b_k$ ($k = 1, 2, \ldots, 11$), computed recurrently from (6), read

$$b_0 = x, \quad b_1 = p, \quad b_2 = -V_x, \quad b_3 = -pV_{xx},$$

$$b_4 = V_x V_{xx} - p^2 V_{xxx},$$

$$b_5 = p(V_x^2 + 3V_x V_{xxx}) - p^3 V_{4x},$$

$$b_6 = -3V_x^2 V_{xxx} - V_x^2 V_{xx} + p^2(5V_x V_{xxx} + 6V_x V_{4x}) - p^4 V_{5x},$$

$$b_7 = -p(V_x^3 + 18V_x^2 V_{xx} V_{xxx} + 15V_x^2 V_{4x})$$

$$+ p^3(5V_{xxx} + 11V_x V_{4x} + 10V_x V_{5x}) - p^5 V_{6x},$$

$$b_8 = V_x V_{xx}^3 + 18V_x^2 V_{xx} V_{xxx} + 15V_x^3 V_{4x}$$

$$- p^2(21V_{xx}^2 V_{xxx} + 33V_x^2 V_{xxx} + 81V_x V_{xx} V_{4x} + 45V_x^2 V_{5x})$$

$$+ p^4(21V_{3x} V_{4x} + 21V_{xx} V_{5x} + 15V_x V_{6x}) - p^6 V_{7x},$$

$$b_9 = p(V_{xx}^4 + 81V_x^2 V_{xx} V_{3x} + 84V_x^2 V_{3x}^2 + 225V_x^2 V_{xx} V_{4x} + 105V_x^3 V_{5x})$$

$$- p^3(75V_{xx}^2 V_{3x}^2 + 102V_x^2 V_{xx} V_{4x} + 231V_x V_{3x} V_{4x} + 255V_x V_{xx} V_{5x}$$

$$+ 105V_x^2 V_{6x}) + p^5(21V_{4x}^2 + 42V_{3x} V_{5x} + 36V_{xx} V_{6x} + 21V_x V_{7x})$$

$$- p^7 V_{8x},$$

$$b_{10} = -(V_x V_{xx}^4 + 81V_x^2 V_{xx}^2 V_{3x} + 84V_x^2 V_{3x}^2 + 225V_x^3 V_{xx} V_{4x}$$

$$+ 105V_x^4 V_{5x}) + p^2(85V_{xx}^3 V_{3x} + 555V_x V_{xx} V_{3x} + 837V_x^2 V_{xx} V_{4x}$$

$$+ 1086V_x V_{3x} V_{4x} + 1305V_x^2 V_{xx} V_{5x} + 420V_x^3 V_{6x})$$

$$- p^4(75V_{3x}^3 + 585V_{xx} V_{3x} V_{4x} + 336V_x V_{4x}^2 + 357V_x^2 V_{5x}$$

$$+ 696V_x V_{3x} V_{5x} + 645V_x V_{xx} V_{6x} + 210V_{xx} V_{7x}) + p^6(84V_{4x} V_{5x}$$

$$+ 78V_{3x}^2 V_{6x} + 57V_{xx} V_{7x} + 28V_x V_{8x}) - p^8 V_{9x},$$

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

Finally, for any fixed \( N \) we define the Taylor scheme of \( N \)th order (shortly: TAY-\( N \)) as follows,
\[ x_{n+1} = \sum_{k=0}^{N} \frac{h^k}{k!} b_k(x_n, p_n), \quad p_{n+1} = \sum_{k=0}^{N} \frac{h^k}{k!} c_k(x_n, p_n), \] (13)
where \( b_k \) and \( c_k \) are defined by (6), (7) and, in particular cases, by (8), (9), (10), (11) and (12).

### 3 Discrete gradient schemes of \( N \)th order

We consider the following family of nonstandard numerical schemes (parameterized by a single function \( \delta_n \)):
\[ \frac{x_{n+1} - x_n}{\delta_n} = \frac{1}{2} (p_{n+1} + p_n) \],
\[ \frac{p_{n+1} - p_n}{\delta_n} = -\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n} \],
where \( \delta_n \) can depend on any variables and parameters, including \( h, x_n, p_n, x_{n+1}, p_{n+1} \). It is enough to assume the consistency condition,
\[ \lim_{h \to 0} \frac{\delta_n(h)}{h} = 1 \].

5
One can easily prove that the total energy is preserved, i.e.,

$$\frac{1}{2}p^2_n + V(x_n) = E = \text{const},$$

(16)

for any choice of the function $\delta_n$. This is an essential generalization of the well known case $\delta_n(h) = h$.

In our recent papers [28, 29, 31] we consider $\delta_n$ of the form

$$\delta_n = \frac{2}{\omega_n} \tan \frac{h\omega_n}{2}, \quad \omega_n = \sqrt{V''(\bar{x}_n)},$$

(17)

where $\bar{x}_n$ may depend on $x_n, x_{n+1}$ but usually does not depend on $h$. Taking $\bar{x}_n = x_0$, where $V'(x_0) = 0$, we get the modified discrete gradient scheme (MOD-GR) [28]. Then, $\bar{x}_n = x_n$ and $\bar{x}_n = \frac{1}{2}(x_n + x_{n+1})$ yield locally exact discrete gradient scheme (GR-LEX) and its symmetric modification (GR-SLEX), see [29, 31]. These three numerical methods are of second, third, and fourth order, respectively [29]. Locally exact (“linearization-preserving”) symplectic schemes of fourth order have been recently considered in [20].

In this paper we will show that the family of numerical integrators of the form (14) contains numerical schemes of any order provided that $V(x)$ is smooth enough. The explicit formulae will be presented up to the order 11.

The system (14) (where $x_n \equiv x$ and $p_n \equiv p$ are given and $\delta_n \equiv \delta$ is a small parameter) implicitly defines $x_{n+1}$ and $p_{n+1}$. Therefore, using implicit differentiation, we can write down the corresponding Taylor series:

$$x_{n+1} = x + p\delta - \frac{1}{2}V'\delta^2 - \frac{1}{4}pV''\delta^3 + \frac{1}{24} (3V'V'' - 2p^2V''')\delta^4 + O(\delta^5),$$

$$p_{n+1} = p - V'\delta - \frac{1}{2}pV''\delta^2 + \frac{1}{12} (3V'V'' - 2V'''p^2)\delta^3$$

$$- \frac{1}{24} (4pV''' + 3p(V'')^2 - p^3V^{(4)})\delta^4 + O(\delta^5),$$

(18)

and then, assuming an analytical dependence of $\delta$ on $h$ and choosing appropriately coefficients of the Taylor expansion of $\delta$, we can assure that $x_{n+1}$, $p_{n+1}$ coincide with the exact solution up to a prescribed order $N$.

In this paper we present another way of approaching this problem. We define

$$\delta_{(\infty)} = \frac{2(x(t+h) - x_n)}{p(t+h) + p_n}, \quad x(t) = x_n, \quad p(t) = p_n.$$  

(19)
Taking into account the energy conservation law, one can easily see that the exact solution satisfies (14) with $\delta_n$ given by (19), i.e.,

$$\frac{x(t+h) - x_n}{\delta_{[\infty]}} = \frac{1}{2} (p(t+h) + p_n) ,$$

$$\frac{p(t+h) - p_n}{\delta_{[\infty]}} = -\frac{V(x(t+h)) - V(x_n)}{x(t+h) - x_n} .$$

(20)

We verify that

$$\lim_{h \to 0} \frac{\delta_{[\infty]}}{h} = \lim_{h \to 0} \frac{2(x(t+h) - x_n)}{h(p(t+h) + p_n)} = \frac{\dot{x}(t)}{p(t)} = 1 ,$$

(21)

because $p(t) = p_n$. Using (5) we can compute the Taylor expansion of $\delta_{[\infty]}$: $\delta_{[\infty]} = \delta_{[\infty]}(x, p, h) = \sum_{k=1}^{\infty} a_k(x, p) h^k$

(22)

(for more details see Section 4). Denoting differentiation with respect to $x$ by a subscript (and using abbreviations like $V_{4x} \equiv V_{xxxx}$) we present a number of coefficients $a_k$ in an explicit form:

$$a_1 = 1 , \quad a_2 = 0 , \quad a_3 = \frac{1}{12} V_{xx} , \quad a_4 = \frac{1}{24} pV_{xxx} ,$$

(23)

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

(24)

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

(25)

$$a_7 = \frac{1}{20160} \left( a_{7,0} + a_{7,2}p^2 + a_{7,4}p^4 \right) ,$$

(26)

$$a_8 = \frac{1}{40320} \left( a_{8,1}p + a_{8,3}p^3 + a_{8,5}p^5 \right) ,$$

$$a_9 = \frac{1}{725760} \left( a_{9,0} + a_{9,2}p^2 + a_{9,4}p^4 + a_{9,6}p^6 \right) ,$$

(27)

$$a_{10} = \frac{1}{7257600} \left( a_{10,1}p + a_{10,3}p^3 + a_{10,5}p^5 + a_{10,7}p^7 \right) ,$$

$$a_{11} = \frac{1}{159667200} \left( a_{11,0} + a_{11,2}p^2 + a_{11,4}p^4 + a_{11,6}p^6 + a_{11,8}p^8 \right) ,$$

7
where the coefficients $a_{j,k}$ depend on $x$ through derivatives of $V$, namely:

\[
\begin{align*}
a_{7,0} &= 17V_{xx}^3 + 45V_x^2V_{4x} - 44V_xV_{xx}V_{xxx}, \\
a_{7,2} &= 20V_{xxx}^2 - 12V_{xx}V_{4x} - 72V_xV_{5x}, \\
a_{7,4} &= 10V_{6x}, \\
a_{8,1} &= 21V_{xx}^2V_{xxx} - 42V_xV_{xx}^2 + 63V_x^2V_{5x}, \\
a_{8,3} &= 14V_{xxx}V_{4x} - 21V_{xx}V_{5x} - 35V_xV_{6x}, \\
a_{8,5} &= 3V_{7x}, \\
a_{9,0} &= 62V_{xx}^4 - 228V_xV_{xx}^2V_{xxx} + 168(V_x^2V_{xxx}^2 - V_x^3V_{5x}) \\
&
+ 90V_x^2V_{xx}V_{4x}, \\
a_{9,2} &= 75V_{xx}V_{xx}^2 + 81V_x^2V_{4x} - 462V_xV_{xxx}V_{4x} \\
&
+ 360V_x^2V_{xx}V_{5x} + 420V_x^2V_{6x}, \\
a_{9,4} &= 42V_{4x}^2 - 120(V_{xx}V_{6x} + V_xV_{7x}), \\
a_{9,6} &= 7V_{8x}, \\
a_{10,1} &= 460V_{xx}^3V_{xxx} - 1170V_xV_{xx}V_{xx}V_{xxx} - 630V_xV_{xx}^2V_{4x} \\
&
+ 2385V_x^2V_{xxx}V_{4x} - 945V_x^2V_{xx}V_{5x} - 1260V_x^3V_{6x}, \\
a_{10,3} &= 150V_{xx}^3 + 15V_xV_{xx}V_{xxx}V_{4x} - 945V_x^2V_{4x} - 456V_xV_{xxx}V_{5x} \\
&
+ 483V_x^2V_{5x} + 1785V_xV_{xx}V_{6x} + 1080V_x^2V_{7x}, \\
a_{10,5} &= 126V_{4x}V_{5x} - 114V_{3x}V_{6x} - 261V_{xx}V_{7x} - 189V_xV_{8x}, \\
a_{10,7} &= 8V_{9x},
\end{align*}
\]
\[ a_{11,0} = 1382V_{xx}^5 - 6448V_xV_{xx}^3V_{3x} + 4140V_x^2V_{xx}^2V_{4x} \]
\[ + 840V_x^3V_{xxx}V_{5x} + 8280V_x^3V_{3x}V_{4x} + 7368V_x^2V_{xx}V_{3x}^2 \]
\[ + 3150V_x^2V_{6x} , \]
\[ a_{11,2} = 3240V_{xx}^2V_{3x}^2 - 6480V_xV_{3x}^3 + 696V_x^3V_{4x} \]
\[ + 4872V_xV_{xxx}V_{4x} + 15660V_x^2V_{4x}^2 - 9144V_xV_{xx}V_{5x} \]
\[ + 11988V_x^2V_{3x}V_{5x} - 21000V_x^2V_{xx}V_{6x} - 10800V_x^3V_{7x} , \]
\[ a_{11,4} = 1710V_{3x}^2V_{4x} - 1803V_{xxx}V_{4x}^2 - 8676V_xV_{4x}V_{5x} \]
\[ + 1260V_xV_{xxx}V_{5x} + 4770V_{xx}^2V_{6x} + 3060V_xV_{3x}V_{6x} \]
\[ + 11400V_xV_{xxx}V_{7x} + 4725V_x^2V_{8x} , \]
\[ a_{11,6} = 336V_{5x}^2 - 780V_{xxx}V_{7x} - 980V_{xxx}V_{6x} - 560V_xV_{9x} \]
\[ + 120V_xV_{6x} , \]
\[ a_{11,8} = 18V_{10x} . \]

We observe that coefficients \( a_k = a_k(x, p) \) are polynomials with respect to \( p \).

In Section 4 we will show that this property holds for any \( k \).

The exact integrator (20) is only a formal construction because \( \delta[\infty] \) is given in a form an infinite series. In order to obtain a practical numerical method we have to approximate \( \delta[\infty] \) by \( h \)-polynomials

\[
\delta[N] = \delta[N](x, p, h) = \sum_{k=1}^{N} a_k(x, p)h^k = h + \sum_{k=3}^{N} a_k(x, p)h^k .
\]

Thus we propose to consider numerical schemes of the form (14), where \( \delta_n = \delta[N](x_n, p_n, h) \) is defined by (33), i.e.,

\[
\frac{x_{n+1} - x_n}{\delta[N]} = \frac{1}{2} (p_{n+1} + p_n) .
\]

\[
\frac{p_{n+1} - p_n}{\delta[N]} = -\frac{V(x_{n+1}) - V(x_n)}{x_{n+1} - x_n} .
\]

The scheme (34) will be referred to as GR-\( N \). In Section 4 we will show that its order is at least \( N \).
4 Theoretical analysis of gradient schemes

The recurrence (6) implies (by mathematical induction) that coefficients $b_k = b_k(x, p)$ are polynomials with respect to $p$ and derivatives of $V$. What is more,

$$b_k(x, 0) = \hat{b}_k(x)V_x,$$

where, by (6), $\hat{b}_k$ is a polynomial with respect to derivatives of $V$ (for any $k > 1$). Actually, $\hat{b}_k(x) \equiv 0$ for $k$ odd. Coefficients $a_k$ can be expressed in terms of $b_j$ using (5), (7) and (19), i.e.,

$$\sum_{k=1}^{\infty} a_k(x, p)h^k \equiv \delta_{[\infty]}(x, p, h) = \frac{2\sum_{k=1}^{\infty} \frac{h^k}{k!} b_k(x, p)}{2p + \sum_{k=1}^{\infty} \frac{h^k}{k!} b_{k+1}(x, p)}. \quad (36)$$

From (36) we immediately obtain

$$2\sum_{j=1}^{\infty} \frac{b_j h^j}{j!} = 2p \sum_{j=1}^{\infty} a_j h^j + \sum_{j=2}^{\infty} h^j \sum_{k=1}^{j-1} \frac{b_{k+1} a_{j-k}}{k!}. \quad (37)$$

Equating coefficients by powers of $h$ we get $pa_1 = b_1$ and the following recursion formula for $j \geq 2$:

$$pa_j = \frac{b_j}{j!} - \frac{1}{2} \sum_{k=1}^{j-1} b_{j-k+1} a_k \frac{1}{(j-k)!}. \quad (38)$$

**Proposition 1.** Coefficients $a_k$ are polynomials with respect to $p$ with coefficients depending on $x$ through derivatives of $V$.

**Proof:** Using mathematical induction, we can show that $a_j$ is a polynomial with respect to $p$ and $p^{-1}$, namely

$$a_j = \sum_{k=-j}^{K} a_{j,k} p^k, \quad (39)$$

where $K \in \mathbb{N}$ and $a_{j,k} = a_{j,k}(x)$ are polynomials with respect to derivatives of $V$. 

10
In order to show that $a_{j,k} = 0$ for $k < 0$ (which means that $\delta_{[\infty]}(x, p, h)$ is analytic at $p = 0$) it is enough to evaluate $\delta_{[\infty]}(x, p, h)$ at $p = 0$. From (36) and (35) we have

$$\delta_{[\infty]}(x, 0, h) = h \left( 1 - 2 \sum_{j=1}^{\infty} \frac{\hat{b}_{2j+2}h^{2j}}{(2j+2)!} \right),$$

or, more explicitly,

$$\delta_{[\infty]}(x, 0, h) = h \left( 1 - \frac{1}{12} V_{xx} h^2 + \frac{1}{360} (V_{xx}^2 + 3V_x V_{xxx}) h^4 + \ldots \right).$$

The finite value of $\delta_{[\infty]}(x, 0, h)$ means that $a_{j,k}$ have to vanish for all $k < 0$.

Proposition 2. The system (34) has a unique solution $(x_{n+1}, p_{n+1})$ provided that $h$ is sufficiently small.

Proof: It is enough to show that the $(x, p) \mapsto (\bar{x}, \bar{p})$ given by

$$\bar{x} = x_n + \frac{1}{2} (p_n + p) \delta_{[N]}(x_n, p_n, h),$$

$$\bar{p} = p_n - \delta_{[N]}(x_n, p_n, h) \frac{V(x) - V(x_n)}{x - x_n},$$

is a contraction in some compact domain. By (23) we have $\delta_{[N]}(x, p, h) = h + O(h^3)$. Therefore, given $x_n, p_n$ there exists $h_1$ such that for $h < h_1$

$$|\delta_{[N]}(x_n, p_n, h)| < 2h.$$ (43)

It is convenient to define

$$\bar{F}(x) := -\frac{V(x) - V(x_n)}{x - x_n}, \quad \bar{F}'(x_n) := -V'(x_n).$$ (44)

We choose any $R > 0$ and denote

$$M = \sup_{x \in B(x_n, R)} |\bar{F}(x)|, \quad M_1 = \sup_{x \in B(x_n, R)} |\bar{F}'(x)|.$$ (45)
where $B(x_n, R)$ is a ball centered at $x_n$ with radius $R$, and, obviously, $M < \infty$ and $M_1 < \infty$. The map (42) satisfies

\[
|\tilde{x} - x_n| \leq \frac{1}{2} |\delta[N]| |p_n + p| < h(2|p_n| + |p - p_n|),
\]

\[
|\tilde{p} - p_n| \leq |\delta[N]| M < 2hM.
\]

Therefore, requiring

\[2h|p_n| + h|p - p_n| \leq R, \quad |p - p_n| \leq 2hM,
\]

we get a map $D \to D$, where $D = B(x_n, R) \times B(p_n, 2hM)$. Conditions (47) are equivalent to

\[
\frac{|p - p_n|}{2M} \leq h \leq \frac{R}{2p_n + |p - p_n|},
\]

which can be satisfied provided that $|p - p_n|$ is small enough, namely

\[
|p - p_n| \leq \sqrt{p_n^2 + 2MR} - |p_n| \equiv \frac{2MR}{|p_n| + \sqrt{p_n^2 + 2MR}}.
\]

Therefore, (47) hold for $h \leq h_2$, where

\[
h_2 = \frac{R}{|p_n| + \sqrt{p_n^2 + 2MR}}.
\]

We will show that the map (42) is a contraction. Given $x_n, p_n$ we apply this map to $x, p$ and $x', p'$ (here the prime is not a derivative), and then we compute:

\[
|\tilde{x}' - \tilde{x}| = \frac{1}{2} |\delta[N]| |p' - p| \leq h|p' - p|,
\]

\[
|\tilde{p}' - \tilde{p}| = |\delta[N]| |F'(\xi)||x' - x| \leq 2hM_1|x' - x|,
\]

where $\xi$ lays between $x$ and $x'$ (we used (43), (45) and Lagrange’s mean value theorem). Therefore, choosing any $L < 1$, we have

\[
\sqrt{(\tilde{x}' - \tilde{x})^2 + (\tilde{p}' - \tilde{p})^2} \leq L \sqrt{(x' - x)^2 + (p' - p)^2}
\]

for $h$ sufficiently small (i.e., $h < L < 1$ and $2hM_1 < L$). Finally, using the Banach fixed point theorem we see that for

\[
h \leq \min \left\{ h_1, h_2, L, \frac{L}{2M_1} \right\}
\]
the system (34) has a unique solution which can be approximated numerically by iterating the map (42).

Proposition 2 assumes that the right-hand side of (42) is evaluated exactly. Unavoidable round-off errors introduce a systematic inaccuracy which cannot be improved by repeating iterations (see, e.g., [32]).

**Proposition 3.** The scheme GR-N is of (at least) $N$th order.

**Proof:** We denote local errors by $\Delta x_n$ and $\Delta p_n$, respectively, i.e.,

$$\Delta x_n = x_{n+1} - x(t + h), \quad \Delta p_n = p_{n+1} - p(t + h),$$

where we took into account $x_n = x(t)$. Moreover, we define

$$r_N := \delta_{[\infty]} - \delta_{[N]} = O(h^{N+1}).$$

We use (20) and (34):

$$x_{n+1} - x_n = \frac{1}{2} \delta_{[N]}(p_{n+1} + p_n),$$

$$x(t + h) - x_n = \frac{1}{2} \delta_{[\infty]} (p(t + h) + p_n),$$

$$p_{n+1} - p_n = \delta_{[N]} \bar{F}(x_{n+1}),$$

$$p(t + h) - p_n = \delta_{[\infty]} \bar{F}(x(t + h)).$$

Hence, we immediately get

$$\Delta x_n = - \frac{1}{2} r_n (p(t + h) + p_n) + \frac{1}{2} \delta_{[N]} \Delta p_n,$$

$$\Delta p_n = \delta_{[N]} (\bar{F}(x_{n+1} - \bar{F}(x(t + h))) + r_n \bar{F}(x(t + h))$$

By Lagrange’s mean value theorem there exists $\xi$ such that

$$\bar{F}(x_{n+1}) - \bar{F}(x(t + h)) = \bar{F}'(\xi) \Delta x_n.$$  

Therefore, solving (57) with respect to $\Delta x_n$ and $\Delta p_n$,

$$\left(1 - \frac{1}{2} \bar{F}'(\xi) \delta_{[N]}^2\right) \Delta x_n = - \frac{1}{2} r_n (p(t + h) + p_n - \bar{F}(x(t + h)) \delta_{[N]}),$$

$$\left(1 - \frac{1}{2} \bar{F}'(\xi) \delta_{[N]}^2\right) \Delta p_n = - \frac{1}{2} r_n ((p(t + h) + p_n) \bar{F}'(\xi) \delta_{[N]} - 2 \bar{F}(x(t + h))).$$
By virtue of (43) and (45) we have \( F'(\xi)\delta^2_{[N]} < 1 \) (for \( h \) sufficiently small). We complete the proof taking into account (55). 

We point out that GR-1 and GR-2 coincide with the discrete gradient scheme (GR) given by (2) (in particular, GR-1 is of 2nd order). In the case of \( V(x) \) linear in \( x \) any method GR-N is exact (i.e., its order becomes infinite).

5 Numerical experiments

In our recent papers we compared several discretizations of the simple pendulum equation \( (V(x) = -\cos x) \) with a special stress on the long-time behaviour, see [28, 29, 33]. Locally exact discrete gradient schemes (GR-LEX and GR-SLEX) turned out to be the best. In some tests their accuracy was better by several orders of magnitude in comparison to standard methods like leap-frog, implicit midpoint rule or the discrete gradient method (GR). GR-LEX and GR-SLEX yield similar results and in this section we confine ourselves to GR-LEX only.

We are going to compare GR-LEX with algorithms of higher order introduced in the present paper, i.e., GR-N and TAY-N. The accuracy of these schemes was tested mainly for the simple pendulum, but other potentials yield similar results. We present some data for the Morse potential, \( V(x) = \frac{1}{2}e^{-2x} - e^{-x} \). For simplicity we always assume the initial position at the stable equilibrium, i.e., \( x_0 = 0 \).

In both cases exact solutions can be expressed in terms of elliptic functions. The exact solutions usually were computed with the accuracy of at least 14 digits. The accuracy is lower only in the close neighbourhood of the separatrix (for \( p_0 = 1.99999 \) we have the accuracy of 12 digits). We point out that the period of the exact solution was computed with a higher accuracy (16 and 15 digits, respectively).

5.1 Global error

Fig. 1, Fig. 2, Fig. 3 and Fig. 4 show the dependence of the global error of the numerical solutions on the time step (the global error was evaluated at \( t = 120T_{ex} \), where \( T_{ex} \) denotes the period of the corresponding exact solution). Some points at Fig. 1 and Fig. 3 form straight lines. These lines are presented in more detail at Fig. 2 and Fig. 4.
GR-3 yields almost the same results as GR-LEX. They are better than GR by several orders of magnitude. GR-N (for $N \geq 5$) are more accurate than GR-LEX by several orders of magnitude. We point out that schemes GR-N are very accurate also for large time steps. In particular, TAY-10 becomes less accurate than GR-7 and TAY-5 for larger $h$.

For small time steps (say, $h < 0.1$) the accuracy of the best discretizations (i.e., GR-7, TAY-10 and GR-11) apparently does not depend on $h$ (actually, it even slightly decreases for smaller $h$). This effect is caused mainly by round-off errors. The accuracy of our computations is not higher than 14 digits and decreases for smaller time steps. The points influenced by this round-off errors are removed from Fig. 2 and Fig. 4, where we left only linear parts of every graph. A slope of every line is closely related to the order of the corresponding method (one can interpret it as an experimental order).

Theoretically all gradient schemes (14) preserve exactly the energy but, of course, round-off errors cause some small inaccuracy, see Fig. 5. We see that the energy error accumulates slowly, almost linearly, with a very very small slope: for $t \approx 300\,000$ we have $\Delta E \approx 10^{-12}$.

### 5.2 Stability and relative error of the period

All gradient schemes have extremely stable period of oscillations. Estimating the numerical value of the period we perform averaging over some number of periods. Thus we define

$$T_{\text{avg}}(N, M) = \frac{1}{M} \left( z_{N+2M} - z_N \right),$$

where $M, N$ are given integers and $z_1, z_2, z_3, \ldots$ are numerical estimates of subsequent zeros of the solution $x(t)$ (we compute them as zeros of interpolating cubic polynomials based on appropriate four neighbouring points), for more details see [28]. The dependence of $T_{\text{avg}}$ on $\varepsilon$ and $p_0$ is omitted for the sake of brevity. Considering very long discrete evolutions (many thousands of periods) we often use another definition of the average period, namely

$$\bar{T}_{\text{avg}}(N, K, L) = \frac{1}{L - K} \sum_{M=K+1}^{L} T_{\text{avg}}(N, M),$$

where $K < M \leq L$. The stability of the discrete gradient scheme (GR) was tested in detail in [28]. Other gradient schemes follow the same pattern.
Fig. 6 compares the average period (more precisely: $T_{\text{avg}}(N, 20)$, see [28]) of numerical solutions produced by GR-7 and TAY-10. If $t$ is not very large, then in both cases the average period oscillates around the exact value $T_{\text{ex}}$. For longer times we clearly see that TAY-10 becomes less and less exact, see Fig. 6, while GR-7 oscillates exactly in the same way, even for very very long times, e.g., $t \approx 30\ 000\ 000$ at Fig. 7.

Fig. 8 and Fig. 9 illustrate the relative error of the period. More precisely, we consider $\bar{T}_{\text{avg}}(0, 100, 200)$. Then, we compare the results with the exact period $T_{\text{ex}}$ (computed as an elliptic integral with an accuracy at least 15 digits). Fig. 8 presents the dependence of the relative period on the time step $h$. We see that GR-7 yields excellent results (better by 3-5 orders of magnitude than GR-LEX) and the accuracy of TAY-10 and GR-11 is almost the same (for $p_0 = 1.95$ and $h < 0.3$). It seems that the accuracy of the period is strongly influenced by the interpolation method of determining zeros, see (60). This effect becomes dominant in the case of the most accurate schemes. The accuracy of Taylor schemes becomes relatively worse for greater $h$.

Similar conclusions can be derived from Fig. 9. Increasing the order of GR-N we increase the accuracy but only to some extent. Indeed, for $h = 0.02$ schemes GR-11 and TAY-10 yield practically the same accuracy as GR-7, i.e., $10^{-13}$ for oscillating motions and $10^{-9}$ for rotating motions, with exception of the region $p_0 \approx 2$, where the accuracy is lower for any numerical scheme. For $p_0 < 2$ (oscillations) GR-7 is more accurate than GR by 7-9 orders of magnitude. For small $p_0$ also GR-LEX and TAY-5 attain such high accuracy. For $p_0 > 2$ the scheme GR-LEX produces almost the same results as GR-3, and both are less accurate than TAY-5 (GR-LEX is more and more accurate for $p_0$ tending to 0). We point out that gradient schemes produce very stable results (i.e., the picture presented at Fig. 9 is time-independent for gradient schemes). The accuracy of Taylor schemes decreases with time, see Fig. 6.

5.3 Neighbourhood of the separatrix

The neighbourhood of the separatrix ($p_0 \approx 2$ for the simple pendulum) is most difficult to be simulated numerically. The discrete gradient method (GR) turns out to be relatively good in this region, see [28], and the locally exact methods (GR-LEX, GR-SLEX) work almost perfectly [29]. Here, we take for comparison also GR-3, GR-7, GR-11 and TAY-5, TAY-10. The Taylor schemes are much worse in this region: for $h = 0.9$ even TAY-10 is not able to reproduce the correct qualitative behaviour, see Fig. 10. Through-
out the first period the scheme GR yields good qualitative behaviour and is better than TAY-10 with a halved time step, see Fig. 10. In the first period GR-3, GR-7 and GR-LEX (and also GR-SLEX and GR-N for $N > 3$) produce similar results. We point out that the exact trajectory is very close to the separatrix ($|p_0 - 2| = 10^{-10}$) and $h$ is very large but, nevertheless, all improved discrete gradient methods simulate very accurately the motion of the pendulum.

Fig. 11 shows the same situation for much longer times ($t > 100 000$). Note that the time step for TAY-10 is much smaller ($h = 0.09$) than the time step for all gradient schemes ($h = 0.9$). In spite of that essential handicap, TAY-10 is only slightly better than GR-7 and less accurate than GR-11. GR-7 is more accurate than GR-LEX.

### 5.4 Computational cost

All gradient schemes are implicit. We apply the fixed point method, see (42), or the Newton method (their computational costs are rather similar), and iterate until $|x_{n+1} - x_n| < 10^{-16}$ is obtained. The obtained accuracy is smaller (14 digits) due to round-off errors. We point out that $\delta_N$ given by (33) depends on $x_n, p_n$ and does not depend on $x_{n+1}, p_{n+1}$. It means that $\delta_N$ is evaluated outside iteration loops, only once at every step. Thus the computational cost of high-order schemes GR-$N$ is not much higher than the cost of the standard discrete gradient method GR (at least for potentials possessing explicit derivatives in a simple form, like the simple pendulum, but this is quite frequent in physical applications). In the case of the simple pendulum the cost of GR-11 is only about two times higher than the cost of the scheme GR. This factor depends on the average number of iterations per step (in the case of higher number of iterations the costs of GR and GR-11 become more similar), compare [33]. The “worst” situation is in the case of one iteration per step, when the scheme GR is 3 times cheaper than GR-11. Therefore, schemes GR-$N$ have the computational cost of the same order as GR, but their accuracy is higher by many orders of magnitude. Fig. 12 presents the number of function evaluations needed to obtain a required accuracy at $t = 50T_{ex}$. In spite of cumbersome formulas schemes TAY-10 and GR-11 have the lowest cost. We point out, however, that the accuracy of TAY-10 constantly decreases with time in comparison to the very stable gradient schemes, see (6).
6 Conclusions

The numerical integrators GR-\(N\), described in this paper, have similar advantages as GR-LEX and GR-SLEX: they preserve exactly the energy integral (i.e., eq. (16) holds) and exact trajectories in the phase space, are extremely stable and have very good long-time behaviour of numerical solutions. They can be constructed for any prescribed order \(N\). In spite of cumbersome formulas the computational cost of high-order gradient methods GR-\(N\) is of the same order as the costs of the standard discrete gradient method GR (provided that the derivatives of potential are relatively simple functions, given in an explicit form).

Therefore, modifications presented in this paper essentially improve the discrete gradient method (at least in the one-dimensional case) keeping all its advantages. Schemes GR-\(N\) (for \(N \geq 7\)) are much more accurate than GR-LEX for most choices of parameters. Only in the region of small \(p_0\) the scheme GR-LEX is comparable with discrete gradient methods of high order.

Numerical schemes (34), like all discrete gradient methods, are neither symplectic nor volume-preserving. Moreover, schemes GR-\(N\) are not time-reversible. Therefore, the conservation of the energy integral plus high order seem to be sufficient to assure outstanding qualitative and quantitative properties of these methods.

Locally exact schemes GR-LEX and GR-SLEX have been recently constructed for multidimensional Hamiltonian systems [34]. In near future we hope to extend results of the present paper on multidimensional Hamiltonian systems and differential equations preserving integrals of motion.

Acknowledgments. We gratefully acknowledge constructive remarks and suggestions of anonymous referees. This research work has been supported by the grant No. N N202 238637 from the Polish Ministry of Science and Higher Education.

References


Figure 1: Global error at $t = 120 \ T_{ex}$ as a function of the time step $h$ for the simple pendulum, $p_0 = 1.8 \ (T_{ex} = 9, 12 \ 219 \ 655 \ 369)$. 
Figure 2: Global error at $t = 120 \, T_{ex}$ as a function of the time step $h$ for the simple pendulum, $p_0 = 1.8$: linear dependence on the logarithmic scale.
Figure 3: Global error at $t = 120 \ T_{ex}$ as a function of the time step $h$ for the Morse potential, for $p_0 = 0.8$ ($T_{ex} = \text{10,4719755}\ 119$).
Figure 4: Global error at $t = 120 \ T_{ex}$ as a function of the time step $h$ for the Morse potential, for $p_0 = 0.8$: linear dependence on the logarithmic scale.
Figure 5: Energy error as a function of time ($t = hN$, $h = 0.25$), for the simple pendulum, $p_0 = 1.8$ ($E_{ex} = 0.62$).
Figure 6: Average period as a function of time ($N$ is a number of half-periods) for the simple pendulum, $p_0 = 1.8$ ($T_{ex} = 9, 12, 219, 655, 369$). Dark points – GR-7, light points – TAY-10, solid straight line – exact period.
Figure 7: Average period as a function of time ($N$ is a number of half-periods) for the simple pendulum, $p_0 = 1.8$, scheme GR-11. Solid straight line – exact period ($T_{ex} = 9, 12 \ 219 \ 655 \ 369$).
Figure 8: Relative error of the period of the simple pendulum as a function of $h$, for $p_0 = 1.95$ ($T_{ex} = 11,6575852844$).
Figure 9: Relative error of the period for the simple pendulum as a function of $p_0$, for $h = 0.02$. Results produced by GR-7, GR-11 and TAY-10 are practically the same. For $p_0 > 2$ GR-LEX and GR-3 yield similar results.
Figure 10: $x_n$ as a function of time ($t = nh$), very near the separatrix ($p_0 = 1.999\,999\,999\,9$), $h = 0.09$ for TAY-5, $h = 0.45$ for TAY-10, $h = 0.9$ for all other discretizations. The solid line corresponds to the exact solution ($T_{ex} = 51.596\,879$).
Figure 11: $x_n$ as a function of time ($t = nh$), very near the separatrix ($p_0 = 1.999\ 999\ 999\ 9$), $h = 0.09$ for TAY-10 and $h = 0.9$ for all other discretizations. The solid line corresponds to the exact solution ($T_{ex} = 51.596\ 879$).
Figure 12: Computational cost of chosen methods. Number of function evaluations ($N$) needed to reach a given global error. The global error is computed at $t = 50 \ T_{ex}$ for the simple pendulum, $p_0 = 1.8$, $T_{ex} = 9,12219,655,369$. 