Control and Cybernetics

vol. **43** (2014) No. 1

A simple proof of the maximum principle with endpoint constraints

by

Adam Korytowski

Department of Automatics and Biomedical Engineering
AGH University, Kraków, Poland
akor@agh.edu.pl

Abstract: The paper presents a new, relatively simple proof of Pontryagin's maximum principle for the *canonical problem* of optimal control, with equality and inequality constraints imposed on the trajectory endpoints. The proof combines together two ideas, which appeared separately in the earlier works: application of the Karush-John conditions for finite-dimensional problems, and using packages of needle variations.

Keywords: optimal control, maximum principle

1. Introduction

Pontryagin's maximum principle is, in its various versions, probably the most important theorem of optimal control, both for theory and applications. It is therefore vital to make it as easy to access as possible, for students, researchers, and engineers. Unfortunately, short and elementary proofs of the maximum principle are known only for very special cases, e.g., with fixed initial state and free final state. Such cases do not include the nonlinear time optimal problem, which can be viewed as a flagship problem of optimal control.

This paper is intended as a step towards improving this situation. It presents a new, relatively short and elementary proof of the maximum principle for the *canonical optimal control problem* (Milyutin et al., 2004), with equality and inequality constraints imposed on the trajectory endpoints. Leaving out mixed or pathwise state constraints, most of the ODE optimal control problems taught at universities, including the time optimal ones, may be easily reduced to this canonical form or treated as its special cases. The simplicity of the proof is achieved by combining together two fundamental ideas, which, in the earlier works, appeared separately: application of the Karush-John conditions for finite-dimensional problems (known also as Fritz John conditions, see Mangasarian and Fromovitz, 1967), and using packages of needle variations. The former idea is due to A.A. Milyutin (Milyutin et al., 2004, see also Dubovitskii and

Milyutin, 1965), the latter dates back to the very beginnings of the maximum principle (Pontryagin et al., 1961).

The paper begins with the problem statement, slightly different from that in Milyutin et al. (2004). For the ease of argument, feasible controls are limited to piecewise continuous functions. Next, necessary facts from the theory of differential equations are recalled. The maximum principle is then formulated, followed by the proof. Lastly, the time variation of the Hamiltonian is characterized in a separate theorem. This well known result is included for completeness, also because its proof is partly new. In the appendix, a C^1 extension from the positive orthant onto the whole space is described.

It is my pleasant duty to thank Nikolai P. Osmolovskii, without whose help this work would not have been possible. He drew my attention to the works of A.A. Milyutin and his coworkers, and pointed out errors and omissions in the manuscript. Thanks are also due to the anonymous reviewers for their helpful comments and suggestions.

2. Canonical optimal control problem and variational equation

Our purpose is to formulate the necessary optimality conditions in an optimal control problem, where the decision variables consist of the initial and terminal time moments t_0 and t_1, the control $u \in PC(t_0, t_1; R^m)$[1], the initial state $x_0 \in R^n$, and the terminal state $x_1 \in R^n$. These variables are subject to constraints: $t_1 > t_0$, $u(t) \in D \subset R^m$ for every $t \in [t_0, t_1]$, $g(t_0, x_0, t_1, x_1) \leqslant 0$, $h(t_0, x_0, t_1, x_1) = 0$, $x_0 = x(t_0)$ and $x_1 = x(t_1)$, where $x : [t_0, t_1] \to R^n$ is a solution of the state equation

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{1}$$

that is, an absolutely continuous function satisfying (1) almost everywhere. D is a given set, the functions $g : R^{2n+2} \to R^r$ and $h : R^{2n+2} \to R^k$ are of class C^1. The function $f : R \times R^n \times R^m \to R^n$ is of class C^1 in the second argument, f and its derivative $\partial_2 f$ are continuous in all their arguments.

A performance index $Q : R^{2n+2} \times PC(t_0, t_1; R^m) \to R$ is to be minimized,

$$Q(t_0, x_0, t_1, x_1, u) = q(t_0, x_0, t_1, x_1),$$

with $q : R^{2n+2} \to R$ being of class C^1. Any tuple (t_0, x_0, t_1, x_1, u) which fulfils the above assumptions is called a *feasible solution.* Any feasible tuple $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, \hat{u})$, such that $Q(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, \hat{u}) \leqslant Q(t_0, x_0, t_1, x_1, u)$ for every feasible (t_0, x_0, t_1, x_1, u), is called an *optimal solution.*

The solution of (1) with an optimal control $\hat{u}$ and condition $x(s) = \xi$ defines a function $P : R^{n+2} \to R^n$,

$$P(t, s, \xi) = x(t), \tag{2}$$

[1] $PC(t_0, t_1; Y)$ denotes the space of all functions $[t_0, t_1] \to Y$ which have a finite number of discontinuities, are right-continuous in $[t_0, t_1[$, left-continuous at t_1, and have a finite left-hand limit at every discontinuity point.

for $t, s \in [\hat{t}_0, \hat{t}_1]$ and ξ from a certain neighborhood of $\hat{x}(s)$. It is well known from the theory of differential equations that P is of class C^1 in the third argument, and continuous and piecewise C^1 w.r.t. the others. Let now x be a fixed, feasible solution of (1) with the control $\hat{u}$. Denote $\Phi(t,s) = \partial_3 P(t, s, x(s))$. From (1) and (2) it immediately follows that the function $t \mapsto \Phi(t,s)$ satisfies a *matrix variational equation*

$$\partial_1 \Phi(t,s) = \Phi(t,s) A(t), \quad \Phi(s,s) = I,$$

where $A(t) = \partial_2 f(t, x(t), \hat{u}(t))$. The matrix $\Phi(t,s)$ is nonsingular and

$$\Phi(t,s)\ \Phi(s,t) = I.$$

By differentiating both sides of this last equality we obtain

$$\partial_2 \Phi(t,s) = -A(s)\, \Phi(t,s). \tag{3}$$

3. Maximum Principle

Define the *Pontryagin function* $H : R \times R^n \times R^m \times R^n \to R$,

$$H(t, x, u, \psi) = \psi^T f(t, x, u).$$

Theorem 1 (Maximum Principle). *Assume that $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, \hat{u})$ is an optimal solution and $\hat{x}$ is the corresponding state trajectory. There then exist $\lambda \in R$, $\mu \in R^r$, $\rho \in R^k$ and an absolutely continuous function $\hat{\psi} : [\hat{t}_0, \hat{t}_1] \to R^n$, such that the following relationships hold*

(i) $\lambda \geqslant 0$, $\mu \geqslant 0$ (*non-negativity conditions*)

(ii) $\lambda + ||\,\mu\,|| + ||\,\rho\,|| \ > 0$ (*non-triviality condition*)

(iii) $\mu^T g(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1) = 0$ (*complementarity condition*)

(iv) $\dot{\hat{\psi}}(t) = -\partial_2 H(t, \hat{x}(t), \hat{u}(t), \hat{\psi}(t))$ a.e. $t \in [\hat{t}_0, \hat{t}_1]$ (*adjoint equation*)

(v) $\hat{\psi}(\hat{t}_0) = \partial_2(\lambda q + \mu^T g + \rho^T h)$

$\hat{\psi}(\hat{t}_1) = -\partial_4(\lambda q + \mu^T g + \rho^T h)$

$H[\hat{t}_0] = -\partial_1(\lambda q + \mu^T g + \rho^T h)$

$H[\hat{t}_1] = \partial_3(\lambda q + \mu^T g + \rho^T h)$ (*transversality conditions*)

(vi) $H(t, \hat{x}(t), v, \hat{\psi}(t)) \leqslant H[t] \quad \forall t \in [\hat{t}_0, \hat{t}_1] \quad \forall v \in D$ (*maximum condition*).

Here, $\partial_i(\lambda q + \mu^T g + \rho^T h)$ stands for the derivative w.r.t. the ith argument of q, gand h, computed at the optimal point, and $H[t] = H(t, \hat{x}(t), \hat{u}(t), \hat{\psi}(t))$.

Proof. The theorem will be proved in stages. First, a family of auxiliary, finite-dimensional optimization problems will be introduced. For each problem of the family, the Karush–John necessary optimality conditions will be formulated. Next, a topological argument will be used to prove that there are Lagrange multipliers common for the whole family. Lastly, an adjoint function will be defined and used for the formulation of necessary conditions.

Note that similar Karush–John conditions and the topological argument appear in Milyutin et al. (2004). However, the auxiliary finite-dimensional problems in that book were constructed differently, in a more complex way and without employing needle variations.

Let us extend the control $\hat{u}$ outside the interval $[\hat{t}_0, \hat{t}_1]$, so that it is continuous at $\hat{t}_0$ and $\hat{t}_1$. For a positive integer s introduce three sequences: a non-decreasing sequence $\theta_1, ..., \theta_s$ contained in $[\hat{t}_0, \hat{t}_1[$, a sequence of non-negative real numbers $\varepsilon_1, ..., \varepsilon_s$, and a sequence $v_1, ..., v_s$, contained in D. Let further u be a control in $PC(\hat{t}_0, \hat{t}_1; D)$,

$$u(t) = \begin{cases} v_i, & t \in [\theta'_i, \theta'_i + \varepsilon_i[\,, \quad i = 1, ..., s \\ \hat{u}(t), & \text{otherwise}, \end{cases}$$

where $\theta'_1 = \theta_1$, and for $i > 1$, $\theta'_i = \theta_i$, if $\theta_i > \theta_{i-1}$, and $\theta'_i = \theta'_{i-1} + \varepsilon_{i-1}$, if $\theta_i = \theta_{i-1}$. Assume that $\theta'_s + \varepsilon_s \leqslant \hat{t}_1$ and $\theta'_i + \varepsilon_i \leqslant \theta'_{i+1}$, $i = 1, ..., s-1$. The initial value problem

$$\dot{x}(t) = f(t, x(t), u(t)), \quad t \in [t_0, t_1], \quad x(t_0) = x_0,$$

defines a function F, $F(t_0, x_0, t_1, \varepsilon) = x(t_1)$, where $\varepsilon = \mathrm{col}(\varepsilon_1, ..., \varepsilon_s)$. An obvious identity $P(t_1, t_0, x_0) \equiv F(t_0, x_0, t_1, 0)$ holds, where $P(t_1, t_0, x_0)$ denotes the value at t_1 of the solution of equation (1) with control $\hat{u}$, if x_0 is the state at t_0. We will now use a well known result of the theory of differential equations. It readily follows from the theorem on solution dependence on problem parameters that F is continuously differentiable in some set of the form $O_1 \times O_2$ where $O_1 \subset R^{n+2}$ is a neighborhood of the point $(\hat{t}_0, \hat{x}_0, \hat{t}_1)$, and $O_2 \subset R^s$ is the intersection of a neighborhood of the origin and the nonnegative orthant. We will calculate the right-hand derivatives of F w.r.t. ε_i, $i = 1, ..., s$, at $(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0)$. Fix i and assume $\varepsilon_j = 0$ for every $j \neq i$. Then, of course, $x(t) = \hat{x}(t)$, $t \leqslant \theta_i$. From the right-differentiability of solutions of (1) it follows that

$$x(\theta_i + \varepsilon_i) = \hat{x}(\theta_i) + \varepsilon_i f(\theta_i, \hat{x}(\theta_i), v_i) + o(\varepsilon_i)$$

$$\hat{x}(\theta_i + \varepsilon_i) = \hat{x}(\theta_i) + \varepsilon_i f(\theta_i, \hat{x}(\theta_i), \hat{u}(\theta_i)) + o(\varepsilon_i).$$

o is a common symbol for all error terms of order higher than one. The function P is of class C^1 in the third argument, and so

$$P(\hat{t}_1, \theta_i + \varepsilon_i, x(\theta_i + \varepsilon_i)) = P(\hat{t}_1, \theta_i + \varepsilon_i, \hat{x}(\theta_i))$$
$$+\varepsilon_i \partial_3 P(\hat{t}_1, \theta_i + \varepsilon_i, \hat{x}(\theta_i))^T f(\theta_i, \hat{x}(\theta_i), v_i) + o(\varepsilon_i)$$

$$P(\hat{t}_1, \theta_i + \varepsilon_i, \hat{x}(\theta_i + \varepsilon_i)) = P(\hat{t}_1, \theta_i + \varepsilon_i, \hat{x}(\theta_i)) \\ + \varepsilon_i \partial_3 P(\hat{t}_1, \theta_i + \varepsilon_i, \hat{x}(\theta_i))^T f(\theta_i, \hat{x}(\theta_i), \hat{u}(\theta_i)) + o(\varepsilon_i).$$

Subtracting the latter equality from the former one and using the continuity of derivative, we obtain

$$F(\hat{t}_0, \hat{x}_0, \hat{t}_1, \varepsilon_i e_i) - F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) = \varepsilon_i \partial_3 P(\hat{t}_1, \theta_i, \hat{x}(\theta_i))^T \Delta f(\theta_i, v_i) + o(\varepsilon_i),$$

where $\Delta f(\theta_i, v_i) = f(\theta_i, \hat{x}(\theta_i), v_i) - f(\theta_i, \hat{x}(\theta_i), \hat{u}(\theta_i))$, and e_i denotes the unit vector of the ith axis in R^s. The function F has been so far determined for $\varepsilon \geqslant 0$ only. We continue it onto the whole space in such a way that it is continuously differentiable in a neighborhood of $(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0)$ (see Appendix). Then

$$\partial_4 F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) = \begin{bmatrix} \partial_{\varepsilon_1} F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) & \cdots & \partial_{\varepsilon_s} F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) \end{bmatrix}^T$$

$$\partial_{\varepsilon_i} F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) = \partial_3 P(\hat{t}_1, \theta_i, \hat{x}(\theta_i))^T \Delta f(\theta_i, v_i), \quad i = 1, ..., s,$$

and it follows from the relationship between P and F that

$$\partial_{\varepsilon_i} F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0) = \partial_2 F(\theta_i, \hat{x}(\theta_i), \hat{t}_1, 0)^T \Delta f(\theta_i, v_i), \quad i = 1, ..., s. \tag{4}$$

We pose an auxiliary, $(2n + s + 2)$-dimensional optimization problem: the performance index

$$Q_a(t_0, x_0, t_1, x_1, \varepsilon) = q(t_0, x_0, t_1, x_1)$$

is minimized subject to

$$g(t_0, x_0, t_1, x_1) \leqslant 0, \quad -\varepsilon \leqslant 0, \quad h(t_0, x_0, t_1, x_1) = 0, \quad x_1 - F(t_0, x_0, t_1, \varepsilon) = 0.$$

The necessary optimality conditions of the point $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, 0)$ in the auxiliary problem are also necessary conditions of optimality of the point $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, \hat{u})$ in the original problem. We now formulate the Karush–John conditions. For $\lambda \in R$, $\mu \in R^r$, $\mu_1 \in R^s$, $\rho \in R^k$, $\rho_1 \in R^n$ define a Lagrange function $L = \lambda q + \mu^T g + \rho^T h - \mu_1^T \varepsilon + \rho_1^T (x_1 - F)$. The Karush–John theorem says that if $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, 0)$ is an optimal solution of the auxiliary problem, then there are $\lambda \in R$, $\mu \in R^r$, $\mu_1 \in R^s$, $\rho \in R^k$, $\rho_1 \in R^n$, such that

$$\lambda \geqslant 0, \quad \mu \geqslant 0, \quad \mu_1 \geqslant 0 \tag{5}$$

$$\lambda + \|\mu\| + \|\mu_1\| + \|\rho\| + \|\rho_1\| > 0 \tag{6}$$

$$\mu^T g(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1) = 0 \tag{7}$$

$$\frac{\partial L}{\partial t_0} = \partial_1(\lambda q + \mu^T g + \rho^T h) - \rho_1^T \partial_1 F = 0 \tag{8}$$

$$\frac{\partial L}{\partial x_0} = \partial_2(\lambda q + \mu^T g + \rho^T h) - (\partial_2 F)\rho_1 = 0 \tag{9}$$

$$\frac{\partial L}{\partial t_1} = \partial_3(\lambda q + \mu^T g + \rho^T h) - \rho_1^T \partial_3 F = 0 \tag{10}$$

$$\frac{\partial L}{\partial x_1} = \partial_4(\lambda q + \mu^T g + \rho^T h) + \rho_1 = 0 \tag{11}$$

$$\frac{\partial L}{\partial \varepsilon} = -\mu_1 - (\partial_4 F)\rho_1 = 0. \tag{12}$$

All derivatives in (5)-(12) are computed at $(\hat{t}_0, \hat{x}_0, \hat{t}_1, \hat{x}_1, 0)$. Condition (6) is equivalent to $\lambda + ||\,\mu\,|| + ||\,\rho\,|| \ > 0$. Indeed, if $\lambda + ||\,\mu\,|| + ||\,\rho\,|| \ = 0$, then $\rho_1 = 0$ by virtue of (5) and (11), and from (12) it follows that $\mu_1 = 0$. Further, the condition $\lambda + ||\,\mu\,|| + ||\,\rho\,|| \ > 0$ may be equivalently replaced by the equality

$$\lambda + ||\,\mu\,|| + ||\,\rho\,|| \ = 1. \tag{13}$$

This is a consequence of the fact that if $T = (\lambda, \mu, \mu_1, \rho, \rho_1)$ satisfies (5)-(12), then T divided by $\lambda + ||\,\mu\,|| + ||\,\rho\,||$ also satisfies (5)-(12).

Let the sequences $\theta_1, ..., \theta_s$ and $v_1, ..., v_s$ be determined as in the beginning of the proof, and let J be the sequence of pairs (θ_i, v_i), $i = 1, ..., s$. Denote the family of all such sequences (finite, of different lengths) by $\Im$. For every $J \in \Im$, denote by $M(J)$ the set of all triples (λ, μ, ρ) satisfying (5), (13), (7) – (12). $M(J)$ is nonempty and compact. For $J^1, J^2 \in \Im$ we write $J^1 \prec J^2$, if J^1 is a subsequence of J^2. In consequence of the obvious implication $J^1 \prec J^2 \Rightarrow M(J^1) \supset M(J^2)$, the family of sets $M(J)$, $J \in \Im$, is centered. Indeed, for any $J^1, ..., J^j \in \Im$, it is straightforward to construct a sequence $J \in \Im$ such that $J^i \prec J$, $i = 1, ..., j$. Then, $M(J) \neq \emptyset$ and $M(J) \subset M(J^i)$, $i = 1, ..., j$. By the Finite Intersection Property known in topology (Edwards, 1995, Ch. 0.1; Yosida, 1965, Ch. 2, p. 6) it is proved that the set

$$\mathcal{M} = \bigcap_{\mathcal{J} \in \Im} \mathcal{M}(\mathcal{J})$$

is nonempty.

We now select $(\lambda, \mu, \rho) \in \mathcal{M}$, determine ρ_1 from (11) and define the adjoint function by the equality $\hat{\psi}(t) = \partial_2 F(t, \hat{x}(t), \hat{t}_1, 0) \cdot \rho_1$. By virtue of the definition of F,

$$\hat{\psi}(t) = \partial_3 P(\hat{t}_1, t, \hat{x}(t))\,\rho_1 = \Phi(\hat{t}_1, t)\rho_1.$$

It follows from (3) that $\hat{\psi}$ satisfies the adjoint equation $\dot{\hat{\psi}}(t) = -A(t)\hat{\psi}(t)$, that is, (iv) holds. From (9) and (11) we obtain the first two transversality conditions (v). To prove the third condition, notice that the derivative of the function $t_0 \mapsto F(t_0, \hat{x}(t_0), \hat{t}_1, 0)$, computed along the trajectory $\hat{x}$ vanishes

$$\frac{\mathrm{d}}{\mathrm{d}t_0} F(t_0, \hat{x}(t_0), \hat{t}_1, 0)\bigg|_{t_0 = \hat{t}_0} = \partial_1 F + (\partial_2 F)^T f[\hat{t}_0] = 0,$$

whence $\rho_1^T\partial_1 F + \rho_1^T(\partial_2 F)^T f[\hat{t}_0] = 0$. Now it suffices to substitute the definition of $\hat{\psi}$ and use (8). The fourth transversality condition results from (10). The maximum condition (vi) is obtained from the relationships (12) and (5) using $s = 1$, $t = \theta_1$, $v = v_1$. They immediately yield $(\partial_4 F)\rho_1 \leqslant 0$, and further, by virtue of (4),
$\rho_1^T\partial_4 F(\hat{t}_0, \hat{x}_0, \hat{t}_1, 0)^T = \rho_1^T\partial_2 F(t, \hat{x}(t), \hat{t}_1, 0)^T\Delta f(t, v) = \hat{\psi}(t)^T\Delta f(t, v) \leqslant 0.$ □

Theorem 1 also applies to problems in which inequality and/or equality constraints do not appear. If there are no inequality constraints, the variable μ is omitted in the formulation of the theorem, together with the complementarity condition and the term $\mu^T g$ in the transversality conditions. To justify this, introduce an always redundant inequality constraint $g(t_0, x_0, t_1, x_1) \equiv -1 \leqslant 0$. Then, $\mu = 0$ by the complementarity condition. If there are no equality constraints, the variable ρ and the term $\rho^T h$ in the transversality conditions are skipped. This can be justified by referring to the Karush–John conditions used in the proof. Irrespective of the fact whether there is a constraint of the form $h(t_0, x_0, t_1, x_1) = 0$ in the original problem, the Karush–John conditions for the finite-dimensional problem with an equality constraint $x_1 = F(t_0, x_0, x_1, \varepsilon)$ are used in the proof. If constraints of both kinds are absent, we additionally put $\lambda = 1$.

4. Time variation of H on extremal solutions

A triple (u, x, ψ) is called an *extremal solution* in $[t_0, t_1]$, if $u \in PC(t_0, t_1; D)$, and u, x and ψ in $[t_0, t_1]$ satisfy the state equation (1), the adjoint equation

$$\dot{\psi}(t) = -\partial_2 H(t, x(t), u(t), \psi(t)),$$

and the maximum condition

$$H(t, x(t), v, \psi(t)) \leqslant H(t, x(t), u(t), \psi(t)) \quad \forall t \in [t_0, t_1] \quad \forall v \in D.$$

Theorem 2. *Assume that f, $\partial_1 f$ and $\partial_2 f$ are continuous. If (u, x, ψ) is an extremal solution in $[t_0, t_1]$, then the function $\chi(t) = H(t, x(t), u(t), \psi(t))$ is continuous in $[t_0, t_1]$, and has a derivative*

$$\dot{\chi}(t) = \partial_1 H(t, x(t), u(t), \psi(t)),$$

continuous in every interval of continuity of u.

Proof. The function χ is obviously continuous in every interval of continuity of the control u. Let θ be an arbitrary point of control discontinuity. Assume, contrary to the claim, that χ has a jump at θ, e.g.,

$$H(\theta, x(\theta), u(\theta-), \psi(\theta)) > H(\theta, x(\theta), u(\theta+), \psi(\theta)).$$

Thus, there are t' and t'' sufficiently close to one another, such that $t' < \theta < t''$ and

$$H(t'', x(t''), u(t'), \psi(t'')) > H(t'', x(t''), u(t''), \psi(t'')),$$

but this is in contradiction with the maximum condition (vi). The function χ is therefore continuous.

Let θ_1 and θ_2 be two time moments from an interval of continuity of u. Denote $\Delta t = \theta_2 - \theta_1$ and $\psi_i = \psi(\theta_i)$, $x_i = x(\theta_i)$, $u_i = u(\theta_i)$,

$$\Delta_i = H(\theta_i, x_i, u_2, \psi_i) - H(\theta_i, x_i, u_1, \psi_i)$$

for $i = 1$, 2. Write the expression $\Delta H = H(\theta_2, x_2, u_2, \psi_2) - H(\theta_1, x_1, u_1, \psi_1)$ in two ways:

$$\begin{aligned}\Delta H &= \Delta_2 + H(\theta_2, x_2, u_1, \psi_2) - H(\theta_1, x_1, u_1, \psi_1)\\ &= \Delta_1 + H(\theta_2, x_2, u_2, \psi_2) - H(\theta_1, x_1, u_2, \psi_1).\end{aligned}$$

For every v, the function $t \mapsto H(t, x(t), v, \psi(t))$ has a continuous derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} H(t, x(t), v, \psi(t)) = \partial_1 H + \partial_2 H^T \dot{x} + \partial_4 H^T \dot{\psi} = \partial_1 H.$$

Then

$$\begin{aligned}\Delta H &= \Delta_2 + \partial_1 H(\theta_1, x_1, u_1, \psi_1)\, \Delta t + o(\Delta t)\\ &= \Delta_1 + \partial_1 H(\theta_1, x_1, u_2, \psi_1)\, \Delta t + o(\Delta t).\end{aligned} \tag{14}$$

The function $v \mapsto \partial_1 H(\theta_1, x_1, v, \psi_1)$ is continuous by assumption, and so

$$\partial_1 H(\theta_1, x_1, u_2, \psi_1)\, \Delta t = \partial_1 H(\theta_1, x_1, u_1, \psi_1)\, \Delta t + o(\Delta t). \tag{15}$$

From (14) and (15) it follows that $\Delta_2 - \Delta_1 = o(\Delta t)$. In virtue of the maximum condition, $\Delta_1 \leqslant 0$ and $\Delta_2 \geqslant 0$. Both these expressions are thus of the order higher than one and in consequence

$$\Delta H = \partial_1 H(\theta_1, x_1, u_1, \psi_1)\, \Delta t + o(\Delta t),$$

which ends the proof. □

5. Conclusions

The here presented proof of the maximum principle is suitable for teaching purposes. It uses classical "textbook" results from the theory of ordinary differential equations, mathematical programming, and, to a smaller extent, topology, to replace the most troublesome parts of typical proofs. The rest of the presented proof is elementary, only basics of differential calculus are required. This is why it comes naturally and without special difficulties in a typical course of study, where the student listening to lectures on optimal control has already mastered the necessary elements of differential equations and finite dimensional optimization. Due to relying on facts from other domains of mathematics, the proof is also relatively short.

Appendix

Let m, n and s be positive integers, $R_+ = [0, \infty[$, and let $F : R^m \times R_+^s \to R^n$ be a C^1 function. Our aim is to construct a C^1 extension of F onto the whole space $R^m \times R^s$. The construction will be done in steps. In the first step F is extended onto $R^m \times R \times R_+^{s-1}$

$$F(x,y) = 2F(x, 0, y_2, y_3, ..., y_s) - F(x, -y_1, y_2, y_3, ..., y_s),$$

$$x \in R^m,\ y_1 < 0,\ y_2 \geqslant 0,\ y_3 \geqslant 0,\ ...,\ y_s \geqslant 0.$$

In step 2, F is extended onto $R^m \times R^2 \times R_+^{s-2}$

$$F(x,y) = 2F(x, y_1, 0, y_3, ..., y_s) - F(x, y_1, -y_2, y_3, ..., y_s),$$

$$x \in R^m,\ y_1 \in R,\ y_2 < 0,\ y_3 \geqslant 0,\ ...,\ y_s \geqslant 0.$$

Generally, in step i, $i = 1, 2, ..., s$,

$$F(x,y) = 2F(x, y_1, ..., y_{i-1}, 0, y_{i+1}, ..., y_s) - F(x, y_1, ..., y_{i-1}, -y_i, y_{i+1}, ..., y_s),$$

$$x \in R^m,\ y_1, ..., y_{i-1} \in R,\ y_i < 0,\ y_{i+1} \geqslant 0,\ ...,\ y_s \geqslant 0. \tag{A1}$$

After the ith step, F is well defined on $R^m \times R^i \times R_+^{s-i}$, and so, after the sth step, F is well defined on the whole space $R^m \times R^s$. The extended function F is obviously continuous. To prove that it is continuously differentiable, it is enough to notice that due to (A1), the right and left partial derivatives of F w.r.t. y_i at $y_i = 0$ are identical.

References

DUBOVITSKII, A.YA. and MILYUTIN, A.A. (1965) Extremum problems with restrictions. *USSR Comput. Math. and Math. Phys.* **5**(3), 1-80.

EDWARDS, R.E. (1995) *Functional Analysis: Theory and Applications.* Dover Publ., New York.

MANGASARIAN, L. and FROMOVITZ, S. (1967) The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints. *J. Mathematical Analysis and Applications* **17**, 37-47.

MILYUTIN, A.A., DMITRUK, A.V. and OSMOLOVSKII, N.P. (2004) *Maximum Principle in Optimal Control.* Moscow State University, Moscow (in Russian).

PONTRYAGIN, L.S., BOLTYANSKII, V.G., GAMKRELIDZE, R.V. and MISH–CHENKO, E.F. (1961) *The Mathematical Theory of Optimal Processes.* Nauka, Moscow (in Russian, first English language edition: John Wiley & Sons, Inc., New York, 1962).

YOSIDA, K. (1965) *Functional Analysis.* Springer, Berlin.