Journal of Comput. & Applied Mathematics **139**(2001), 197–213

DIRECT APPROACH TO CALCULUS OF VARIATIONS VIA NEWTON-RAPHSON METHOD

YURI LEVIN, MIKHAIL NEDIAK, AND ADI BEN-ISRAEL

Abstract. Consider m functions $f_i(x_1, \cdots, x_n)$, the system of equations $f_i = 0$, $i = 1, \cdots, m$ and the Newton iterations for this system that use the Moore–Penrose inverse of the Jacobian matrix. Under standard assumptions, the Newton iterations converge quadratically to a stationary point of the sum-of-squares $\sum f_i^2$. Approximating derivatives $\dot{x}$ as differences $\frac{\Delta x}{\Delta t}$ with $\Delta t = h$, we apply the Newton method to the system obtained by discretizing the integral $\int_{t_0}^{t_1} L(t, x, \dot{x})dt$. The approximate solutions y_h of the discretized problem are shown to converge to a solution of the Euler-Lagrange boundary value problem $\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}$ with the degree of approximation linear in h, if the Lagrangian $L(t, x, \dot{x})$ is twice continuously differentiable. Higher continuous derivatives of L guarantee higher orders of approximation.

1. Introduction

The Newton-Raphson method for solving a system of equations

$$\begin{array}{rcl} f_1(x_1, \ldots, x_n) & = & 0 \\ \ldots\ldots\ldots\ldots & & \\ f_m(x_1, \ldots, x_n) & = & 0 \end{array} \qquad \text{or} \quad \mathbf{f}(\mathbf{x}) = \mathbf{0}\,, \tag{1}$$

uses the iterations

$$\mathbf{x}^{k+1} := \mathbf{x}^k - (D\mathbf{f}(\mathbf{x}^k))^{-1}\,\mathbf{f}(\mathbf{x}^k)\,,\ k = 0, 1, \ldots \tag{2}$$

if $m = n$ and the Jacobian $D\mathbf{f}(\mathbf{x})) := \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x})\right)$ is nonsingular throughout the iterations. A solution of (1) can be found by minimizing the sum of squares

$$\sum_{i=1}^{m} f_i^2(\mathbf{x}) \ \rightarrow\ \min\ . \tag{3}$$

On the other hand, if the system (1) does not have a solution (e.g. if $m > n$), a minimizer of (3) may be a reasonable substitute. Such minimizers satisfy the necessary condition

$$\nabla \sum_{i=1}^{m} f_i^2(\mathbf{x}) \ =\ 0\ , \quad \text{or} \quad (D(f)(\mathbf{x}))^T\ \mathbf{f}(\mathbf{x}) = \mathbf{0}\,. \tag{4}$$

Replacing the inverse in (2) by the Moore-Penrose inverse $D\mathbf{f}(\mathbf{x})^\dagger$, we obtain a Newton method (see [1])

$$\mathbf{x}^{k+1} := \mathbf{x}^k - \left(D\mathbf{f}(\mathbf{x}^k)\right)^\dagger\,\mathbf{f}(\mathbf{x}^k)\,,\ k = 0, 1, \ldots \tag{5}$$

that can be shown to converge, at a quadratic rate, to a solution of (4), see [2].

In this paper we propose a direct method for solving calculus of variations problems with fixed end–points

$$\int_{\tau_0}^{\tau_1} L(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t))dt \ \rightarrow\ \min\,, \tag{6}$$

1991 *Mathematics Subject Classification.* Primary 65H10, 49M15; Secondary 49M25, 65L12.

Key words and phrases. Newton-Raphson methods, Euler-Lagrange equations, systems of equations, discrete approximations, stability, consistency.

The work of the first author was supported by DIMACS.

subject to
$$\boldsymbol{x}(\tau_0) = \boldsymbol{x}_0 \ , \ \boldsymbol{x}(\tau_1) = \boldsymbol{x}_1 \ , \tag{7}$$

where $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{R}^p$, $\boldsymbol{x}(\cdot) = (x_1(\cdot), \ldots, x_p(\cdot))$ and $x_i(\cdot) \in C^1([\tau_0, \tau_1]),\ i = 1, \ldots, p$. The method is based on the Newton iterations (5).

Assuming the Lagrangian L in (6) is positive throughout $[\tau_0, \tau_1]$ (since the end-points are fixed, adding a sufficiently large constant gives an equivalent problem), we can write it as a square

$$L = F^2(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) \ . \tag{8}$$

A regular partition of $[\tau_0, \tau_1]$,

$$[\tau_0,\ \tau_0 + h,\ \tau_0 + 2h, \cdots, \tau_0 + (N-1)\,h,\ \tau_1 = \tau_0 + N\,h]$$

is then used to approximate the derivative $\dot{x}(\cdot)$ of $x(\cdot)$ by a difference, say

$$\dot{x}(t) := \frac{x(t+h) - x(t)}{h} \ . \tag{9}$$

The integral of L in (6) is thus approximated by a sum of squares

$$\int_{\tau_0}^{\tau_1} L(t, \boldsymbol{x}, \dot{\boldsymbol{x}})\, dt \ \approx \ h \sum_k F^2 \left(t_k, \boldsymbol{x}(t_k), \frac{\boldsymbol{x}(t_k + h) - \boldsymbol{x}(t_k)}{h} \right) \tag{10}$$

where $t_k = \tau_0 + h\,k$. In this way the problem (6)–(7) can be approximated by the least squares problems

$$\sum_k f_k^2(\mathbf{x}) \ \rightarrow \ \min \tag{11}$$

where

$$f_k := F \left(t_k, \boldsymbol{x}(t_k), \frac{\boldsymbol{x}(t_k + h) - \boldsymbol{x}(t_k)}{h} \right)$$

and $\mathbf{x}$ is the vector with components $x_i(t_k)$.

In some problems it is natural to represent the Lagrangian as a sum of squares

$$L = \sum_i F_i^2(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) \tag{12}$$

rather than a single square as in (8). Examples are

$$L = \frac{1}{2} \left(x^2 + \dot{x}^2 \right) \ , \text{ or } L = M + \frac{1}{2} \left(x^2 - \dot{x}^2 \right) \ ,$$

where $M > 0$ is sufficiently large to make L positive throughout $[\tau_0, \tau_1]$. Also, other difference schemes may be used instead of (9). We prove that under some reasonable conditions (on the functions f_k in (12) and the difference scheme), the limit points as $\rightarrow \infty$ of the Newton method (5) are vectors $\mathbf{x}$ that correspond to extremals $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))$, i.e. curves satisfying the Euler–Lagrange equation

$$\frac{\partial}{\partial x_i} L(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) = \frac{d}{dt} \left(\frac{\partial}{\partial \dot{x}_i} L(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) \right), \ i = 1, \ldots, p \tag{13}$$

Thought the Newton-Raphson method has already been used in some approaches to the variational problems (see [3]), it is usually applied to the solution of the discretized version of (13) which requires the use of higher order derivatives of L. On the other hand, the approach presented in this paper applies the Newton-Raphson method directly to the least squares problem (11) which only requires the knowledge of the gradient.

2. Discretization Scheme

Let $t_k := \tau_0 + k\,h$, $k = 0, \ldots, N$ and $t_N := \tau_1$, where N (number of points) and $h = (\tau_1 - \tau_0)/N$ (step size) are parameters of the discretization scheme. Given p-dimensional vector function $\boldsymbol{x}(\cdot) = (x_1(\cdot), \ldots, x_p(\cdot))$ and a set $\mathcal{K} \subseteq \{0, \ldots, N\}$, denote by $\mathbf{P}_h^{\mathcal{K}}\boldsymbol{x}$ the $|\mathcal{K}| \times p$ matrix with rows $(x_1(t_k), \ldots, x_p(t_k))$, $k \in \mathcal{K}$. For $\mathcal{K} = \{0, \ldots, N\}$ we write $\mathbf{P}_h$, omitting the superscript $\mathcal{K}$.

Approximating a derivative $\dot{x}$ by a difference, e.g. (9) throughout an interval $[\tau_0, \tau_1]$, may cause problems in one of the endpoints τ_0 or τ_1, because of the need for an additional point outside $[\tau_0, \tau_1]$. The following definition sets the notation used in the sequel.

Definition 1. *Let $\{t_k := \tau_0 + k\,(\tau_1 - \tau_0)/N : k = 0, \ldots, N\}$ be a regular partition of $[\tau_0, \tau_1]$ with step size $h := (\tau_1 - \tau_0)/N$. Select a subset $\mathcal{J} \subset \{0, 1, \ldots, N\}$ such that the set of points $\{t_j : j \in \mathcal{J}\}$ excludes one or more of the endpoints. Typical choices of $\mathcal{J}$ are*

$$\{0, 1, \ldots, N-1\}\ ,\ \{1, 2, \ldots, N\}\ ,\ \text{or}\ \{1, 2, \ldots, N-1\}\ .$$

Let $x \in C^2([\tau_0, \tau_1])$. The discretization scheme for the derivative $\dot{x}$ is represented by an $|\mathcal{J}| \times (N+1)$ matrix A such that

$$A\mathbf{P}_h x = \mathbf{P}_h^{\mathcal{J}}\dot{x} + \varphi_x^{\mathcal{J}}(h)h, \tag{14}$$

$$(A^{\mathcal{I}})^T\mathbf{P}_h^{\mathcal{J}} x = -\mathbf{P}_h^{\mathcal{I}}\dot{x} + \psi_x^{\mathcal{I}}(h)h, \tag{15}$$

where

- *$\mathcal{I} = \mathcal{J} \setminus \{0, N\}$ is the index set of internal grid points,*
- *$A^{\mathcal{I}}$ is the submatrix of A with columns in $\mathcal{I}$, and*
- *$\varphi_x^{\mathcal{J}}(h)$ and $\psi_y^{\mathcal{I}}(h)$ are bounded in some neighborhood $\{h : |h| < \epsilon\}$ of zero.*

Relations (14)–(15) of the Definition refdef:A are not easy to verify directly. On the othe hand, the lemma below provides conditions that are easier to check and apply for a fairly large class of matrices A.

Lemma 1. *Let $A = (a_{kl})$, $\mathcal{J}$, $\mathcal{I}$ be as in Definition 1, and let the entries of hA be bounded functions of h in some neighborhood of 0. Then A satisfies (14)–(15) if and only if the following two conditions hold: (a) for all $k \in \mathcal{J}$*

$$\sum_{l=0}^{N} a_{kl} = 0\ ,\ \sum_{l=0}^{N} a_{kl}(l-k) = \frac{1}{h}\ , \tag{16}$$

(b) for all $l \in \mathcal{I}$

$$\sum_{k\in\mathcal{J}} a_{kl} = 0\ ,\ \sum_{k\in J} a_{kl}(k-l) = \frac{1}{h}\ . \tag{17}$$

Proof. Necessity follows by applying (14)–(15) to an arbitrary affine function $y(t) = a + bt$, $a, b \in \mathbb{R}$.

For sufficiency, we show first that (16) implies (14). For all $k \in \mathcal{J}$, applying the Taylor series expansion with residual term in Cauchy form and (16),

$$\begin{aligned} A_k\mathbf{P}_h y &= \sum_{l=0}^{N} a_{kl}y(t_l) \\ &= \sum_{l=0}^{N} a_{kl}(y(t_k) + \dot{y}(t_k)(l-k)h) + \sum_{l=0}^{N} a_{kl}\ddot{y}(t_k + \theta_{kl}(t_l - t_k))(l-k)^2h^2 \\ &= \dot{y}(t_k) + \sum_{l=0}^{N} a_{kl}\ddot{y}(t_k + \theta_{kl}(t_l - t_k))(l-k)^2h^2, \end{aligned}$$

where all θ_{kl} are constants in the interval $[0, 1)$.

The residual term here can be written in the form

$$\left(\sum_{l=0}^{N} ha_{kl}\ddot{y}(t_k+\theta_{kl}(t_l-t_k))(l-k)^2\right)h=\varphi_y^k(h)h,$$

with $\varphi_y^k(h)$ bounded in some neighborhood of 0 because hA is assumed to have the same property. Therefore (14) holds for any A satisfying (16).

We next show that (17) implies (15). For all $l\in\mathcal{I}$, applying (17)

$$\begin{aligned}(A^l)^T\mathbf{P}_h^J y &= \sum_{k\in J} a_{kl}y(t_k)\\ &= \sum_{k\in J} a_{kl}(y(t_l)+\dot{y}(t_l)(k-l)h)+\sum_{k\in J} a_{kl}\ddot{y}(t_l+\theta_{kl}(t_k-t_l))(k-l)^2h^2\\ &= \dot{y}(t_l)+\left(\sum_{k\in J} ha_{kl}\ddot{y}(t_l+\theta'_{kl}(t_k-t_l))(k-l)^2\right)h\\ &= \dot{y}(t_l)+\psi_y^l(h)h,\end{aligned}$$

where all θ'_{kl} are constants in the interval $[0,1)$. Again, $\psi_y^l(h)$ is bounded in some neighborhood of 0 because hA is assumed to have the same property. Thus (15) holds for such A. □

Now, let us show that a wide-known Euler approximation scheme satisfies the condition of Definition 1.

Example 1 (Euler Difference Approximation Scheme)**.** The Euler difference approximation scheme is defined by the $N\times(N+1)$ matrix $A=(a_{kl})$, where

$$a_{kl}=\begin{cases}-\dfrac{1}{h} & , \quad k=l\ ,\\ \dfrac{1}{h} & , \quad k=l-1\ ,\\ 0 & , \quad \text{otherwise,}\end{cases}$$

for $k\in\mathcal{J}=\{0,...,N-1\}$, $l\in\mathcal{I}=\{0,...,N\}$.

Obviously, in this case the conditions (16) and (17) hold. Therefore, by lemma 1, the Euler approximation scheme satisfies (14)–(15).

3. Solution of the Discretized Problem

Consider a twice continuously differentiable Lagrangian $L(t,u,v)$, represented as a sum of squares

$$L(t,u,v)=\sum_{i=1}^{m} F_i^2(t,u,v),$$

where the functions F_i , $i=1,\dots,m$ are (at least) continuously differentiable functions.

Combining (12) and (14), we conclude that the problem (6)–(7) can be approximated by the discretized problem

$$h\sum_{k\in\mathcal{J}}\sum_{i=1}^{m} F_i^2(t_k,Z_k,A_kZ) \quad\to\quad \min, \tag{18}$$

$$Z_0=\boldsymbol{x}_0,\ Z_N=\boldsymbol{x}_1, \tag{19}$$

where A_k and Z_k are k^{th} rows of the matrices A and Z respectively.

Consider the application of the Newton method (5) to the problem

$$F_i(t_k,Z_k,A_kZ)=0\ ,\ i=1,\dots,m\ ,\ k\in\mathcal{J} \tag{20}$$

obtained from the left sides of (18). If the method converges, we thus obtain a stationary point of the sum of squares of (18), i.e. a solution $\mathbf{z} = (z_{kj})$ of

$$\sum_{i=1}^{m}\left[F_i(t_k, Z_k, A_kZ)\frac{\partial F_i}{\partial x_j}(t_k, Z_k, A_kZ) + \sum_{l\in J} F_i(t_l, Z_l, A_lZ)\frac{\partial F_i}{\partial \dot{x}_j}(t_l, Z_l, A_lZ)a_{lk}\right] = 0, k \in \mathcal{I} \tag{21}$$

with Z_0 and Z_N given by (19) and $j = 1, \ldots, p$.

Introduce the following notation: for any function $G(\cdot,\cdot,\cdot)$, vector $\mathbf{s} \in \mathbb{R}^{|K|}$ and matrices $U, V \in \mathbb{R}^{|K|\times p}$, with rows indexed by the set $\mathcal{K}$, let $G(\mathbf{s}, U, V)$ be a vector in $\mathbb{R}^{|\mathcal{K}|}$ with coordinates $G(s_k, U_k, V_k)$, $k \in \mathcal{K}$.

The system of equations (21) can be written as

$$\sum_{i=1}^{m}\left[\mathrm{diag}\left(\frac{\partial F_i}{\partial x_j}(\boldsymbol{t}^{\mathcal{I}}, Z^{\mathcal{I}}, (AZ)^{\mathcal{I}})\right) F_i(\boldsymbol{t}^{\mathcal{I}}, Z^{\mathcal{I}}, (AZ)^{\mathcal{I}}) + \right. \tag{22}$$

$$\left.(A^{\mathcal{I}})^T\mathrm{diag}\left(\frac{\partial F_i}{\partial \dot{x}_j}(\boldsymbol{t}^{\mathcal{J}}, Z^{\mathcal{J}}, AZ)\right) F_i(\boldsymbol{t}^{\mathcal{J}}, Z^{\mathcal{J}}, AZ)\right] = 0,\ j = 1, \ldots, p \tag{23}$$

that is equivalent to

$$\frac{\partial L}{\partial x_j}\left(\boldsymbol{t}^{\mathcal{I}}, Z^{\mathcal{I}}, (AZ)^{\mathcal{I}}\right) + (A^{\mathcal{I}})^T\frac{\partial L}{\partial \dot{x}_j}\left(\boldsymbol{t}^{\mathcal{J}}, Z^{\mathcal{J}}, AZ\right) = 0,\ j = 1, \ldots, p\,. \tag{24}$$

Note that (21) can be derived directly from the stationarity condition (4) satisfied by the limit points $\mathbf{z}^* = (z^*_{ik})$ of the Newton method (5), applied to (20). Indeed, (21) is a rewriting of

$$[\mathbf{D}F(\mathbf{z}^*)]^T F(\mathbf{z}^*) = 0\,. \tag{25}$$

4. Convergence of the Newton Method to the Solution of the Euler-Lagrange Equation

The Euler–Lagrange boundary value system

$$\frac{\partial}{\partial x_j}L(t, \boldsymbol{y}, \dot{\boldsymbol{y}}) - \frac{d}{dt}\frac{\partial}{\partial \dot{x}_j}L(t, \boldsymbol{y}, \dot{\boldsymbol{y}}) = 0\,,\ j = 1, \ldots, p \tag{26}$$

$$\boldsymbol{y}(t_0) = \boldsymbol{x}_0\,,\ \boldsymbol{y}(t_1) = \boldsymbol{x}_1\,, \tag{27}$$

is a necessary condition for $\boldsymbol{y}$ to be an extremal of (6)–(7).

For further discussion, introduce an operator $B_h : \mathbb{R}^{(N+1)\times p} \to \mathbb{R}^{|\mathcal{I}|}$ as

$$B_h(Z) = \left(\frac{\partial}{\partial x_j}L\left(\boldsymbol{t}^{\mathcal{I}}, Z^{\mathcal{I}}, (AZ)^{\mathcal{I}}\right) + (A^{\mathcal{I}})^T\frac{\partial}{\partial \dot{x}_j}L(\boldsymbol{t}^{\mathcal{J}}, Z^{\mathcal{J}}, AZ),\ j = 1, \ldots, p\right), \tag{28}$$

for any $Z \in \mathbb{R}^{(N+1)\times p}$. Then (24) can be written as

$$B_h(Z) = 0\,. \tag{29}$$

Following the treatment in [4, § 20.1, p. 195], we now formulate the conditions used in the main convergence result.

(P1) Existence and Uniqueness: The problem (26)–(27) has a unique solution $\boldsymbol{y}(\cdot)$, and for all sufficiently large N the problem (29)-(19) has the unique solution Y_h.

(P2) Stability: There exists positive constants s and γ such that the inequality

$$\|V - U\|^s \le \gamma\|B_h(V) - B_h(U)\| \tag{30}$$

holds for any matrices $U, V \in \mathbb{R}^{(N+1)\times p}$ satisfying boundary condition (19) and where $\|\cdot\|$ is a componentwise l_∞ norm (i.e., for an arbitrary matrix W, $\|W\| = \max_{i,j}|w_{ij}|$).

(P3) Consistency: The matrix A representing the difference scheme satisfies conditions (14)–(15).

A special case when condition (P2) holds is when the operator B_h is linear. This happens when both $\frac{\partial L}{\partial x}(t,u,v)$, and $\frac{\partial L}{\partial x}(t,u,v)$ are linear in u and v. This, in turn, is true when $L(t,u,v)$ is a quadratic function in u,v. Conditions for (P3) to hold are given in § 2.

Conditions (P1) and (P3) allow proving the following lemma, needed in the sequel.

Lemma 2. *Under assumptions (P1) and (P3) there exists a constant $\alpha > 0$ such that*

$$\|B_h(\mathbf{P}_h\boldsymbol{y})\| \le \alpha\, h\ , \tag{31}$$

for sufficiently small h.

Proof. For simplicity, let $p = 1$. The general case can be shown in a similar way.

Since $y(\cdot)$ is the solution of (26)–(27), the components of $B_h(\mathbf{P}_h y)$ corresponding to the boundary conditions are zeros. We may therefore concentrate on the components indexed by the set $\mathcal{I}$.

Condition (14) implies that for any $k \in \mathcal{J}$ there exists $\theta_k \in (0,1)$ such that

$$\begin{aligned}\frac{\partial L}{\partial x}(t_k, y(t_k), A_k\mathbf{P}_h y) &= \frac{\partial L}{\partial x}(t_k, y(t_k), \dot y(t_k) + \varphi_y^k(h)h)\\ &= \frac{\partial L}{\partial x}(t_k, y(t_k), \dot y(t_k)) + \frac{\partial^2 L}{\partial x \partial \dot x}(t_k, y(t_k), \dot y(t_k) + \theta_k\varphi_y^k(h)h)\varphi_y^k(h)h\ .\end{aligned} \tag{32}$$

Similarly, for any $k \in \mathcal{J}$ there exists $\theta'_k \in (0,1)$ such that

$$\begin{aligned}\frac{\partial L}{\partial \dot x}(t_k, y(t_k), A_k\mathbf{P}_h y) &= \frac{\partial L}{\partial \dot x}(t_k, y(t_k), \dot y(t_k) + \varphi_y^k(h)h)\\ &= \frac{\partial L}{\partial \dot x}(t_k, y(t_k), \dot y(t_k)) + \frac{\partial^2 L}{\partial \dot x^2}(t_k, y(t_k), \dot y(t_k) + \theta'_k\varphi_y^k(h)h)\varphi_y^k(h)h\ .\end{aligned} \tag{33}$$

Now we apply $(A^{\mathcal{I}})^T$ to the vector formed by the first term of (33) with different k. Condition (15) implies that

$$(A^{\mathcal{I}})^T\mathbf{P}_h^{\mathcal{J}}\frac{\partial L}{\partial \dot x}(t,y,\dot y) = -\mathbf{P}_h^{\mathcal{I}}\frac{d}{dt}\frac{\partial L}{\partial \dot x}(t,y,\dot y) + \psi^{\mathcal{I}}_{\frac{\partial L}{\partial \dot x}(t,y,\dot y)}(h)h. \tag{34}$$

Observe,

$$\mathbf{P}_h^{\mathcal{I}}\frac{\partial L}{\partial x}(t,y,\dot y) = \mathbf{P}_h^{\mathcal{I}}\frac{d}{dt}\frac{\partial L}{\partial \dot x}(t,y,\dot y). \tag{35}$$

Adding the right-hand side of (32) with the right-hand side of (33) premultiplied by $(A^{\mathcal{I}})^T$, and using (34) and (35) we obtain

$$B_h(\mathbf{P}_h y) = R_h(y)h,$$

where $R_h(y)$ is bounded in h in some neighborhood of 0. This completes the proof of the lemma. □

Theorem 1 (Convergence Theorem)**.** *Let $\boldsymbol{y}(\cdot)$ be a solution of the Euler-Lagrange system (26)–(27) and Y_h be a solution of (19)-(29). Then conditions (P1)-(P3) imply the existence of positive constants γ , s and α such that*

$$\|Y_h - P_h\boldsymbol{y}\|^s \le \alpha\gamma h\ , \tag{36}$$

for all sufficiently small h. This implies the convergence of the approximate solution Y_h to $\boldsymbol{y}(\cdot)$ as $h \to 0$.

Proof. By Lemma 2 there exists a positive constant α such that (31) holds for sufficiently small h. Using Condition (P2) with $V = Y_h$ and $U = \mathbf{P}_h\boldsymbol{y}$, we obtain the existence of positive s and γ such that

$$\begin{aligned}\|Y_h - \mathbf{P}_h\boldsymbol{y}\|^s &\le \gamma\|B_h(Y_h) - B_h(\mathbf{P}_h\boldsymbol{y})\|\\ &= \gamma\|B_h(\mathbf{P}_h\boldsymbol{y})\|\ ,\ \text{since}\ \ B_h(Y_h) = \mathbf{0}\\ &\le \alpha\gamma h\ ,\ \text{by (31)}\ ,\end{aligned}$$

proving the existence of positive constants α, γ and s such that (36) holds for sufficiently small h. □

5. Generalizations

In practice it is often important to obtain better precision of the discretization scheme. Theorem 1 can only guarantee precision of the order $h^{1/s}$. One straightforward way to improve on that is to impose stricter condition on the degree of approximation of derivative provided by matrix A. We reformulate (14) and (15) as follows:

Let ν be an integer ≥ 1. For any $y \in C^{\nu+1}([\tau_0, \tau_1])$

$$AP_h y = \mathbf{P}_h^J \dot{y} + \varphi_y^J(h) h^\nu, \tag{37}$$

$$(A^{\mathcal{I}})^T \mathbf{P}_h^J y = -\mathbf{P}_h^{\mathcal{I}} \dot{y} + \psi_y^I(h) h^\nu, \tag{38}$$

where

- $\mathcal{I} = \{1, \ldots, N-1\}$ is a set of internal grid points,
- $A^{\mathcal{I}}$ is the submatrix of A with columns indexed by $\mathcal{I}$, and
- $\varphi_y^{\mathcal{J}}(h)$ and $\psi_y^{\mathcal{I}}(h)$ are bounded in h in some neighborhood of 0.

Condition (P3) will be reformulated now as

(P3′) Consistency: The matrix A representing the difference scheme satisfies conditions (37)–(38).

As an immediate result of these more restrictive conditions we obtain the following generalization of the Lemma 2:

Lemma 3. *Under the assumptions of $\nu+1$ times continuous differentiability of L, existence of unique solution $\boldsymbol{y}(\cdot)$ of (26)-(27) and (P3′) there exists $\alpha > 0$ such that $\|B_h(\mathbf{P}_h \boldsymbol{y})\| \leq \alpha h^\nu$ for any sufficiently small h.*

The proof of Lemma 3 uses the Taylor series development up to degree $\nu+1$ and the consequences of (26) which are obtained by differentiating it up to ν times.

Theorem 1 is now generalized to

Theorem 2 (Convergence Theorem)**.** *Let $\boldsymbol{y}(\cdot)$ be a solution of the Euler-Lagrange system (26)–(27) and Y_h be a solution of (19)-(29). Then conditions (P1),(P2) and (P3′) with an additional requirement of $(\nu+1)$–times continuous differentiability of L imply that there exist positive constants γ , s and α such that*

$$\|Y_h - \mathbf{P}_h \boldsymbol{y}\|^s \leq \alpha\gamma h^\nu \,, \tag{39}$$

for all sufficiently small h. This implies the convergence of the approximate solution Y_h to $\boldsymbol{y}(\cdot)$ as $h \to 0$.

The proof of the Theorem 2 follows along the lines of the proof of Theorem 1.

We can also generalize Lemma 1 with an analogous proof:

Lemma 4. *Let A , $\mathcal{J}$ and $\mathcal{I}$ be as in Definition 1, and let the entries of $h^\nu A$ be bounded in h in some small neighborhood of 0. Then A satisfies the consistency condition (P3′) if and only if the following two conditions hold:*
(a) for all $k \in \mathcal{J}$

$$\sum_{l=0}^{N} a_{kl}(l-k)^n = 0 \,,\ 0 \leq n \leq \nu \,,\ n \neq 1 \,,\ \sum_{l=0}^{N} a_{kl}(l-k) = \frac{1}{h} \,, \tag{40}$$

(b) for all $l \in \mathcal{I}$

$$\sum_{k \in J} a_{kl}(k-l)^n = 0 \,,\ 0 \leq n \leq \nu \,,\ n \neq 1 \,,\ \sum_{k \in J} a_{kl}(k-l) = \frac{1}{h} \,. \tag{41}$$

6. Numerical results

Recall that the integral in (6) is approximated by a sum-of-squares (18)

$$\int_{t_0}^{t_1} L(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t))dt \;\approx\; h \sum_{k\in\mathcal{J}} \sum_{i=1}^{m} F_i^2(t_k, Z_k, A_k Z)$$

and a minimum of the sum-of-squares is obtained by attempting to solve the over-determined system (20),

$$F_i(t_k, Z_k, A_k Z) = 0 \ , \ i = 1, \dots, m \ , \ k \in \mathcal{J}.$$

In some cases it is advantageous to add a constant θ to the Lagrangian, obtaining the problem,

$$\int_{t_0}^{t_1} (L(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) + \theta)dt \;\rightarrow\; \min \ , \tag{42}$$

subject to (7). This problem is equivalent to (6)–(7), since the endpoints are fixed. The approximation (18) now becomes

$$\int_{t_0}^{t_1} (L(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) + \theta)dt \;\approx\; h \sum_{k\in\mathcal{J}} \sum_{i=1}^{m} (F_i^2(t_k, Z_k, A_k Z) + \theta) \tag{43}$$

and the over–determined system (20)

$$\sqrt{F_i(t_k, Z_k, A_k Z)^2 + \theta} = 0 \ , \ i = 1, \dots, m \ , \ k \in \mathcal{J} \ . \tag{44}$$

For $\theta = 0$ this reduces to the original system (20). In some problems a small negative θ resulted in faster convergence. Note, if $F^2 > 0$ then a negative θ would bring $F^2 + \theta$ closer to zero, resulting in smaller Newton steps. A positive θ is useful if it is necessary to keep the integrand of (42) positive.

The numerical examples below were computed using MATLAB . In examples 2-5 we minimize an integral $\int_0^1 L(t, y, \dot{y})\, dt$ subject to the same boundary conditions,

$$y(0) = 0 \ , \ y(1) = 1 \ . \tag{45}$$

In each table, the last column (labeled $y(t_i)$) is the discretized theoretical solution. The initial approximation z_1 is a random perturbation of the discretized theoretical solution.

In example 6 we minimize an integral $\int_0^1 L(t, x, y, \dot{x}, \dot{y})\, dt$ (two dimensional system).

Example 2 (Shortest path problem)**.** The shortest path between $(0, 0)$ and $(1, 1)$ is found by minimizing an integral of

$$\sqrt{1 + \dot{y}^2}dt \ , \tag{46}$$

subject to (45). The solution is the straight line

$$y(t) = t \ , \ 0 \le t \le 1 \ . \tag{47}$$

The Newton method is applied to the system

$$\sqrt{\theta + \sqrt{1 + N^2(z_{i+1} - z_i)^2}} = 0, \ i = 0, \dots, N-1 \tag{48}$$

with $z_0 = 0$, $z_N = 1$. Table 1 gives the first five odd iterations, with $N = 10$ and $\theta = -0.5$. The solution (47) is obtained in 9 iterations, see column z_9.

Example 3 (A mechanical system with a single particle)**.** Consider the Lagrangian

$$L(t, y, \dot{y}) = \dot{y}^2 + y^2 \ . \tag{49}$$

The corresponding Euler-Lagrange equation is

$$\ddot{y} = y$$

i	t_i	z_1	z_3	z_5	z_7	z_9	$x(t_i)$
1	0.10	0.06509	0.09610	0.09965	0.09997	0.10000	0.10000
2	0.20	0.21979	0.19784	0.19982	0.19998	0.20000	0.20000
3	0.30	0.28784	0.29360	0.29944	0.29995	0.30000	0.30000
4	0.40	0.43600	0.39623	0.39968	0.39997	0.40000	0.40000
5	0.50	0.53537	0.50024	0.50001	0.50000	0.50000	0.50000
6	0.60	0.60936	0.59669	0.59970	0.59997	0.60000	0.60000
7	0.70	0.69966	0.69834	0.69986	0.69999	0.70000	0.70000
8	0.80	0.83998	0.80201	0.80017	0.80002	0.80000	0.80000
9	0.90	0.93216	0.90423	0.90038	0.90003	0.90000	0.90000

TABLE 1. Numerical results for Example 2 with $\theta = -0.5$

i	t_i	z_1	z_4	z_7	z_{11}	z_{15}	$y(t_i)$
1	0.10	0.10689	0.08453	0.08536	0.08526	0.08525	0.08523
2	0.20	0.17245	0.16990	0.17158	0.17137	0.17134	0.17132
3	0.30	0.28676	0.25696	0.25951	0.25919	0.25916	0.25912
4	0.40	0.34845	0.34671	0.35004	0.34960	0.34956	0.34952
5	0.50	0.41200	0.44013	0.44407	0.44351	0.44346	0.44341
6	0.60	0.56180	0.53712	0.54255	0.54186	0.54179	0.54174
7	0.70	0.69376	0.63904	0.64648	0.64563	0.64554	0.64549
8	0.80	0.78637	0.75035	0.75668	0.75584	0.75575	0.75571
9	0.90	0.89384	0.86897	0.87437	0.87361	0.87351	0.87348

TABLE 2. Numerical results for Example 3 with $\theta = 0$

i	t_i	z_1	z_4	z_7	z_{11}	z_{15}	$y(t_i)$
1	0.10	0.07336	-0.02437	0.04157	0.04173	0.04175	0.01505
2	0.20	-0.16079	-0.02964	0.21988	0.22286	0.22326	0.03254
3	0.30	0.25444	-0.03928	0.22029	0.21997	0.21993	0.05531
4	0.40	0.21432	-0.24122	0.21921	0.21894	0.21890	0.08705
5	0.50	0.27846	-0.21274	0.48561	0.48521	0.48515	0.13290
6	0.60	0.35778	-0.29254	0.29230	0.29213	0.29211	0.20030
7	0.70	0.38518	0.73170	0.30523	0.30509	0.30507	0.30018
8	0.80	0.29917	1.43842	0.25401	0.25393	0.25392	0.44873
9	0.90	0.55659	1.21308	0.52032	0.52029	0.52028	0.67004

TABLE 3. Numerical results for Example 4, with $L = F^2$ and $\theta = 0$

whose solution, subject to (45), is

$$y(t) = \frac{e^t - e^{-t}}{e - e^{-1}} \tag{50}$$

discretized in the last column of Table 2. A reasonable approximation is obtained in 15 iterations.

Example 4. Consider the Lagrangian $L = \dot{y}^2 + 16y^2$. Table 3 gives the results for $\theta = 0$, if the Lagrangian is expressed as a single square

$$L = F^2 \ .$$

The iterations do not converge to the solution $y(t_i)$ in the last column. Changing to a sum of two squares

$$L = F_1^2 + F_2^2 \ , \text{ with } F_1^2 = \dot{y}^2 \ , \ F_2^2 = 16y^2 \ ,$$

i	t_i	z_1	z_4	z_7	z_{11}	z_{15}	$y(t_i)$
1	0.10	0.07336	0.01504	0.01537	0.01535	0.01534	0.01505
2	0.20	-0.16079	0.03264	0.03318	0.03315	0.03314	0.03254
3	0.30	0.25444	0.05568	0.05629	0.05625	0.05625	0.05531
4	0.40	0.21432	0.08784	0.08838	0.08835	0.08835	0.08705
5	0.50	0.27846	0.13419	0.13461	0.13459	0.13459	0.13290
6	0.60	0.35778	0.20208	0.20238	0.20236	0.20236	0.20030
7	0.70	0.38518	0.30233	0.30252	0.30251	0.30251	0.30018
8	0.80	0.29917	0.45096	0.45107	0.45107	0.45107	0.44873
9	0.90	0.55659	0.67174	0.67179	0.67179	0.67179	0.67004

TABLE 4. Numerical results for Example 4, with $L = F_1^2 + F_2^2$

i	t_i	z_1	z_4	z_7	z_{11}	$y(t_i)$
1	0.10	0.12005	0.11863	0.11866	0.11866	0.11864
2	0.20	0.34083	0.23607	0.23613	0.23613	0.23610
3	0.30	0.31564	0.35116	0.35124	0.35124	0.35119
4	0.40	0.36509	0.46273	0.46284	0.46284	0.46278
5	0.50	0.41457	0.56969	0.56981	0.56981	0.56975
6	0.60	0.51773	0.67096	0.67108	0.67108	0.67102
7	0.70	0.85670	0.76554	0.76564	0.76565	0.76559
8	0.80	0.75388	0.85247	0.85255	0.85255	0.85250
9	0.90	0.95174	0.93088	0.93093	0.93093	0.93090

TABLE 5. Numerical results for Example 5

and using $\theta_1 = 0$ and $\theta_2 = 0.1$, the method converges in about 15 iterations, see Table 4.

Example 5 (Harmonic oscillator)**.** Consider the harmonic oscillator with Lagrangian $L = \dot{y}^2 - y^2$. We solve it here using the factorization

$$1 + L = F_1^2 + F_2^2 \ , \ \text{where } F_1^2 = \dot{y}^2 \ , \ F_2^2 = 1 - y^2 \ .$$

The Euler–Lagrange equation is

$$\ddot{y} = -y \ ,$$

whose solution subject to (45) is $y(t) = \sin t / \sin(1)$. Table 5 shows convergence in 11 iterations.

Example 6 (Two dimensional system)**.** Consider a two dimensional system with Lagrangian

$$L = \dot{x}^2 + \dot{y}^2 + x^2 + 2y^2 + xy.$$

on the interval $[0, 1]$. Numerical procedure uses the factorization

$$L = F_1^2 + F_2^2 + F_3^2 + F_4^2 + F_5^2$$

where

$$\begin{aligned} F_1^2 &= \dot{x}^2, \\ F_2^2 &= \dot{y}^2, \\ F_3^2 &= x^2, \\ F_4^2 &= 2y^2, \\ F_5^2 &= xy. \end{aligned}$$

i	t_i	z_1	z_4	z_7	z_{11}	$x(t_i)$
1	0.10	0.02885	0.07398	0.07399	0.07399	0.07395
2	0.20	0.17841	0.15293	0.15294	0.15294	0.15287
3	0.30	0.27635	0.23695	0.23695	0.23695	0.23687
4	0.40	0.30478	0.32625	0.32626	0.32626	0.32616
5	0.50	0.39621	0.42119	0.42119	0.42119	0.42109
6	0.60	0.56539	0.52220	0.52221	0.52221	0.52211
7	0.70	0.59289	0.62987	0.62987	0.62987	0.62979
8	0.80	0.78887	0.74485	0.74485	0.74485	0.74479
9	0.90	0.88808	0.86793	0.86793	0.86793	0.86790

TABLE 6. Numerical results for Example 6: x component

i	t_i	z_1	z_4	z_7	z_{11}	$y(t_i)$
1	0.10	0.87945	0.84477	0.84477	0.84477	0.84468
2	0.20	0.67758	0.70680	0.70680	0.70680	0.70665
3	0.30	0.57906	0.58374	0.58374	0.58374	0.58355
4	0.40	0.43144	0.47353	0.47353	0.47353	0.47333
5	0.50	0.40934	0.37443	0.37442	0.37442	0.37423
6	0.60	0.29094	0.28492	0.28491	0.28491	0.28473
7	0.70	0.18549	0.20372	0.20371	0.20371	0.20356
8	0.80	0.11711	0.12974	0.12974	0.12974	0.12962
9	0.90	0.09879	0.06208	0.06208	0.06208	0.06201

TABLE 7. Numerical results for Example 6: y component

Boundary values are $x(0) = y(1) = 0$ and $x(1) = y(0) = 1$. Tables 6 and 7 show convergence in 11 iterations.

In the examples above we can see that the initial solution was fairly close to the optimal. However, these examples were provided mostly to illustrate the idea of the method. We have also made experiments where the initial solution was very far from the optimal. In these cases the numerical procedure had to use adaptive θ. Since we currently do not have a solid theoretical framework for this approach we reserve its presentation for the future.

REFERENCES

[1] A. Ben-Israel, *A Newton-Rapshon method for the solution of systems of equations*, J. Math. Anal. Appl. **15**(1966), 243–252.
[2] Y. Levin and A. Ben-Israel, *Convergence of the Newton-Raphson Method*, to appear
[3] H. Kagiwada, R. Kalaba, N. Rasakhoo and K. Spingarn, *Numerical Derivatives and Nonlinear Analysis*, Plenum Press, 1986.
[4] E. Zeidler, *Nonlinear Functional Analysis and its Applications*, Vol II, Springer-Verlag, 1990.

YURI LEVIN, RUTCOR–RUTGERS CENTER FOR OPERATIONS RESEARCH, RUTGERS UNIVERSITY, 640 BARTHOLOMEW RD, PISCATAWAY, NJ 08854-8003, USA
E-mail address: ylevin@rutcor.rutgers.edu

MIKHAIL NEDIAK, RUTCOR–RUTGERS CENTER FOR OPERATIONS RESEARCH, RUTGERS UNIVERSITY, 640 BARTHOLOMEW RD, PISCATAWAY, NJ 08854-8003, USA
E-mail address: msnediak@rutcor.rutgers.edu

ADI BEN-ISRAEL, RUTCOR–RUTGERS CENTER FOR OPERATIONS RESEARCH, RUTGERS UNIVERSITY, 640 BARTHOLOMEW RD, PISCATAWAY, NJ 08854-8003, USA
E-mail address: bisrael@rutcor.rutgers.edu