Note del corso of Calcolo delle Variazioni, a.a. 2012-13 — A. Visintin
Redazione del febbraio 2013

Queste note costituiscono solo una traccia delle lezioni. Per una presentazione più completa il lettore è rimandato alle opere in bibliografia. L'asterisco indica i complementi.

Table of Contents

I. Classical one-dimensional calculus of variations
I.1. Introduction
I.2. Integral functionals
I.3. Classical conditions for minimization
I.4. First integral and Lagrange multipliers
I.5. Examples
I.6. Boundary conditions
I.7. Noether Theorem
I.8. Legendre transformation and canonical equations

II. Analytical mechanics and optics
II.1. Analytical mechanics
* II.2. Analytical optics
* II.3. From the Maxwell equations to geometrical optics
II.4. Hamilton-Jacobi equation

III. Minimization, variational inequalities and Γ-Convergence
III.1. Direct method of minimization (Tonelli theorem)
III.2. Variational inequalities (Lions-Stampacchia theorem)
III.3. The obstacle problem
* III.4. De Giorgi's Γ-convergence in metric spaces
* III.5. Ekeland's minimization principle

IV. Optimal control
IV.1. Control problems
* IV.2. Linear optimal control problems (Lions's theory)
IV.3. Dynamic programming and time-discrete Bellman's theory
IV.4. Time-continuous Bellman's theory
IV.5. Pontryagin's maximum principle
IV.6. Pontryagin's equations
* IV.7. Differential games (Isaacs's theory)

V. Elements of convex calculus
V.1. Convex lower semicontinuous functions
* V.2. Fenchel's transformation
* V.3. Subdifferential

Acronyms

B-function: Bellman function
B-equation: Bellman equation
E-L equation: Euler-Lagrange equation
H-J equation: Hamilton-Jacobi equation
H-J-B equation: Hamilton-Jacobi -Bellman equation
PDP: principle of dynamic programming
ODE: ordinary differential equation
PDE: partial differential equation

We shall denote by $D_1, ..., D_N$ the partial derivatives with respect to $x_1, .., x_N$. If the latter are M-dimensional vectors, we shall denote by D_{ij} the partial derivatives with respect to the jth component of x_i, for $i = 1, .., N$ and $j = 1, .., M$.

I. ONE-DIMENSIONAL CALCULUS OF VARIATIONS

I.1. Introduction

This course will touch several issues of the classical and the modern theory of calculus of variations, which also includes aspects that are at the interface with optimization, convex analysis, variational inequalities, and other subjects. This field if relevant in itself as well as for applications.

We start with the classical one-dimensional theory, including fundamental results of Euler, Lagrange, and others. We then illustrate applications to analytical mechanics and analytical optics, including the Hamilton and Jacobi theories.

Afterwards we introduce Tonelli's direct method of the calculus of variations and Stampacchia's variational inequalities. We also outline De Giorgi's theory of Γ-convergence and Ekeland's variational principle in metric spaces.

We then deal with optimal control, including fundamental results of Lions, Bellman and Pontryagin; we also outline differential games. Notice that differential games generalize optimal control, which in turn generalizes the basic Bolza problem of calculus of variations. A common thread may be found in the calculus of variations, optimal control, differential games and their applications to physics and economics, and so on: the possibility of formulating a Hamilton-Jacobi-type PDE and the equivalent approach in terms of a canonical-type system of ODEs.

In the final part we introduce some basic elements of convex calculus: the Legendre-Fenchel transformation and the notion of subdifferential.

I.2. Integral functionals

Let us define the Lagrange functional

$$\begin{aligned} &F(u) = \int_a^b f(x, u(x), u'(x))\, dx \\ &\forall u \in V := \{v \in C^1([a,b]) : v(a) = \alpha, v(b) = \beta\}, \end{aligned} \tag{2.1}$$

with $a < b$, $\alpha, \beta \in \mathbf{R}$ prescribed, and $f \in C^0([a,b] \times \mathbf{R} \times \mathbf{R})$ a given function.

More generally, one may deal with:

(i) vector-valued functions $[a,b] \to \mathbf{R}^N$, for some integer $N \geq 1$. In this case one assumes that $f \in C^0([a,b] \times \mathbf{R}^N \times \mathbf{R}^N)$, and $\alpha, \beta \in \mathbf{R}^N$;

(ii) either scalar- or vector-valued functions u that depend on several variables. In this case the ordinary derivative u' is replaced by the gradient ∇u;

(iii) functions u that belong to other function spaces, in particular Sobolev spaces;

(iv) more general boundary conditions;

(v) more general functionals than F; and so on.

On the history of the Calculus of Variations. Classical one-dimensional problems include Dido's isoperimetric problem, the problem of the curve of minimal length joining two prescribed points, the problem of the curve *brachystochrone,* and so on (see references).

Among the multidimensional problems we mention the classical *Dirichlet problem,* which in its original formulation consisted in minimizing the integral

$$I(u) = \int_\Omega |\nabla u(x)|^2 \, dx \qquad u \in C^1(\Omega) \cap C^0(\bar{\Omega}) \tag{2.2}$$

(with Ω domain of $\mathbf{R}^3$), and prescribing the value of u on (a part of) the boundary. One may show that any minimizing function of (2.2) is harmonic, namely fulfills the equation $\Delta u = 0$ in Ω. Problems like this need not have a solution in C^k spaces.

Traditionally one distinguishes between:

(i) classical or indirect methods (essentially before 1900): study of either necessary or sufficient conditions for minimizers, typically formulated in terms of equations or inequalities; this assumed the existence of a minimizer;

(ii) direct methods (essentially after 1900): these are aimed to establish the existence of minimizers, typically via the use of minimizing sequences and the topological notions of compactness and lower semicontinuity. Here larger spaces than the C^ks are usually involved, e.g., L^p and Sobolev spaces.

The Calculus of Variations was developed by analysts and mathematical-physicists, at a lucky time when there was no sharp distinction between these disciplines. This left traces in terminology and notation.

Examples. Let us define the functionals

$$\begin{aligned} &F_1(u) = \int_{-1}^{1} x^2 u'(x)^2 \, dx \\ &\forall u \in V_1 := \{v \in C^1([-1,1]) : v(-1) = 0, v(1) = 1\}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} &F_2(u) = \int_0^1 \frac{1}{1+u'(x)^2} \, dx \\ &\forall u \in V_2 := \{v \in C^1([0,1]) : v(0) = 0, v(1) = 1\}. \end{aligned} \tag{2.4}$$

It is easy to see that both functionals have no minimum and no maximum.

Two lemmata.

Lemma 2.1. *(Fundamental lemma of the Calculus of Variations) If $g \in C^0([a,b])$ is such that*

$$\int_a^b g(x)v(x) \, dx = 0 \qquad \forall v \in C^0([a,b]), v(a) = v(b) = 0, \tag{2.5}$$

then $g \equiv 0$ in $[a,b]$.

Proof. By contradiction, let $\bar{x} \in [a,b]$ exist such that $g(\bar{x}) \neq 0$. Then g has constant sign in a suitable neighborhood U of $\bar{x}$. Selecting a function v which is positive in U and vanishes elsewhere, we get $\int_a^b g(x)v(x)\,dx \neq 0$, at variance with the assumption. □

Lemma 2.2. *(Du Bois-Reymond lemma) If $g \in C^0([a,b])$ is such that*

$$\int_a^b g(x)v'(x)\,dx = 0 \qquad \forall v \in C^1([a,b]), v(a) = v(b) = 0, \tag{2.6}$$

then g is constant in $[a,b]$.

If $g \in C^1([a,b])$ this assertion is a simple consequence of the previous lemma. If g is just continuous, this result may be proved via a small trick, or by a regularization procedure.

Variations. Let $a, b, \alpha, \beta \in \mathbf{R}$ with $a < b$, assume that $f \in C^1([a,b]\times\mathbf{R}\times\mathbf{R})$, and search for a minimizer of the functional

$$\begin{aligned} &F(u) = \int_a^b f(x, u(x), u'(x))\,dx \\ &\forall u \in V_{\alpha,\beta} := \{v \in C^1([a,b]) : v(a) = \alpha, v(b) = \beta\}. \end{aligned} \tag{2.7}$$

The linear space $V_{0,0}$ is associated to the affine space $V_{\alpha,\beta}$. We shall deal with the minimization of this functional in this (affine) function space; this is known as the *Lagrange problem.*.

Let us introduce the auxiliary function

$$\varphi_{u,v}(t) := F(u + tv) \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}, \forall t \in \mathbf{R}, \tag{2.8}$$

and set

$$\delta F(u,v) := \varphi'_{u,v}(0) \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}. \tag{2.9}$$

When existing, $\delta F(u,v)$ is called the first variation of F in u, with respect to the (infinite-dimensional) vector v. This extends the notion of derivative with respect to finite-dimensional vectors, that is already known to the reader for functions $\mathbf{R}^N \to \mathbf{R}$.

Similarly, we set

$$\delta^2 F(u,v) := \varphi''_{u,v}(0) \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}; \tag{2.10}$$

if existing, this is called the second variation of F at u with respect to the (infinite-dimensional) vector v.

It is easily checked that

$$\delta F(u,v) = \int_a^b (vD_2 f + v' D_3 f)\,dx \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}, \tag{2.11}$$

$$\delta^2 F(u,v) = \int_a^b [v^2 D_2^2 f + 2vv' D_2 D_3 f + (v')^2 D_3^2 f]\,dx \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}. \tag{2.12}$$

Exercise. Let $f : \mathbf{R}^2 \to \mathbf{R}$ be of class C^2. If $(0,0)$ is a point of relative minimum for the restrictions of f to all straight lines through the origin, does it follow that $(0,0)$ is also a local minimizer of f in the plane? (i.e., is f minimized in some neighborhood of $(0,0)$?)

Hint: consider Peano's function $f(x,y) = (y - x^2)(y - 2x^2)$ for any $(x,y) \in \mathbf{R}^2$.

I.3. Classical conditions for minimization

Proposition 3.1. *(Generalized Fermat principle) Let the functional F be defined as in (2.7), with $f \in C^1([a,b]\times\mathbf{R}\times\mathbf{R})$. For any extremum u of F (i.e., a point of either maximum or minimum), then*

$$\delta F(u,v) = 0 \qquad \forall v \in V_{0,0}. \tag{3.1}$$

Proof. By the definition (2.9) of the first variation, it suffices to apply the classical Fermat principle to the function $\varphi_{u,v}$ (cf. (2.8)). □

Because of the theorem of passage to the limit under integral, it is easy to see that (3.1) entails the *Euler-Lagrange equation* (more shortly, "E-L equation") in *weak form:*

$$(\delta F(u;v) =) \int_a^b [D_2 f(x,u(x),u'(x))v(x) + D_3 f(x,u(x),u'(x))v'(x)]\,dx = 0 \qquad \forall v \in V_{0,0}. \tag{3.2}$$

This equation is also equivalent to the following integral equation (named after Du Bois-Reymond, and easily deducible from the Du Bois-Reymond Lemma 2.2), for a suitable real constant C:

$$D_3 f(x,u(x),u'(x)) = C + \int_a^x D_2 f(s,u(s),u'(s))\,ds \qquad \forall x \in [a,b]. \tag{3.3}$$

If $f \in C^2([a,b]\times\mathbf{R}\times\mathbf{R})$ and $u \in V_{\alpha,\beta} \cap C^2([a,b])$, differentiating (3.3) with respect to x one gets the *E-L equation in strong form:* [(1)]

$$\frac{d}{dx} D_3 f(x,u(x),u'(x)) = D_2 f(x,u(x),u'(x)) \qquad \forall x \in [a,b]. \tag{3.4}$$

Conversely, if $f \in C^2([a,b]\times\mathbf{R}\times\mathbf{R})$, $u \in V_{\alpha,\beta} \cap C^2([a,b])$ and the strong form of the E-L equation is fulfilled, then obviously the weak form of the E-L equation is also satisfied.

By what we saw, any extremum fulfills the E-L equation in weak form if $u \in C^1([a,b])$, and that in strong form if $f \in C^2([a,b]\times\mathbf{R}\times\mathbf{R})$ and $u \in C^2([a,b])$. Any solution of the E-L equation in weak (strong, resp.) form is called a weak (strong, resp.) *extremal.* [(2)] Therefore, under suitable regularity conditions,

$$\text{any extremum is an extremal, but not conversely.} \tag{3.5}$$

However we have the next statement.

Proposition 3.2.

$$\textit{If } F \textit{ is convex, then any extremal is a minimizer (so it is an extremum).} \tag{3.6}$$

(1) In alternative, (3.4) may be derived directly from (3.1) via the fundamental lemma of the calculus of variations.

(2) So there are extrema, extremals, and also ... the extremes of the interval.

If $f(x,\cdot,\cdot)$ is convex for any x, then F is convex. If $f(x,\cdot,\cdot)$ is strictly convex for any x, then the minimizer (if existing) is unique.

Proof. (3.6) is obvious for functions $\mathbf{R}\to\mathbf{R}$. In our case it suffices to notice that for any $u\in V_{\alpha,\beta}$ and any $v\in V_{0,0}$ the auxiliary function $\mathbf{R}\to\mathbf{R}: t\mapsto \varphi_{u,v}(t):=F(u+tv)$ is convex. The proof of the second part is straightforward. □

Classically the E-L equation has been used in the search for extrema: if an extremum exists then it fulfills the E-L equation, and may thus be searched for among the solutions of that equation.

*** On the integral formulation of the Euler-Lagrange equation.** The integral equation (3.3) and the equivalent weak equation (3.4) characterize extremals of F also if $C^1([a,b])$ is replaced either by $C^1_{pw}([ab])$ (namely the space of continuous functions $[a,b]\to\mathbf{R}$ that are piecewise of class C^1), or by the wider space of Lipschitz functions. In this case (3.3) just holds a.e. in $]a,b[$.

In order to write the strong equation (3.4) it is not needed that f,u be of class C^2. [(3)] The hypothesis that $D_3f(x,u(x),u'(x))$ be of class C^1 is actually needed for the application of the fundamental lemma; anyway, instead of that lemma one may use Du Bois-Reymond's lemma.

Moreover existence and continuity of the left member of (3.4) follow from (3.4) itself, by comparing the terms of this equation. In any case, without further regularity of f and u, the derivative at the left side cannot be developed. One may then write (3.4) in either weak or integral form.

E-L equation for vector functions. Let $f\in C^1([a,b]\times\mathbf{R}^N\times\mathbf{R}^N)$ and $u\in C^1([a,b])^N$, for some integer $N\geq 1$. With reference to the function

$$[a,b]\times\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{R}:(x,u,\xi)\mapsto f(x,u,\xi),$$

let us denote by D_{2j} (D_{3j}, resp.) the partial derivative with respect to the j-th component of the vector u (ξ, resp.), for $j=1,...,N$. By an obvious extension of the previous procedure, if u is an extremum of F (defined as in (2.1)) we infer the E-L equation in weak form

$$\int_a^b [D_{2j}f(x,u(x),u'(x))v_j(x)+D_{3j}f(x,u(x),u'(x))v_j'(x)]\,dx=0 \qquad \forall v\in C^1([a,b])^N \text{ such that } v(0)=v(b)=0, \tag{3.7}$$

implying the sum (from 1 to N) over repeated indices (Einstein's convention). This weak equation is also equivalent to an integral equation which is analogous to (3.3).

If $f\in C^2([a,b]\times\mathbf{R}^N\times\mathbf{R}^N)$ and $u\in V_{\alpha,\beta}\cap C^2([a,b])^N$, then here also the weak E-L equation is equivalent to the strong E-L equation

$$\frac{d}{dx}D_{3j}f(x,u(x),u'(x))=D_{2j}f(x,u(x),u'(x)) \qquad \forall x\in[a,b],\ \text{for } j=1,...,N. \tag{3.8}$$

Also in this case one speaks of weak and strong extremals, and the above results are extended to vector-valued functions in an obvious way.

(3) Anyway so is written in several textbooks, where by writing (3.4) it is implied that the derivative at the left side is developed via the theorem of the composed function.

The E-L equation can also be extended to functionals that depend on higher-order derivatives of u, to different boundary conditions, to functions u of several variables, and so on. For instance, let us consider the problem

$$\begin{aligned} &\widehat{F}(u) = \int_a^b f(x, u(x), u'(x), u''(x))\, dx \\ &\forall u \in V := \{v \in C^2([a,b]) : v(a) = \alpha, v(b) = \beta\} \end{aligned} \tag{3.9}$$

for a given function $f \in C^4([a,b]\times\mathbf{R}\times\mathbf{R}\times\mathbf{R})$, and prescribed $a < b$, $\alpha, \beta \in \mathbf{R}$. (Different boundary conditions may also be considered.) It is easy to see that any extremal $u \in V \cap C^4([a,b])$ of $\widehat{F}$ fulfills the following E-L equation in strong form (written implying the argument $(x, u(x), u'(x), u''(x))$ of f and of its derivatives):

$$\frac{d}{dx} D_3 f - \frac{dx^2}{dx^2} D_4 f = D_2 f \qquad \forall x \in [a,b], \text{ for } j = 1, ..., N. \tag{3.10}$$

More generally, if the integrand depends on derivatives up to an order k and $f \in C^{2k}$, then the E-L equation in strong form is an ODE of order $2k$.

A necessary condition for minimization.

Theorem 3.3. *(Legendre condition) If $f \in C^2([a,b]\times\mathbf{R}^N\times\mathbf{R}^N)$, then for any minimizer u of F*

$$\textit{the Hessian matrix } D_3^2 f(x, u, u') \textit{ is positive semidefinite.} \tag{3.11}$$

Proof. If u is a minimizer of F then

$$\delta^2 F(u, v) \geq 0 \qquad \forall v \in V_{0,0}. \tag{3.12}$$

If $f \in C^3([a,b]\times\mathbf{R}^N\times\mathbf{R}^N)$ and $u \in V_{\alpha,\beta}\cap C^2([a,b])$, noticing that $2vv' = (v^2)'$ and integrating by parts the second addendum of (2.12), we get (implying the argument $(x, u(x), u'(x))$ of f and its derivatives)

$$\delta^2 F(u, v) = \int_a^b \Big[v^2 \Big(D_2^2 f - \frac{d}{dx} D_2 D_3 f \Big) + (v')^2 D_3^2 f \Big]\, dx \qquad \forall u \in V_{\alpha,\beta}, \forall v \in V_{0,0}. \tag{3.13}$$

These derivatives of f are obviously bounded. As $(v')^2$ may be arbitrarily *large* also where v^2 is *small,* (3.11) must then hold for any minimizer (i.e., point of minimum) of F. □

The Legendre condition (3.11) is just a necessary condition for the minimization of F; actually, freely speaking, the v^2-term may prevail over the $(v')^2$-term in (3.13): $(v')^2$ may be arbitrarily *small* also where v^2 is *large*.. Even if the Hessian $D_3^2 f(x, u, u')$ were positive definite for any $x \in [a,b]$, one could not infer that u is a minimizer.

A counterexample is provided by the quadratic functional $F(u) = \int_0^{2\pi} [(u')^2 - u^2]\, dx$; here the minimization sees the first-order term competing with the zero-order one. In this case $u_0 \equiv 0$ fulfills the E-L equation, and is thus an extremal of F. On the other hand (as the integrand is a homogeneous function of degree two)

$$\delta^2 F(u_0, v) = \int_0^{2\pi} [(v')^2 - v^2]\, dx \qquad \forall v \in V_{0,0}; \tag{3.14}$$

setting $v(x) = \sin(x/2)$, it is easily checked that

$$\delta^2 F(u_0, v) = \int_0^{2\pi} \left\{ \frac{1}{4} \cos^2(x/2) - \sin^2(x/2) \right\} dx < 0. \tag{3.15}$$

Exercise. Study the minimization of the functionals

$$\begin{aligned} &G_1(u) = \int_{-1}^{1} u(x)^2 [1 - u'(x)]^2 \, dx \\ &\forall u \in W_1 := \{v \in C^1_{pw}([-1,1]) : v(-1) = 0, v(1) = 1\}, \end{aligned} \tag{3.16}$$

$$\begin{aligned} &G_2(u) = G_1(u) \\ &\forall u \in W_2 := \{v \in C^1([-1,1]) : v(-1) = 0, v(1) = 1\}. \end{aligned} \tag{3.17}$$

$$\begin{aligned} &G_3(u) = \int_0^1 u(x)^2 x \, dx \\ &\forall u \in W_3 := \{v \in C^1([0,1]) : v(0) = 0, v(1) = 1\}. \end{aligned} \tag{3.18}$$

(G_1 has a minimum in $u_0(x) = (x-1)^+$. As $u_0 \notin C^1([-1,1])$, one may see that $\inf G_2 = 0$ but G_2 has no minimizer. G_3 has no minimizer too, also if $\inf G_3 = 0$.)

Relative minimizers. So far, we just dealt with *absolute* (or global) minimizers. Next we consider *relative* (or local) minimizers, with respect to a prescribed topology; here we confine ourselves to the scalar setting (i.e., $N = 1$). Any $u_0 \in C^1([a,b])$ is called a *strong (weak, resp.) relative minimizer* if and only if there exists a neighborhood U of u_0 in the topology of $C^0([a,b])$ [(4)] ($C^1([a,b])$, resp.) such that u_0 is an absolute minimizer for the restriction of F to U.

As the topology of $C^1([a,b])$ is strictly finer than that of $C^0([a,b])$, every strong relative minimizer is also a weak relative minimizer. The converse fails; see e.g. the counterexample in [De 100]. Using this terminology (which is standard in the Calculus of Variations), thus

absolute minimizer $\Rightarrow$ *strong relative minimizer* $\Rightarrow$ *weak relative minimizer,*

and all of these implications are strict.

Other second order conditions. [(5)] Let us consider the functional F defined in (2.1). In analogy with what is known from the basic course of analysis for functions of $C^2(\mathbf{R}^N)$, any strong relative minimizer u of F fulfills:

(i) the first-order extremality condition

$$\delta F(u, v) = 0 \qquad \forall v \in V_{0,0} \qquad \text{(equivalent to the E-L equation)}, \tag{3.19}$$

(ii) the second-order condition [(6)]

$$\delta^2 F(u, v) \geq 0 \qquad \forall v \in V_{0,0}. \tag{3.20}$$

[(4)] This is the topology that $C^0([a,b])$ induces onto the subset $C^1([a,b])$.

[(5)] This outline is especially synthetical. A more detailed presentation may be found e.g. in [De 98-113].

[(6)] here we just deal with scalar functions u scalari.

The converse fails, even if (at variance with what is known from the basic course of analysis for functions of $C^2(\mathbf{R}^N)$) the latter condition is replaced by the stronger

$$\delta^2 F(u,v) > 0 \qquad \forall v \in V_{0,0}; \tag{3.21}$$

see e.g. the counterexample in [De 104]. At the end of this section we shall see that however a suitable strengthening of this inequality provides a sufficient condition for an extremal to be a strong relative minimizer.

In any case, if the functional F is convex then any extremal is a minimizer.

Theorem 3.4. *(Legendre) If $f \in C^2$ and u_0 is a weak relative minimizer, then (compare with Theorem 3.3 for $N = 1$)*

$$D_3^2 f(x, u_0(x), u_0'(x)) \geq 0 \qquad \forall x \in]a,b[. \tag{3.22}$$

The converse fails (we already saw a counterexample).

In order to display the next result, first for any $f \in C^1$ let us define the Weierstrass *excess function*

$$\begin{aligned} E(x,u,\xi,\eta) := f(x,u,\eta) - f(x,u,\xi) - (\eta - \xi) D_3 f(x,u,\xi) \\ \forall (x,u,\xi,\eta) \in [a,b] \times \mathbf{R}^3. \end{aligned} \tag{3.23}$$

Theorem 3.5. *(Weierstrass) If $f \in C^1$ and u_0 is a strong relative minimizer, then*

$$E(x, u_0(x), u_0'(x), \eta) \geq 0 \qquad \forall x \in]a,b[, \forall \eta \in \mathbf{R}. \tag{3.24}$$

This fails for weak relative minimizers. The converse implication also fails.

Notice that (3.24) means that, for any x, in the plane (u, f) the straight line that is tangent to the graph of $f(x, u_0(x), \cdot)$ at the point $(u_0'(x), f(x, u_0(x), u_0'(x))$ stays below that graph everywhere. (3.24) then entails that $D_3^2 f(x, u_0(x), u_0'(x)) \geq 0$, whenever $D_3^2 f(x, u_0(x), u_0'(x))$ exists. If $f \in C^2$, then the Weierstrass condition (3.24) entails the Legendre condition (3.22); in this case the Weierstrass Theorem 3.5 thus follows from the Legendre Theorem 3.4.

A sufficient condition: If $f \in C^2$, u_0 fulfills the weak E-L equation, and

$$\exists \lambda > 0 : \forall v \in V_{0,0}, \quad \delta^2 F(u,v) \geq \lambda \int_a^b [v(x)^2 + v'(x)^2]\, dx, \tag{3.25}$$

then u_0 is a weak relative minimizer. The converse obviously fails. (7) On the other hand, the weaker condition (3.21) does not even entail that u_0 is a weak relative minimizer; for a counterexample see e.g. [De 104].

Synthesis. (See above for the regularity assumptions)
— For absolute minimizers (for any N):
if F is convex, then any extremal is a minimizer,

(7) This result is at the basis of a sophisticated theory of Legendre and Jacobi (see e.g. [De 115-124]).

for any minimizer u, the Hessian matrix $D_3^2 f(x,u,u')$ is positive semidefinite.
— For relative minimizers (for $N=1$):
for any weak relative minimizer u, $D_3^2 f(x,u,u') \geq 0$,
for any strong relative minimizer u, $E(x,u_0(x),u_0'(x),\eta) \geq 0$,
any extremal u that fulfills (3.25) is a weak relative minimizer.

The converse of these statements fails.

I.4. First integral and Lagrange multipliers

First integrals. For any $f \in C^2$ and any solution $u \in V_{\alpha,\beta} \cap C^2([a,b])$ of the strong E-L equation of the functional (2.1), a simple computation [De 70] shows that

$$\frac{d}{dx}[f(x,u(x),u'(x)) - u'(x)D_3 f(x,u(x),u'(x))] = D_1 f(x,u(x),u'(x)) \qquad \forall x \in [a,b]. \tag{4.1}$$

In analogy with what we saw for the integral formulation of the E-L equation, by integrating this equation we get

$$f(x,u(x),u'(x)) - u'(x)D_3 f(x,u(x),u'(x)) = C + \int_a^x D_1 f(s,u(s),u'(s))\,ds \qquad \forall x \in [a,b], \tag{4.2}$$

and this holds also if $f \in C^1$ and $u \in C^1([a,b])$. More generally, if $u \in C^1_{pw}([a,b])$ then (4.2) holds a.e. in $[a,b]$.

If f does not explicitly depend on x then, defining the *Hamiltonian function,*

$$H(u,\xi) := \xi D_3 f(u,\xi) - f(u,\xi) \qquad \forall (u,\xi) \in \mathbf{R}^2, \tag{4.3}$$

by (4.1) we get

$$\frac{d}{dx} H(u(x),u'(x)) = 0 \qquad \forall x \in [a,b].$$

That is, for any extremal $u = u(x)$ of F,

$$H(u(x),u'(x)) = C : \text{constant} \qquad \forall x \in [a,b]. \tag{4.4}$$

The constant C may be determined by the boundary conditions. (4.4) is also named a *conservation law,* and $H(u,u')$ is also called a *first integral* (of the motion).

Establishing a first integral may be useful for the integration of the E-L equation. For instance the equation $H(u,u') = C$ is of first order, whereas the E-L equation is of second order. If the equation $H(u,u') = C$ can be made explicit with respect to u', to wit if it may be rewritten in the *normal form* $u' = g(u,C)$, then a standard theory may be applied to the study of this equation.

Analogous remarks apply in the case of $D_2 f(x,u,u') \equiv 0$, that is the function $f(...)$ does not depend on u. The E-L equation is then reduced to $D_2 \tilde{f}(x,u') = C$. If this equation can be made explicit with respect to u', then also in this case one is reduced to a first-order equation.

Functionals of a different form from (2.1) may have several first integrals, or no one at all.

Method of Lagrange multipliers. [De 70] Let $f, g \in C^1([a,b]\times\mathbf{R}\times\mathbf{R})$ be given functions. Let us consider a functional F defined as in (2.1), and search for a minimizer subject to the constraint

$$G(u) = \int_a^b g(x, u(x), u'(x))\, dx = C \ (\in \mathbf{R} \text{ prescribed}). \tag{4.5}$$

That is, setting

$$V_{\alpha,\beta,G} := \{v \in C^1([a,b]) : v(a) = \alpha, v(b) = \beta, G(u) = C\}, \tag{4.6}$$

we search for $u \in V_{\alpha,\beta,G}$ such that

$$F(u) \le F(v) \qquad \forall v \in V_{\alpha,\beta,G}. \tag{4.7}$$

Theorem 4.1. *(Lagrange multipliers) Let $f, g \in C^1([a,b]\times\mathbf{R}\times\mathbf{R})$ and $u \in V_{\alpha,\beta,G}$ fulfill (4.7). If the constraint is not degenerate at u, in the sense that*

$$\exists \psi \in C^1([a,b]): \quad \psi(a) = \psi(b) = 0, \quad \delta G(u,\psi) \neq 0, \tag{4.8}$$

then

$$\exists \lambda \in \mathbf{R} : \forall v \in V_{0,0}, \qquad \delta(F + \lambda G)(u, v) = 0. \tag{4.9}$$

For the argument see e.g. [De 70].

By (4.9) the function $\varphi := F + \lambda G$ fulfills the E-L equation in weak form — hence also that in strong form if $f, g \in C^2([a,b]\times\mathbf{R}\times\mathbf{R})$ and $u \in C^2([a,b]) \cap V_{\alpha,\beta,G}$.

The function φ and the scalar λ are named the *Lagrangian function* and the *Lagrange multiplier* (associated to the given problem). Anyway the integrand f is often also called the Lagrangian, extending the terminology of analytical mechanics.

Exercises. (i) Extend (4.3) to vector-valued functions.

(ii) May Theorem 4.1 be inverted?

I.5. Examples

Let us define the functionals

$$\begin{aligned} &F_1(u) = \int_{-1}^1 u(x)^2 [2x - u'(x)]^2\, dx \\ &\forall u \in V_1 := \{v \in C^1([-1,1]) : v(-1) = 0, v(1) = 1\}, \end{aligned} \tag{5.1}$$

$$\begin{aligned} &F_2(u) = \int_{-1}^1 [u'(x)^2 - 1]^2\, dx \\ &\forall u \in V_2 := \{v \in C^1_{pw}([-1,1]) : v(-1) = v(1) = 0\}, \end{aligned} \tag{5.2}$$

$$F_3(u) = F_2(u) \qquad \forall u \in V_3 := \{v \in C^1([-1,1]) : v(-1) = v(1) = 0\}, \tag{5.3}$$

$$F_4(u) = \int_{-1}^1 \exp\left[-u'(x)^2\right] dx \qquad \forall u \in V_2. \tag{5.4}$$

The functional F_1 has just the minimizer

$$u(x) = \begin{cases} 0 & \text{if } -1 \le x \le 0, \\ x^2 & \text{if } 0 \le x \le 1; \end{cases} \tag{5.5}$$

as $u \in C^1([-1,1]) \setminus C^2([-1,1])$, u is a weak (not strong) extremal.

It is easily seen that F_2 has an infinity of minimizers, all $u \in V_2$ with $u' = \pm 1$ a.e. in $]-1,1[$; this class includes many piecewise monotone functions (although not all functions with this property are piecewise monotone!). None of these functions is of class C^1, and actually the problem of minimizing F_3 has no solution. This suggests that the C^ks might not be the most appropriate spaces for the study of several variational problems.

For the functional F_2, the strong form of the E-L equation reads

$$\frac{d}{dx} D_3 f(u'(x)) = 0 \qquad \text{i.e.} \qquad \frac{d}{dx}\{u'(x)[u'(x)^2 - 1]\} = 0, \tag{5.6}$$

and this is solved by $u_0 \equiv 0$ ($\in V_2$), which is no minimizer. The same holds for the weak form of the E-L equation, which reads

$$u'(x)[u'(x)^2 - 1] = \text{ constant in }]0,1[. \tag{5.6'}$$

For the functional F_4, the E-L equation in strong form reads

$$\frac{d}{dx}\{-u'(x)\exp[-u'(x)^2]\} = 0 \qquad \forall x \in]-1,1[, \tag{5.7}$$

and this is solved by $u_0 \equiv 0$ ($\in V_2$). It is easily seen that this is a maximizer. The same holds for the weak form of the E-L equation, which reads

$$\{-u'(x)\exp[-u'(x)^2]\} = \text{ constant in }]-1,1[. \tag{5.7'}$$

Some classical variational problems.

(i) *Problem of the brachystochrone.* [BuGiHi 44], [Br 7, 40], [De 55] Given two distinct points (x_0, y_0), (x_1, y_1) sitting on a vertical plane, let us join them by a curve $y = u(x)$, and assume that a point mass slides frictionless along this curve under the action of gravity. It is required to determine a function u such that, starting form (x_0, y_0), the moving point reaches (x_1, y_1) in the shortest possible time. Let us denote by v the scalar velocity of the point mass. Because of the conservation of the total mechanical energy, we have

$$\frac{1}{2} m v(x)^2 + mgu(x) = c \text{ (constant)} \qquad \text{for } x \in [x_0, x_1], \tag{5.8}$$

whence $v(x) = \sqrt{\frac{2c}{m} - 2gu(x)}$. Denoting by s the arc-length parameter, we have $ds(x) = \sqrt{1 + u'(x)^2}\, dx$; hence

$$T = \int_0^T dt = \int_0^L \frac{ds}{v(x)} = \int_{x_0}^{x_1} \sqrt{\frac{1 + u'(x)^2}{\frac{2c}{m} - 2gu(x)}}\, dx. \tag{5.9}$$

One may show that this functional is minimized by a curve having the following parametric representation:

$$x(t) = x_0 + k(t - \sin t), \qquad u(t) = y_0 - k(1 - \cos t) \qquad 0 \le \tau \le \bar{\tau}, \tag{5.10}$$

with k and $\bar{\tau}$ positive constants that are determined by the conditions $x(\bar{\tau}) = x_1$ and $u(\bar{\tau}) = y_1$. This is an arc of cycloid. As

$$x(t) - x_0 = \frac{kt^3}{6} + o(t^4), \qquad u(t) - y_0 = \frac{kt^2}{2} + o(t^3), \tag{5.10'}$$

it is easily seen that this arc starts vertically.

One may show that this curve is also *isochrone* (also named *tautochrone):* if the point mass is released with null velocity from any point $(x, y) \neq (x_1, y_1)$ of the curve, then the time that it takes to reach (x_1, y_1) does not depend on (x, y). Note that the arc of cycloid between (x, y) and (x_1, y_1) is not the brachystochrone between these two points, which is actually an arc of cycloid but starts vertically.

(ii) *Problem of the catenary.* [BuGiHi 46], [De 74] We want to represent the equilibrium configuration of a string (or a chain) of length L subject to the gravity, having extremes (x_0, y_0) and (x_1, y_1). Any equilibrium configuration minimizes the potential energy

$$F(u) = g\rho \int_{x_0}^{x_1} u\sqrt{1 + u'(x)^2}\, dx \qquad \forall u \in C^1([x_0, x_1]),\ u(x_0) = y_0, u(x_1) = y_1 \tag{5.11}$$

(g: gravity acceleration, ρ: linear density of the string) among the functions u subject to the length constraint:

$$G(u) := \int_{x_0}^{x_1} \sqrt{1 + u'(x)^2}\, dx = L. \tag{5.12}$$

One may then use the method of Lagrange multipliers (see Sect. I.4), defining the Lagrange functional $\mathcal{L}(u, \lambda) := F(u) + \lambda G(u)$. For any $\lambda > 0$ this functional is convex, so that any extremal minimizes F along the constraint. One may show that this functional is minimized by a function (called *catenary)* of the form

$$u(x) = \frac{1}{\alpha}[\cosh(\alpha x + \beta) + \gamma] \qquad \forall x \in [x_0, x_1], \tag{5.13}$$

for suitable real constants α, β, γ. In alternative as a parameter one may use the arc-length:

$$s(x) = \int_{x_0}^{x} \sqrt{1 + u'(x)^2}\, dx \qquad \forall x \in [x_0, x_1], \tag{5.14}$$

and represent the catenary as $x = x(s)$ and $y = u(s)$, for $0 \leq s \leq L$.

(iii) *Problem of the elastic string.* [BuGiHi 49] Let us consider an elastic string, that we represent as the graph of a function $y = u(x)$ ($x \in [-1, 1]$) with $u(-1) = u(1) = 0$. As a first approximation, the total mechanical energy (= elastic energy + potential energy) equals

$$F(u) = \rho \int_{-1}^{1} \Big(\frac{k}{2} u'(x)^2 + gu(x)\Big)\, dx \qquad \forall u \in C^1([-1, 1]), \tag{5.15}$$

(g: gravity acceleration, k: elasticity coefficient, ρ: linear density of the string). Any extremal $u \in C^2([-1, 1])$ with $u(-1) = u(1) = 0$ fulfills the corresponding E-L equation

$$-ku''(x) + g = 0 \qquad \forall x \in\,]-1, 1[, \tag{5.16}$$

which is solved by

$$u(x) = \frac{g}{2k}(x^2 - 1) \qquad \forall x \in [-1, 1]. \tag{5.17}$$

(iv) *Problem of the clamped elastic beam.* [BuGiHi 50] If instead of a string we consider an elastic beam, then the total energy reads

$$\begin{aligned} &F(u) = \rho \int_{-1}^{1} \Big\{\frac{H}{2} u''(x)^2 + gu(x)\Big\}\, dx \\ &\forall u \in C^2([-1, 1]),\ u(-1) = u(1) = 0,\ u'(-1) = u'(1) = 0, \end{aligned} \tag{5.18}$$

with H a positive constant. Any extremal u fulfills the E-L equation

$$H\,D^4u(x)+g=0 \quad \text{with} \quad u(-1)=u(1)=0,\ u'(-1)=u'(1)=0; \tag{5.19}$$

this has the solution

$$u(x)=\frac{g}{24\,H}(x^2-1)^2 \qquad \forall x\in[-1,1]. \tag{5.20}$$

(v) *Harmonic oscillator.* The evolution of an oscillating point mass (a spring) is represented by a function $u=u(t)$ for $t\in[0,T]$, and leads to the formulation of a (Lagrangian) functional of the form

$$\int_0^T f(t,u,u')\,dt=\int_0^T\Big\{\frac{m}{2}(u')^2-\frac{\omega}{2}u^2\Big\}\,dt \qquad \forall u\in C^1([0,T]), \tag{5.21}$$

with ω a positive constant. Any extremal $u\in C^2([0,T])$ fulfills the E-L equation

$$mu''(t)+\omega u(t)=0 \qquad \forall t\in[0,T], \tag{5.22}$$

which is known as the equation of the harmonic oscillator.

(vi) *Wave equation.* The evolution of an elastic string is represented by a function $u=u(x,t)$ for $(x,t)\in Q:=[a,b]\times[0,T]$. The transversal oscillations of the string lead us to define a (Lagrangian) functional of the form

$$F(u):=\iint_Q[\rho(D_tu)^2-\tau(D_xu)^2+gu]\,dxdt \qquad \forall u\in C^1(Q), \tag{5.23}$$

with ρ,τ positive constants. Any extremal $u\in C^2(Q)$ of this functional fulfills the E-L equation

$$\rho D_t^2u(x,t)-\tau D_x^2u(x,t)=g \qquad \forall(x,t)\in Q, \tag{5.24}$$

which is known as the one-dimensional wave equation.

(vii) *Curve of prescribed curvature.* Let a function $H\in C^0(]a,b[\times\mathbf{R})$ be prescribed, and consider the functional

$$F(u):=\int_a^b\big[\sqrt{1+u'(x)^2}+H(x,u(x))\big]\,dx \qquad \forall u\in C^1([a,b]). \tag{5.25}$$

The corresponding E-L equation is

$$\frac{d}{dx}\frac{u'(x)}{\sqrt{1+u'(x)}}=D_2H(x,u(x)) \qquad \forall x\in\,]a,b[. \tag{5.26}$$

If $D_2H=C$ constant, then the graphic of u has constant curvature. If $2C\le(b-a)$, then by integrating (5.26) one may check that the solution consists of arcs of circle of radius $(b-a)/2C$. If instead $2C>(b-a)$ then the solution fails to exist.

(viii) *Newton dynamics.* Let us consider a system of n point masses $m_1,...,m_n$, respectively sitting at the points $X_1(t),...,X_n(t)\in\mathbf{R}^3$ at the generic instant $t\in[0,T]$, and subject to a conservative field of force:

$$f(X_j(t))=-\nabla V(X_j(t)) \qquad \forall t\in[0,T], j=1,...,n, \tag{5.27}$$

for a prescribed potential $V \in C^1(\mathbf{R}^3)$. Let us define the velocities $Y_j(t) := X_j'(t)$, and introduce the Lagrangian function (= kinetic energy minus potential energy)

$$\begin{aligned} f(X,Y) &:= \frac{1}{2}\sum_{j=1}^n m_j|Y_j|^2 - \sum_{j=1}^n V(X_j) \\ \forall X = (X_1,...,X_n), Y &= (Y_1,...,Y_n) \in (\mathbf{R}^3)^n. \end{aligned} \tag{5.28}$$

For any process $X = X(t)$, let us then define the *action* functional

$$\begin{aligned} F(X) = \int_0^T f(X(t),X'(t))\,dt = \sum_{j=1}^n \int_0^T \Big(\frac{1}{2}m_j|X_j'(t)|^2 - V(X_j(t))\Big)\,dt \\ \forall X \in [C^1([0,T])^3]^n. \end{aligned} \tag{5.29}$$

The classical *Hamilton principle* prescribes that the evolution occurs along an extremal of the action. The corresponding E-L equation coincides with Newton's motion law:

$$mX_j''(t) = -\nabla V(X_j(t))\ (= f(X_j(t))) \qquad \forall t \in [0,T],\ \text{for } j = 1,...,n. \tag{5.30}$$

Exercise. How must the latter example be modified, if the potential energy $\sum_{j=1}^n V(X_j)$ is replaced by a more general one of the form $V(X_1,...,X_n)$?

I.6. Boundary conditions

Let us still define the functional F as in (2.1), and fix any $u \in C^2([a,b])$. Here we shall not prescribe any boundary conditions, and let vary not only the function u but also the extremes $x_0 = a$, $x_1 = b$, $y_0 = \alpha$, $y_1 = \beta$. We shall still denote the first variation of the function u by δu, and the differential of the numbers x_0, x_1, y_0, y_1 by dx_0, dx_1, dy_0, dy_1. [8] We shall confine the interval $[x_0,x_1]$ to a fixed interval $[\bar a,\bar b]$, and assume that $y_i = u(x_i)$ for $i = 0,1$, so that

$$dy_i(x_i) = (\delta u)(x_i) + u'(x_i)\,dx_i + o(dx_i) \qquad (i = 0,1). \tag{6.1}$$

Let us write $dy(x_i) := dy_i(x_i)$ for $i = 0,1$, and define the *conjugate momentum* of u':

$$p(x) := D_3 f(x,u(x),u'(x)) \qquad \forall x \in [\bar a,\bar b]. \tag{6.2}$$

We shall often omit to display the argument $(x,u(x),u'(x))$, as we did above.

By varying the curve of equation $u = u(x)$ and also its extremes, one can prove the following general variation formula:

$$\begin{aligned} \delta F(u,\delta u) &= \int_{x_0}^{x_1} (D_2 f\,\delta u + p\,\delta u')\,dx + f\,dx\Big|_{x_0}^{x_1} \\ &= \int_{x_0}^{x_1} (D_2 f - p')\,\delta u\,dx + p\,\delta u\Big|_{x_0}^{x_1} + f\,dx\Big|_{x_0}^{x_1} \\ &\overset{(6.1)}{=} \int_{x_0}^{x_1} (D_2 f - p')\,\delta u\,dx + p\,dy\Big|_{x_0}^{x_1} + (f - p\,u')\,dx\Big|_{x_0}^{x_1}, \end{aligned} \tag{6.3}$$

[8] The following *thumb rule* here applies: variations are denoted by δ for functions, by d for numbers.

up to infinitesimals of higher order than one in δu, dx_i and dy_i, for $i = 0, 1$. (On the left side of this formula we omitted to display the dependence of the variation of the extremes.)

Extending the previous definition, we say that u is an extremal for the functional F whenever

$$\delta F(u, \delta u) = 0 \qquad \text{for any admissible variation } \delta u, \tag{6.3$'$}$$

including variations of the boundary values x_0, x_1, y_0 and y_1. We may also vary u just at the internal of the interval $[x_0, x_1]$ keeping the boundary values fixed, namely with $dy_i = dx_i = 0$ for $i = 0, 1$. The formula (6.3) then yields the E-L equation, which obviously also reads

$$p'(x) = D_2 f(x, u(x), u'(x)) \qquad \forall x \in [\bar{a}, \bar{b}]. \tag{6.4}$$

Writing $f_i := f(x_i, u(x_i), u'(x_i))$ for $i = 0, 1$, (6.3') then yields the *transversality condition*

$$\begin{aligned} \delta F(u, \delta u) &= p\, dy\Big|_{x_0}^{x_1} + (f - p\, u')\, dx\Big|_{x_0}^{x_1} \\ &= p(x_1)\, dy_1 - p(x_0)\, dy_0 + \\ &\quad [f_1 - p(x_1)\, u'(x_1)]\, dx_1 - [f_0 - p(x_0)\, u'(x_0)]\, dx_0 = 0, \end{aligned} \tag{6.5}$$

up to infinitesimals of higher order than one in dx_0, dx_1 and δu. We have thus derived the following statement.

Proposition 6.1. *A function u is extremal for F if and only if it fulfills the E-L equation and the transversality condition*

$$p(x_1)\, dy_1 - p(x_0)\, dy_0 + [f_1 - p(x_1)\, u'(x_1)]\, dx_1 - [f_0 - p(x_0)\, u'(x_0)]\, dx_0 = 0. \tag{6.6}$$

For each of the four variables x_0, x_1, y_0 and y_1, we must decide whether we prescribe its value or let it free to vary; in any case we get four boundary conditions. Let us see some examples:

(i) By prescribing $x_0 = a$, $x_1 = b$, $y_0 = \alpha$ and $y_1 = \beta$, we obviously have

$$dx_0 = 0, \quad dx_1 = 0, \quad dy_0 = 0, \quad dy_1 = 0,$$

so that (6.6) provides no further boundary condition.

(ii) By prescribing no one of the values x_0, x_1, y_0 and y_1, by (6.6) we get four boundary conditions:

$$p(x_1) = 0, \quad p(x_0) = 0, \quad p(x_1)\, u'(x_1) = f_1, \quad p(x_0)\, u'(x_0) = f_0,$$

that is,

$$p(x_1) = 0, \quad p(x_0) = 0, \quad f_1 = 0, \quad f_0 = 0. \tag{6.7}$$

(iii) By prescribing $x_0 = a$ and $x_1 = b$ and allowing y_0 and y_1 to vary, we have $dx_0 = 0$ and $dx_1 = 0$. Moreover the condition (6.6) entails $p(x_0) = 0$ and $p(x_1) = 0$. In this case the boundary conditions read

$$x_0 = a, \qquad x_1 = b, \qquad p(x_0) = 0, \qquad p(x_1) = 0. \tag{6.8}$$

The first two are named *forced conditions,* the other two *natural conditions.*

(iv) By prescribing $x_0 = a$ and $y_1 = \beta$ and letting x_1 and y_0 free, we have $dx_0 = 0$ and $dy_1 = 0$. Moreover the condition (6.6) entails $p(x_1)\,u'(x_1) = f_1$ and $p(x_0) = 0$. In this case the boundary conditions read

$$x_0 = a, \qquad y_1 = \beta, \qquad p(x_1)\,u'(x_1) = f_1, \qquad p(x_0) = 0. \tag{6.8$'$}$$

The general solution of the E-L equation contains two arbitrary constants, as it is natural since this is a second-order differential equation. Moreover in general the extremes x_0 and x_1 are not determined. Four boundary conditions match these four degrees of freedom.

These developments are easily extended to vector-valued functions.

Extremes on prescribed curves. A variant of the previous setting consists in fixing two disjoint curves Γ_0, Γ_1 in the plane (x, y), and then prescribing that the points of departure and arrival of the extremal arc respectively lie on these curves. Let Γ_i be the graphic of a function $\psi_i \in C^1([\bar a, \bar b])$, for $i = 0, 1$. Notice that

$$dy_i = \psi_i'(x_i)\,dx_i + o(dx_i) \qquad (i = 0, 1). \tag{6.9}$$

Up to infinitesimals of higher order than one, because of (6.3) any extremal u of F fulfills the E-L equation and (with synthetical notation)

$$\begin{aligned} \delta F(u, \delta u) &\overset{(6.5)}{=} p\,dy\Big|_{x_0}^{x_1} + (f - p\,u')\,dx\Big|_{x_0}^{x_1} \overset{(6.9)}{=} [p\,\psi_i' + f - p\,u']\,dx\Big|_{x_0}^{x_1} \\ &= \{f_1 + \big[\psi_1'(x_1) - u'(x_1)\big]\,p(x_1)\}\,dx_1 \\ &\quad - \{f_0 + \big[\psi_0'(x_0) - u'(x_0)\big]\,p(x_0)\}\,dx_0, \end{aligned} \tag{6.10}$$

up to infinitesimals of higher order than one in dx_0, dx_1 and δu. Here we just get two boundary conditions. As dx_1 is independent of dx_0, these read

$$\begin{aligned} f_0 + \big[\psi_0'(x_0) - u'(x_0)\big]\,p(x_0) = 0, \\ f_1 + \big[\psi_1'(x_1) - u'(x_1)\big]\,p(x_1) = 0. \end{aligned} \tag{6.11}$$

These are named *transversality conditions* (with reference to the constraints Γ_0, Γ_1).

Exercises. (i) Extend the previous developments to vector-valued functions.

(ii) Write the general variation formula for the functional

$$\begin{aligned} \widehat F(u) = \int_a^b f(x, u(x), u'(x))\,dx + G(x_0, x_1, y_0, y_1) \\ \forall u \in C^2([a, b]), \forall (x_0, x_1, y_0, y_1) \in \mathbf{R}^4, \end{aligned} \tag{6.12}$$

with $f \in C^2([a, b]\times\mathbf{R}\times\mathbf{R}\times\mathbf{R})$ and $G \in C^2(\mathbf{R}^4)$ given functions.

(iii) Write the general variation formula for the functional

$$\widehat F(u) = \int_a^b f(x, u(x), u'(x), u''(x))\,dx \qquad \forall u \in C^4([a, b]), \tag{6.13}$$

with $f \in C^4([a, b]\times\mathbf{R}\times\mathbf{R}\times\mathbf{R})$ given function.

I.7. Noether Theorem

Essentially this theorem tells us that

symmetries in the universe are responsible for conservation laws.

(Of course this presumes that the laws of the universe admit a variational formulation, and this assumption is far from being obvious.) We prove this statement within the Euler-Lagrange formulation; however the same result might be proved in the framework of the Hamiltonian (i.e., canonical) theory.

Let us consider a family of transformations of the form

$$\begin{aligned} (\widetilde{x} :=)\ \widetilde{x}_\varepsilon &= \varphi(x,u;\varepsilon) \\ (\widetilde{u} :=)\ \widetilde{u}_\varepsilon &= \psi(x,u;\varepsilon) \end{aligned} \qquad \forall (x,u,\varepsilon) \in [a,b]\times\mathbf{R}\times]-1,1[, \tag{7.1}$$

with φ, ψ mappings of class C^1 with respect to the parameter ε. For any ε, any curve $u = u(x)$ is thus transformed into a curve $\widetilde{u} = \widetilde{u}(\widetilde{x},\varepsilon)$. Let us assume that

$$\begin{aligned} \varphi(x,u;0) &= x \\ \psi(x,u;0) &= u \end{aligned} \qquad \forall (x,u) \in [a,b]\times\mathbf{R}, \tag{7.2}$$

and denote by the index ε the partial derivative with respect to this parameter; we thus get

$$\begin{aligned} \widetilde{x} &= x + \varepsilon\varphi_\varepsilon(x,u;0) + o(\varepsilon) \\ \widetilde{u} &= u + \varepsilon\psi_\varepsilon(x,u;0) + o(\varepsilon) \end{aligned} \qquad \forall (x,u) \in [a,b]\times\mathbf{R}. \tag{7.3}$$

The pair of functions $(x,u) \mapsto (\varphi_\varepsilon(x,u;0), \psi_\varepsilon(x,u;0))$ is named the *infinitesimal generator* of the family of transformations (7.1). Let us set [9]

$$\begin{aligned} \delta x &:= \widetilde{x} - x = \varepsilon\varphi_\varepsilon(x,u;0) + o(\varepsilon) \\ \delta u &:= \widetilde{u} - u = \varepsilon\psi_\varepsilon(x,u;0) + o(\varepsilon). \end{aligned} \tag{7.4}$$

As the variable x is unidimensional, varying it is tantamount to varying the extremes of the domain of the function $u = u(x)$. The general variation formula (6.3) is here fulfilled with du in place of dy at the extremes of the interval, up to infinitesimals of higher order than one in dx and du. We shall say that the functional (2.1) is invariant with respect to the above family of transformations if, setting $\widetilde{x}_i(\varepsilon) := \varphi(x_i, u(x_i);\varepsilon)$ for $i = 0, 1$,

$$F(u) := \int_{x_0}^{x_1} f(x,u(x),u'(x))\,dx = \int_{\widetilde{x}_0(\varepsilon)}^{\widetilde{x}_1(\varepsilon)} f(\widetilde{x},\widetilde{u}(\widetilde{x},\varepsilon),\widetilde{u}'(\widetilde{x},\varepsilon))\,d\widetilde{x} =: \widetilde{F}(\widetilde{u}(\cdot,\varepsilon)) \qquad \forall\varepsilon. \tag{7.5}$$

In this case, defining the momentum p as in (6.2), for any extremal $u = u(x)$ of F by (6.3) we have

$$\begin{aligned} \delta F(u,\delta u) &= p\,du\Big|_{x_0}^{x_1} + (f - p\,u')\,dx\Big|_{x_0}^{x_1} \\ &\overset{(7.4)}{=} \varepsilon p\,\psi_\varepsilon(x,u(x);0) + \varepsilon(f - p\,u')\,\varphi_\varepsilon(x,u(x);0)\Big|_{x_0}^{x_1} \qquad \forall\varepsilon, \end{aligned} \tag{7.6}$$

up to infinitesimals of higher order than one in dx_0, dx_1 and δu. The invariance condition $\delta F(u,\delta u) = 0$ then reads

$$p\,\psi_\varepsilon(x,u(x);0) + (f - p\,u')\,\varphi_\varepsilon(x,u(x);0) = 0 \qquad \text{for } x = x_0 \text{ and } x = x_1. \tag{7.6'}$$

[9] We still use the letter "d" ("δ" , resp.) to denote variations of numbers (functions, resp.).

Being x_0 and x_1 arbitrary points of $[a,b]$, we then infer the following classical result.

Theorem 7.1. *(Noether) If the functional (2.1) is invariant for the family of transformations (7.1) and these fulfill (7.2), then any extremal $u = u(x)$ of F fulfills the two following boundary conditions*

$$p\,\psi_\varepsilon(x,u(x);0) + (f - p\,u')\,\varphi_\varepsilon(x,u(x);0) = \textit{constant in } [a,b]. \tag{7.7}$$

($p\,u' - f$ coincides with the Hamiltonian.)

These developments may easily be extended to vector-valued functions. In this case

$$\begin{aligned} \widetilde{x} &= x + \varepsilon\varphi_\varepsilon(x,\vec{u};0) + o(\varepsilon) \\ \tilde{\vec{u}} &= \vec{u} + \varepsilon\vec{\psi}_\varepsilon(x,\vec{u};0) + \vec{o}(\varepsilon) \end{aligned} \tag{7.7'}$$

and along any extremal (summing over repeated indices)

$$p_j\,\psi_{j\varepsilon}(x,\vec{u};0) + (f - p_j\,u_j')\,\varphi_\varepsilon(x,\vec{u};0) = \text{ constant in } [a,b]. \tag{7.8}$$

Examples. (i) *Conservation of the total energy.* For the family of transformations

$$\begin{aligned} \widetilde{x} &= \varphi(x,u;\varepsilon) = x + \varepsilon \\ \widetilde{u} &= \psi(x,u;\varepsilon) = u \end{aligned} \qquad \forall (x,u) \in [a,b]\times\mathbf{R}, \tag{7.9}$$

we have

$$\varphi_\varepsilon(x,u;0) = 1, \qquad \psi_\varepsilon(x,u;0) = 0. \tag{7.10}$$

Therefore, for any extremal of F, (7.7) yields

$$f(x,u,u') - p\,u' = \text{ constant in } [a,b]. \tag{7.11}$$

In this way we retrieve a known property:

if the Lagrangian (equivalently, the Hamiltonian) does not explicitly depend on time, then the Hamiltonian is constant along extremal curves.

This has a straightforward extension to vector-valued functions.

(ii) *Conservation of the (linear) momentum.* Let us consider the family of transformations

$$\begin{aligned} \widetilde{x} &= \varphi(x,u;\varepsilon) = x \\ \widetilde{u}_1 &= \psi_1(x,\vec{u};\varepsilon) = u_1 + \varepsilon \\ \widetilde{u}_j &= \psi_j(x,\vec{u};\varepsilon) = u_j \quad (j = 2,...,N) \end{aligned} \tag{7.12}$$

(this for $N \geq 2$; if instead $N = 1$ it suffices to drop the final line). We have

$$\begin{aligned} &\varphi_\varepsilon(x,u;0) = 0, \\ &\psi_{1\varepsilon}(x,u;0) = 1 \\ &\psi_{j\varepsilon}(x,u;0) = 0 \quad (j = 2,...,N), \end{aligned} \tag{7.13}$$

so that for any extremal of F (7.8) yields

$$p_1 = \text{constant in } [a,b]. \tag{7.14}$$

We conclude that:

if the Lagrangian (equivalently, the Hamiltonian) does not explicitly depend on a component of $\vec{u}$ (here u_1), then the corresponding component of the conjugate momentum is constant along extremal curves.

(iii) *Conservation of the angular momentum.* Let us consider the family of transformations

$$\begin{aligned} \widetilde{x} &= \varphi(x,u;\varepsilon) = x \\ \widetilde{u}_1 &= \psi_1(x,\vec{u};\varepsilon) = u_1\,\cos\varepsilon + u_2\,\sin\varepsilon \\ \widetilde{u}_2 &= \psi_2(x,\vec{u};\varepsilon) = -u_1\,\sin\varepsilon + u_2\,\cos\varepsilon \\ \widetilde{u}_j &= \psi_j(x,\vec{u};\varepsilon) = u_j \qquad (j=3,...,N) \end{aligned} \tag{7.15}$$

(this if $N \geq 3$; if instead $N = 2$ it suffices to drop the final line). As

$$\begin{aligned} &\varphi_\varepsilon(x,u;0) = 0, \\ &\psi_{1\varepsilon}(x,u;0) = u_2, \\ &\psi_{2\varepsilon}(x,u;0) = -u_1 \\ &\psi_{j\varepsilon}(x,u;0) = 0 \qquad (j=3,...,N), \end{aligned} \tag{7.16}$$

for any extremal of F (7.8) yields

$$p_1\,u_2 - p_2\,u_1 = \text{ constant in } [a,b]. \tag{7.17}$$

For instance, if $N = 2$ we conclude that:

if the Lagrangian (equivalently, the Hamiltonian) explicitly depends on $\vec{u}$ just via its modulus (this is the case e.g. in a central field of force), then the angular momentum is constant along extremal curves.

I.8. Legendre transformation and canonical equations

Moments. Assuming that $f \in C^2([a,b]\times\mathbf{R}\times\mathbf{R})$ and $u \in C^2([a,b])$, we already defined the momentum

$$p(x) := D_3 f(x,u(x),u'(x)) \qquad \forall x \in [a,b]. \tag{8.1}$$

If $D_3^2 f(x,u(x),u'(x)) \neq 0$ for any x, then this algebraic equation may be solved with respect to $u'(x)$: there exists a function $g \in C^1([a,b]\times\mathbf{R}\times\mathbf{R})$ such that

$$p(x) = D_3 f(x,u(x),u'(x)) \quad \Leftrightarrow \quad u'(x) = g(x,u(x),p(x)) \qquad \forall x \in [a,b]. \tag{8.2}$$

Hence

$$f(x,u(x),u'(x)) = f(x,u(x),g(x,u(x),p(x))) =: \widetilde{f}(x,u(x),p(x)) \qquad \forall x \in [a,b]. \tag{8.3}$$

The definition (8.1) allows us to reformulate the strong E-L equation as

$$D_2 f(x,u(x),u'(x)) = p'(x) \qquad \forall x \in [a,b] \tag{8.4}$$

(but not $D_2\widetilde{f}(x,u(x),p(x)) = p'(x)$: why?). We then get the system

$$\begin{aligned} p(x) &= D_3 f(x,u(x),u'(x)) \\ p'(x) &= D_2 f(x,u(x),u'(x)) \end{aligned} \qquad \forall x \in [a,b], \tag{8.5}$$

thus a system of two equations of first order, instead of one of second order.

The first of these equations is just the definition of p, the second one is a reformulation of the E-L equation (the equation of motion in a dynamical problem, with x representing the time variable). It is more convenient to proceed in a different way, by applying to $f(x,u,u')$ the Legendre transform with respect to u'.

Legendre transform. Here we replace u' by the variable z, as its connection with u is immaterial; actually in this transformation u will not play any role. Let us set

$$p := D_3 f(x,u,z) \qquad \forall (x,u,z) \in [a,b]\times\mathbf{R}\times\mathbf{R}. \tag{8.6}$$

If $D_3^2 f(x,u,z) \neq 0$ for any (x,u,z), then this equation may be solved with respect to z, entailing

$$p = D_3 f(x,u,z) \quad \Leftrightarrow \quad z = g(x,u,p). \tag{8.7}$$

More specifically we shall assume that the Legendre transform

$$(x,u,z) \mapsto (x,u,D_3 f(x,u,z)) =: (x,u,p) \tag{8.7'}$$

is a diffeomorphism. Let us define the Hamiltonian function

$$H(x,u,p) := p\, g(x,u,p) - f(x,u,g(x,u,p)) \qquad \forall (x,u,p). \tag{8.8}$$

Assuming the relation $z = g(x,u,p)$ (i.e., $p = D_3 f(x,u,z)$), we may also write

$$H(x,u,p) = p\, z - f(x,u,z) \tag{8.8'}$$

with no dependence of H on z (!); indeed

$$\frac{\partial}{\partial z}[p\, z - f(x,u,z)] = p - D_3 f(x,u,z) = 0 \qquad \forall (x,u,p,z). \tag{8.9}$$

Dually, assuming that the Hamiltonian function $H(x,u,p)$ is given, we may set $z := D_3 H(x,u,p)$. Assuming that $D_3^2 H(x,u,p) \neq 0$ for any (x,u,p), we may represent p as a function of z:

$$z = D_3 H(x,u,p) \quad \Leftrightarrow \quad p = k(x,u,z). \tag{8.10}$$

We shall assume that the Legendre transform

$$(x,u,p) \mapsto (x,u,D_3 H(x,u,p)) =: (x,u,z)$$

is a diffeomorphism. Let us define the Lagrangian function

$$f(x,u,z) := k(x,u,z)\, z - H(x,u,k(x,u,z)) \qquad \forall (x,u,z). \tag{8.11}$$

Implying the relation $p = k(x,u,z)$ (i.e., $z = D_3 H(x,u,p)$), we may also write

$$f(x,u,z) = p\, z - H(x,u,p)$$

with no dependence of f on p (!); indeed

$$\frac{\partial}{\partial p}[p\, z - H(x,u,p)] = z - D_3 H(x,u,p) = 0 \qquad \forall (x,u,p,z). \tag{8.12}$$

In conclusion, the three following equations are mutually equivalent

$$\begin{aligned} &p = D_3 f(x,u,z) \\ &z = D_3 H(x,u,p) \qquad\qquad \forall (x,u,p,z). \\ &f(x,u,z) + H(x,u,p) = p\,z \end{aligned} \tag{8.13}$$

This final relation is the Legendre transform between f and H. Moreover from (8.8) it follows that, implying the relations (8.13),

$$\begin{aligned} &D_1 f(x,u,z) = -D_1 H(x,u,p) \\ &D_2 f(x,u,z) = -D_2 H(x,u,p) \end{aligned} \qquad \forall (x,u,p,z). \tag{8.14}$$

Notice that these are just definitions, and do not represent any property of the system: in the mechanical case, they convey no physical meaning.

Canonical equations (Hamilton's theory). Let us replace z by u' and imply the relation

$$p(x) = D_3 f(x,u(x),u'(x)) \quad \text{ that is } \quad u'(x) = g(x,u(x),p(x)) \qquad \forall x \in [a,b],$$

using the notation of (8.7). (8.10) entails that, denoting by (H_x, H_u, H_p) the gradient of H,

$$H_p(x,u(x),p(x)) = u'(x) \qquad \forall x \in [a,b]. \tag{8.15}$$

On the other hand (8.7) yields

$$H_u(x,u(x),p(x)) = -D_2 f(x,u(x),u'(x)) \qquad \forall x \in [a,b]. \tag{8.16}$$

We may thus rewrite the E-L equation (8.4) as

$$H_u(x,u(x),p(x)) = -p'(x) \qquad \forall x \in [a,b]. \tag{8.17}$$

Notice that, at variance with (8.15), this is no identity: it is rather an equation that must be solved by any extremal u. Moreover, using the transform $u' \mapsto p$, the extremal curve is not determined by the function $x \mapsto u(x)$ (i.e., $x \mapsto (u(x), u'(x))$), but by the pair $x \mapsto (u(x), p(x))$.

We have thus derived the Hamilton system of *canonical equations:*

$$\begin{aligned} &u'(x) = H_p(x,u(x),p(x)) \\ &p'(x) = -H_u(x,u(x),p(x)) \end{aligned} \qquad \forall x \in [a,b]. \tag{8.18}$$

These developments may easily be extended to vector-valued functions u and p. Notice that, as the E-L equation characterizes extremals (and thus is just a necessary condition for extrema) the same holds for the canonical equations.

Lagrangian as a function of the canonical variables. As (still assuming the relation $p = D_3 f(x,u,z)$)

$$f(x,u,z) = p\,z - H(x,u,p), \tag{8.19}$$

the Lagrangian (2.1) may be expressed in terms of the canonical variables (u,p), defining the following functional

$$\begin{aligned} J(u,p) :=& \int_a^b f(x,u,u')\,dx = \int_a^b \big[p(x)\,u'(x) - H(x,u(x),p(x))\big]\,dx \\ =& \int_a^b \Big[-p'(x)\,u(x) - H(x,u(x),p(x))\big]\,dx + p\,u\Big|_{x=a}^{x=b} \\ & \forall (u,p) \in C^1([a,b]). \end{aligned} \tag{8.20}$$

This is known as the *Poincaré-Cartan integral* (or the *canonical integral).* The corresponding E-L vector-equation is thus equivalent to the canonical system (8.18).

Canonical formulation of Newtonian mechanics. Let us revisit the example that we outlined at the end of Section I.5. Let us define the Hamiltonian

$$H(X,P) := Y\cdot P - f(X,Y) \qquad \forall X,P\in(\mathbf{R}^3)^n; \tag{8.21}$$

here also H does not depend on Y because of the definition of the vector of the momenta: $P := D_Y f(X,Y)$, that is in terms of components (cf. (5.28))

$$P_j = \frac{1}{m_j}Y_j \qquad j=1,...,n. \tag{8.22}$$

The two latter formulae entail that

$$H(X,P) = \sum_{j=1}^n \frac{1}{2m_j}|P_j|^2 + \sum_{j=1}^n V(X_j) \qquad \forall X,P\in(\mathbf{R}^3)^n. \tag{8.23}$$

The Hamiltonian thus coincides withe the sum of kinetic and potential energy, namely the total (mechanical) energy.

*** Poisson Parentheses.** For any pair of functions $f,g : \mathbf{R}^{2n+1}\to\mathbf{R}$ of class C^1, the Poisson parentheses are defined as follows:

$$\{f,g\} := \sum_{i=1}^n D_{q_i}f\ D_{p_i}g - D_{p_i}f\ D_{q_i}g \ \ \big(=: D_qf\cdot D_pg - D_pf\cdot D_qg\big). \tag{8.24}$$

Obviously the outcome $\{f,g\}$ is also a function $\mathbf{R}^{2n+1}\to\mathbf{R}$, and $\{g,f\} = -\{f,g\}$.

The canonical equations may be written in terms of the Poisson parentheses in a more symmetrical form:

$$\dot q = \{q,H\}, \qquad \dot p = \{p,H\}. \tag{8.25}$$

Moreover the following identities hold:

$$\{q_i,q_j\}=0, \qquad \{p_i,p_j\}=0, \qquad \{q_i,p_j\}=\delta_{ij} \qquad \forall i,j=1,...,n. \tag{8.26}$$

It is easily checked that a (time-independent) function $\xi : \mathbf{R}^{2n+1}\to\mathbf{R}$ is a first integral of the motion if and only if

$$\{\xi,H\}\equiv 0 \qquad \text{in } \mathbf{R}^{2n+1}. \tag{8.27}$$

Exercises. (i) Extend the previous developments to vector-valued functions. Extending the Legendre transform, notice that the condition of no-vanishing of the second derivative is here replaced by the no-vanishing of the Hessian determinant.

(ii) Extend the previous developments to second-order Lagrange functions $f(x,u,u',u'')$. In particular reformulate the corresponding (fourth order) E-L equation as a system of two equations of second order, defining the momentum p as above and also setting $\tilde p := D_{u''}f(x,u,u,u'')$.

(iii) We saw that if the Lagrangian does not explicitly depend on a component of $\vec u$, then the corresponding component of the momentum is constant along the extremal curves. Derive this property from the canonical equations.

[Notice that if the Lagrangian $f(x,\vec u,\vec u')$ does not depend on u_j, then the same holds for the associated Hamiltonian $H(x,\vec u,\vec p)$.]

II. ANALYTICAL MECHANICS AND OPTICS

II.1. Analytical mechanics

In this section we shift to a notation that is typically used in mechanics.
We replace x by t, u by q (generalized coordinate), f by L (Lagrangian function), and

$$F(u) := \int_{x_0}^{x_1} f(...)\, dx \qquad \text{by} \qquad A(q) := \int_{t_0}^{t_1} L(...)\, dt \quad \textit{(action} \text{ functional)}.$$

We denote derivatives with respect to t by the dot: we thus write $\dot{q}$ instead of q'. We also assume that the state is characterized by n generalized coordinates: $q(t) \in \mathbf{R}^n$ for any t.

Aside the *space of configurations,* namely of the $q \in \mathbf{R}^n$, we introduce the *phase space,* which consists of the coordinate-momentum pairs $(q, p) \in (\mathbf{R}^n)^2$. By appending the time coordinate t, we get the Cartan *space of states,* which consists of the triplets $(q, p, t) \in \mathbf{R}^{2n+1}$. The evolution of a mechanical system may thus be represented by the motion of a point in the phase space, or by a curve in the space of the states; as the canonical equations are of first order, this evolution is determined by prescribing any point of that curve. The set of all mechanically admissible evolutions may thus be represented as the flow of a fictitious fluid in the phase space. (Hamiltonian mechanics is thus represented via those coordinates that in fluid mechanics are labelled as Eulerian.)

If $L_t = 0$, or equivalently $H_t = 0$, then the canonical equations are *autonomous,* so that in the phase space the flow is stationary.

One may choose between a Lagrangian representation, in terms of the pair $(q, \dot{q})$ and of the E-L equation, and a Hamiltonian representation in terms of the pair (q, p) and of the canonical equations. The latter approach looks more convenient, also because of the next results.

Incompressibility theorem (of Liouville). Let us denote the divergence in the phase space by $\nabla_{q,p} \cdot$. By the canonical equation, along the motion [10]

$$\begin{aligned} \nabla_{q,p} \cdot (\dot{q}, \dot{p}) := \nabla_q \cdot \dot{q} + \nabla_p \cdot \dot{p} &= \frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} \\ &= \frac{\partial}{\partial q_i} H_{p_i}(t, q(t), p(t)) - \frac{\partial}{\partial p_i} H_{q_i}(t, q(t), p(t)) = 0 \qquad \forall t \in \mathbf{R}. \end{aligned} \tag{1.1}$$

Because of Gauss's theorem of the divergence, the total flux of the field $(\dot{q}, \dot{p})$ through any closed hypersurface in the phase space is then null. Therefore the flux in this space may be regarded as incompressible, as it preserves the volume.

Theorem of energy conservation. By the canonical equations, along the motion

$$\frac{dH}{dt} = H_{q_i} \dot{q}_i + H_{p_i} \dot{p}_i + H_t = H_{q_i} H_{p_i} - H_{p_i} H_{q_i} + H_t = H_t \qquad \forall t \in \mathbf{R}. \tag{1.2}$$

If $H_t = 0$ it then follows that H is constant along the motion:

$$H_t = 0 \quad \Rightarrow \quad H(q(t), p(t)) = \text{ constant } (=: E) \qquad \forall t \in \mathbf{R}. \tag{1.3}$$

[10] These passages are not rigorous (why?); a more precise justification would actually be needed.

In geometrical terms this means that, if $H_t = 0$, then the flux in the phase space is confined to a hypersurface $H(q,p) = E$: constant. In physical terms this accounts for the principle of conservation of the mechanical energy.

Circuitation theorem (of Helmholtz). The action may also be written in terms of the canonical variables (q,p), instead of $(q,\dot q)$. Denoting by A the action in a fixed time interval $[t_0,t_1]$, we thus have $A(t) = A(q(t),p(t))$. Assuming the E-L equation (or the equivalent canonical equations), keeping p fixed and varying q at the extremes $t = t_0$ and $t = t_1$, the general variation formula (I.6.3) yields the variation of the action functional:

$$\delta A(q,p;dq) = p\,dq\Big|_{t_0}^{t_1} \qquad \forall t_0,t_1 \in \mathbf{R}\ (t_0 < t_1). \tag{1.4}$$

Let us next fix t_0 and vary t_1, that we shall denote by t. For any t, let us consider a closed regular curve Γ in the phase space $[0,L] \to \mathbf{R}^{2n} : s \mapsto (q(s,t),p(s,t))$. Denoting by D_s the partial derivative with respect to s, along that curve we may replace dq by $D_s q(s,t)\,ds$; we then define the circuitation

$$I(t) := \int_\Gamma p(\cdot,t)\,dq(\cdot,t) = \int_0^L p(s,t)\,D_s q(s,t)\,ds \qquad \forall t \in \mathbf{R}. \tag{1.5}$$

As the curve Γ is closed, the total variation along Γ (i.e., the integral of the infinitesimal increments) of the action vanishes:

$$\int_0^L \delta A(q,p;dq(s))\ = 0. \tag{1.6}$$

Therefore

$$\begin{aligned} I(t_1) - I(t_0) &\overset{(1.5)}{=} \int_0^L p(s,t)\,dq(s,t)\Big|_{t=t_0}^{t=t_1} \\ &\overset{(1.4)}{=} \int_0^L \delta A(q,p;dq(s)) \overset{(1.6)}{=} 0 \qquad \forall t_0,t_1 \in \mathbf{R}. \end{aligned} \tag{1.7}$$

We have thus proved the following statement.

Theorem 1.1. *The circuitation $I(t)$ is time invariant.*

This is reminiscent of a classical theorem of Helmholtz for ideal (hence irrotational) fluids, [(11)] which asserts that the circuitation of the velocity is constant in time; this also accounts for the persistence of vortices in time.

Elimination of the time variable. Let us set $q_{n+1} := t$, and define the corresponding momentum $p_{n+1} := -H$, so that $p_{n+1}\,\dot q_{n+1} = -H$. The definition of the Lagrangian may thus be modified as follows (here we use the notation $\vec v$ for elements of $\mathbf{R}^{n+1}$)

$$L(t,q,p,\dot q,\dot p) := \sum_{i=1}^n p_i\,\dot q_i - H(t,q,p,\dot q,\dot p) \quad \Leftrightarrow \quad \widetilde L(\vec q,\vec p,\dot{\vec q},\dot{\vec p}) := \sum_{i=1}^{n+1} p_i\,\dot q_i. \tag{1.8}$$

[(11)] This is a model of physical fluids in the Euclidean space, rather than the fictitious fluid in the phase space!

The former of these representations is associated with the minimization of the action functional

$$A(q,p) := \int_{t_0}^{t_1} L(q,p,\dot q,\dot p)\,dt \qquad \text{without constraints.} \tag{1.9}$$

Equivalently one may minimize

$$\widetilde A(\vec q,\vec p) := \int_{t_0}^{t_1} \widetilde L(\vec q,\vec p,\dot{\vec q},\dot{\vec p})\,dt \quad \text{under the constraint} \quad p_{n+1}+H=0. \tag{1.10}$$

In either case it is promptly checked that the corresponding E-L equation is equivalent to the system of the canonical equations, cf. (I.8.20).

Jacobi principle and Riemannian metric. Descartes discovered that geometry may be treated analytically. Descartes's geometry was Euclidean: in the '600 there was no reason to think of any alternative. In his general theory of relativity Einstein introduced the idea of representing the dynamics via the Riemannian geometry, that had been developed in the '800. In that new theory at any point x the elementary length ds is defined by an equality of the form

$$(ds)^2 = g_{ij}(x)\,dx^i\,dx^j \qquad \text{(repeated indices are summed from 1 to } n\text{)}, \tag{1.11}$$

with $g(x)=\{g_{ij}(x)\}$ a symmetric and positive-definite matrix, which depends with continuity on x. More precisely, for any x, $g(x)$ is a tensor and is named a *Riemannian metric* (or Riemannian metric tensor); the Euclidean geometry is retrieved if $g(x)$ coincides with the identity tensor. For instance in Einstein's general theory of relativity the force of gravity is accounted for by the space-time metric tensor (which however in this case is not positive-definite); in this way that force is reduced to the geometrical structure.

More than half a century before Einstein, Jacobi had developed an analogous point of view for the geometrization of classical mechanics, in the case in which

(i) the Lagrangian L does not explicitly depend on time, and

(ii) the potential energy U does not depend on the (generalized) velocities.

We know that under the assumption (i), denoting by K the kinetic energy, the total energy $H := K+U$ is constant; let us denote by E this constant. Therefore

$$L = K - U = 2K - E. \tag{1.12}$$

As E is constant, we may replace the action functional $A := \int_{t_0}^{t_1} L\,dt$ by

$$\widetilde A := \int_{t_0}^{t_1} L\,dt + E(t_1-t_0) = \int_{t_0}^{t_1} (L+E)\,dt = \int_{t_0}^{t_1} 2K\,dt. \tag{1.13}$$

Moreover, under rather general conditions,

$$K(x,\dot x) = \tfrac12 m_{ij}(x)\,\dot x^i\,\dot x^j, \tag{1.14}$$

with $\{m_{ij}(x)\}$ a positive-definite symmetric matrix, which we may assume to depend with continuity on x. Let us then introduce the Riemannian metric $g_{ij}(x) = m_{ij}(x)$, that is (denoting by s the curvilinear abscissa)

$$(ds)^2 := m_{ij}(x)\,dx^i\,dx^j = 2K\,(dt)^2. \tag{1.15}$$

Therefore, for any curve $x : [t_0, t_1] \mapsto \mathbf{R}^n$ in the space of the configurations, denoting by ℓ the Riemannian length of the curve,

$$\begin{aligned}\widetilde{A} &= \int_{t_0}^{t_1} 2K\,dt = \int_{t_0}^{t_1} \sqrt{2K}\,\sqrt{2K}\,dt = \int_{t_0}^{t_1} \sqrt{2K}\,\frac{ds}{dt}\,dt \\ &= \int_0^{\ell} \sqrt{2K}\,ds = \int_0^{\ell} \sqrt{2(E-U)}\,ds \;\; \Big(= \int_{t_0}^{t_1} \sqrt{2(E-U(x))\,m_{ij}(x)\,\dot{x}^i\,\dot{x}^j}\,dt\Big).\end{aligned} \tag{1.16}$$

We infer that the trajectories of the system are geodesic curves of the configuration space equipped with the Riemannian metric (1.15), that are covered with time-law

$$\frac{ds}{dt}(t) = \sqrt{2K} = \sqrt{2[E - U(x(t))]} \qquad \forall t. \tag{1.17}$$

Under the hypotheses (i) and (ii), we have thus geometrized the dynamics of the Lagrangian system.

II.2. Analytical optics

The laws of light propagation, in particular of refraction and reflexion (and also a quantitative description of the mirages), can be derived from variational principles that are similar to those of classical mechanics. These principles were actually introduced for geometrical optics already in the second half of the 17th century, prior to the formulation of the analogous principles for mechanics by Lagrange and Hamilton. In the next section we shall also see that the laws of geometrical optics may be derived from the system of Maxwell laws for electromagnetism; they actually represent wave propagation, in the limit as the wave length vanishes.

Fermat's variational principle. Let us denote by n the *refraction index* of an isotropic and inhomogeneous medium, that is the ratio between the speed of light c in the vacuum and that in the medium $v = v(x)$:

$$n = n(x) := \frac{c}{v(x)}\ (\geq 1) \qquad \forall x \in \mathbf{R}^3. \tag{2.1}$$

We shall assume that the function $n : \mathbf{R}^3 \to \mathbf{R}$ is regular enough (essentially it will suffice that all the functions and derivatives that we shall write be either continuous or piecewise continuous). If dt is the time needed by light to cover an elementary distance ds in the medium, then obviously

$$dt = \frac{ds}{v(x)} = \frac{n(x)\,ds}{c}, \qquad c\,dt = \frac{c\,ds}{v(x)} = n(x)\,ds. \tag{2.2}$$

Therefore:

(i) in the medium the light travels the distance ds in the time $dt = n(x)\,ds/c$;

(ii) in vacuum in the time dt the light covers the distance $c\,dt = n(x)\,ds$, that is named elementary *optical length.*

In 1662 Fermat stated the principle that a light ray between two points P and Q follows a curve that is extremal for the (total) *optical length* [12]

$$A(\vec{r}) := \int_P^Q n(\vec{r})\,|d\vec{r}| = \int_0^{\widehat{L}} n(\vec{r}(\sigma))\,|\vec{r}\,'(\sigma)|\,d\sigma. \tag{2.3}$$

[12] For historical exactness, Fermat actually assumed minimality.

By (2.2) this is a travel time (up to the constant c), rather than a length.

Here $[0, \widehat{L}] \to \mathbf{R}^3 : \sigma \mapsto \vec{r}(\sigma)$ is the piecewise regular parameterization of any curve that joins P with Q; that is, $\vec{r}(0) = P$ and $\vec{r}(\widehat{L}) = Q$. The length $\widehat{L}$ of the parameteric interval is not known a priori, but this raises no difficulty.

Setting $T(\vec{r})$ equal to the time that is taken by the light ray $\vec{r}(\cdot)$ to go from P to Q, by (2.2) we have $A(\vec{r}) = c\,T(\vec{r})$. [13] Hence $\vec{r}(\cdot)$ minimizes $A(\vec{r})$ if and only if it minimizes the travel time between those two points. A similar conclusion holds for extremal paths.

Next we show some alternative parameterizations of the optical length functional.

First reparameterization: arc-length. If $\sigma = s$ (the arc-length parameter) then $|\vec{r}\,'(s)| = 1$; denoting by $\widehat{L}$ the unknown length of the curve, then

$$A(\vec{r}) = \int_0^{\widehat{L}} n(\vec{r}(s))\, ds. \tag{2.4}$$

If n is uniform, then $A(\vec{r})$ is obviously proportional to $\widehat{L}$. [14] When minimizing the optical length written in the form (2.4), account must be taken of the constraint $|\vec{r}\,'(s)| = 1$, which is implicit in the definition of the parameter s; here a Lagrange multiplier must thus be used for the search for extremals of the functional A. The representation (2.3) then looks more convenient.

Eikonal equation. [15] If we regard the integrand of (2.3)

$$f(\vec{r}, \vec{r}\,') := n(\vec{r})\, |\vec{r}\,'| \tag{2.5}$$

as a Lagrangian function, then the optical length $A(\vec{r})$ may be interpreted as an action functional. As in (2.5) f does not explicitly depend on the integration variable σ, the problem may be geometrized, as now we see.

If $n \in C^1(\mathbf{R}^3)$ and $\vec{r}$ is of class C^2, then the extremals of A fulfill the E-L equation in strong form. A simple computation shows that this equation reads

$$\frac{d}{d\sigma}\Big\{ n(\vec{r}(\sigma)) \frac{\vec{r}\,'(\sigma)}{|\vec{r}\,'(\sigma)|} \Big\} = |\vec{r}\,'(\sigma)| \nabla n(\vec{r}(\sigma)) \qquad \forall \sigma \in [0, \widehat{L}]. \tag{2.6}$$

This ODE is called the *eikonal equation.* Notice that the term $\vec{r}\,'(\sigma)/|\vec{r}\,'(\sigma)|$ is homogeneous of degree 0. As the curve is piecewise regular, $|\vec{r}\,'(\sigma)| \neq 0$ for a.e. σ.

By using the parameterization by arc-length $\sigma = s$, we have $|\vec{r}\,'(s)| = 1$. We may then rewrite the equation (2.6) in the form

$$\frac{d}{ds}\Big\{ n(\vec{r}(s)) \frac{d\vec{r}(s)}{ds} \Big\} = \nabla n(\vec{r}(s)) \qquad \forall s \in [0, \widehat{L}]. \tag{2.7}$$

In this case the Lagrangian is reduced to $n(\vec{r}(s))$. This seems to simplify the analysis, but (as we already noticed discussing (2.4)) in the search for extremals one must then prescribe the constraint $|\vec{r}\,'(s)| = 1$, and use a Lagrange multiplier.

---

[13] Up to the factor c the optical length is the travel time of the light ray, rather than the Euclidean distance: one might thus refer to it as the *optical time.* We shall see that the optical length may actually be interpreted as a Riemannian distance.

[14] So in this case the optical length does represent a length (rather than a travel time)...

[15] The term eikonal is used for several objects (functions, equations, and so on) that concern geometrical optics.

In a discontinuous medium the field n is just piecewise C^0, and in general an extremal $\vec{r}$ of A will just be piecewise C^1. We may then derive the E-L equation in weak form, or equivalently in the integral form of Du Bois-Reymond: for a suitable $\vec{C} \in \mathbf{R}^3$,

$$n(\vec{r}(s))\,\frac{d\vec{r}(s)}{ds} = \vec{C} + \int_0^s \nabla n(\vec{r}(\sigma))\,d\sigma \qquad \text{for a.e. } s \in [0, \widehat{L}]. \tag{2.8}$$

A different Lagrangian. Notice that the optical Lagrangian (2.5) is positively homogeneous of degree 1 in the variable $\vec{r}\,'$. This corresponds to the *optical momentum*

$$\vec{p} := \nabla_{\vec{z}} f(\vec{r}, \vec{r}\,')\big|_{\vec{z}=\vec{r}\,'} = \nabla_{\vec{z}} \left\{ n(\vec{r})\,|\vec{z}| \right\}\big|_{\vec{z}=\vec{r}\,'} = n(\vec{r})\,\frac{\vec{r}\,'}{|\vec{r}\,'|}. \tag{2.9}$$

We may thus rewrite the optical length (2.3) in the form

$$\widehat{A}(\vec{r}) := \int_0^L \vec{p}(\sigma)\cdot d\vec{r}(\sigma); \tag{2.10}$$

here $\vec{p}$ clearly appears as the conjugate variable of $\vec{r}$. In this case the Legendre transform cannot be applied, since the Hessian matrix of the function $\vec{z} \mapsto n(\vec{r})\,|\vec{z}|$ is singular. (Formally we would get $H = \vec{p}\cdot\vec{r}\,' - n(\vec{r})\,|\vec{r}\,'| \equiv 0$.)

Second reparameterization and Hamiltonian formulation. Let us now reparameterize the light ray via a parameter ρ such that $ds = n\,d\rho$. [16] Multiplying both members of (2.7) by $n(\vec{r}(\rho))$, we get the equivalent equation

$$\frac{d^2}{d\rho^2}\vec{r}(\rho) = \nabla\,\frac{n(\vec{r}(\rho))^2}{2} \qquad \forall \rho. \tag{2.11}$$

This equation has the same structure as Newton's law for conservative fields (!), which also reads $\vec{r}\,''(t) = \nabla[-V(\vec{r}(t))/m]$. The equation (2.11) may thus be associated to the new Lagrangian

$$\widetilde{L}(\vec{r}, \vec{r}\,') := \frac{|\vec{r}\,'|^2}{2} + \frac{n(\vec{r})^2}{2}; \tag{2.12}$$

this is homogeneous of degree 2 in the variable $\vec{r}\,'$, at variance with $n(\vec{r})|\vec{r}\,'|$ (cf. (2.3)) which is homogeneous of degree 1. This function $\widetilde{L}$ is analogous to the Lagrangian that we met in mechanics.

As the Hessian matrix of the function $\vec{z} \mapsto \widetilde{L}(\vec{r}, \vec{z})$ is nonsingular for any $\vec{z}$ and any $\vec{r} \neq \vec{0}$, we may apply the Legendre transform to this function. By setting

$$\vec{p} := D_{\vec{r}\,'}\widetilde{L}(\vec{r}, \vec{r}\,') = \vec{r}\,', \qquad \widetilde{H} := \vec{p}\cdot\vec{r}\,' - \widetilde{L}, \tag{2.12$'$}$$

we get the Hamiltonian

$$\widetilde{H}(\vec{r}, \vec{p}) = \frac{|\vec{p}|^2}{2} - \frac{n(\vec{r})^2}{2}. \tag{2.13}$$

As $\widetilde{H}_t = 0$, this is a first integral of the eikonal equation.

[16] Despite of the reparameterization, we keep on using the notation $\vec{r}(\cdot)$.

*** Third reparameterization: optical length.** Let us reparameterize the light-ray curve by means of the parameter *optical length:*

$$\tau(s) := \int_0^s n(\vec{r}(v))\, dv \qquad \forall s \in [0, \widehat{L}], \tag{2.14}$$

so that (as $(ds)^2 = d\vec{r}\cdot d\vec{r}$)

$$d\tau(s) = n(\vec{r}(s))\, ds = n(\vec{r}(\tau))^2\, \vec{r}\,'(\tau)\cdot\vec{r}\,'(\tau)\, d\tau.$$

Selecting the parameter $\sigma = \tau$ in (2.3) and setting $\bar{L} := \tau(\widehat{L})$, we may then represent the action functional A as follows

$$\widetilde{A}(\vec{r}) := \int_0^{\bar{L}} n(\vec{r}(\tau))^2\, \vec{r}\,'(\tau)\cdot\vec{r}\,'(\tau)\, d\tau. \tag{2.15}$$

The E-L equation of this functional yields once more the eikonal equation (2.11).

In conclusion, by changing the parameterization in the action functional, different Lagrangians were obtained. All of them however correspond to the same optical length, and to the same E-L equation: the eikonal equation.

Wave fronts and Huygens law. A and $\widetilde{A}$ depend on the function r, and thus are two functionals. After fixing the point P in which the light source stands, let us now consider the optical length as a function of the arrival point $Q = x$ of the light ray:

$$S(x) := \inf\Big\{ \int_0^{\widetilde{L}} n(\vec{\ell}(\sigma))|\vec{\ell}\,'(\sigma)|\, d\sigma : \vec{\ell} \in C^1([0,\widehat{L}]), \vec{\ell}(0) = P,\ \vec{\ell}(\widehat{L}) = x\Big\} \qquad \forall x \in \mathbf{R}^3. \tag{2.16}$$

We know that $S(x)$ is proportional to the time taken by the ray to go from P to x. Each level surface Σ of this function thus consists of the points that are simultaneously reached by the light ray. Σ is thus a *wave front.*

For an isotropic medium it is then clear that any light ray is orthogonal to Σ in the intersection point of the ray with this surface. Therefore, denoting by $\vec{r}$ a light ray through the point x, [17]

$$\nabla S(x)\cdot d\vec{r}(x) = |\nabla S(x)|\, ds = n(x)\, ds. \tag{2.17}$$

Let us define the vector field

$$\vec{n}(x) := n(x)\frac{d\vec{r}}{|d\vec{r}|}\ \Big(= n(x)\frac{d\vec{r}}{ds}\Big), \tag{2.18}$$

which depends on the (known) field n, and on the (a priori arbitrary) direction of the light ray. For an isotropic medium we may thus refine (2.17) by formulating the classical *Huygens law:*

$$\nabla S(x) = \vec{n}(x) \qquad \forall x \in \mathbf{R}^3. \tag{2.19}$$

Huygens's principle. [18] Let $\Phi_{\vec{r}_0}(t)$ be the wave front generated at time t by a source that sits at a point $\vec{r}_0$ at the time 0. Similarly, let $\Phi_{\vec{r}}(s)$ be the wave front generated by

[17] It is usual to write no arrow over x.

[18] This dates back to the '600. This principle was then refined by Fresnel, who introduced the notion of interference.

$\vec r \in \Phi_{\vec r_0}(t)$ after a time s. Then $\Phi_{\vec r_0}(t+s)$ (i.e., the wave front generated by $\vec r_0$ after a time $t+s$) is the envelope of the wave fronts $\Phi_{\vec r}(s)$, as $\vec r$ ranges in $\Phi_{\vec r_0}(t)$. This is not difficult to be proved.

Isotropic medium. In this case the latter result also stems from the stated orthogonality between rays and wave fronts. (If the medium is also homogeneous then the wave fronts are envelopes of Euclidean spheres, and the result is trivial.) Notice that

$$(2.19) \qquad \Leftrightarrow \qquad \begin{cases} (2.17) \text{ and} \\ \text{light rays are orthogonal to wave fronts.} \end{cases} \tag{2.20}$$

Refraction law. (2.19) entails $\nabla \times \vec n = \vec 0$ in $\mathbf{R}^3$. Therefore if Σ is a regular surface of discontinuity for the field $\vec n$ and $\vec \nu$ is normal to the surface, then $\vec r$ is just piecewise C^1 and the Huygens law entails

$$\vec \nu \times \vec n \ \text{ is continuous across } \Sigma; \tag{2.21}$$

that is, the tangential component of $\vec n$ is continuous across Σ. This is equivalent to the classical *refraction law of Ibn Shal and Snell:* denoting by θ the angle between the light ray and the normal to Σ,

$$n \sin\theta \ \text{ is continuous across } \Sigma. \tag{2.22}$$

A similar argument applies to the reflection of rays by a mirror: also in that case $\vec r$ is piecewise C^1, and the reflection preserves the tangential component of $\vec n$.

The Huygens law also provides a quantitative description of mirages (here omitted).

Optical Hamilton-Jacobi equation. (2.19) obviously also entails the following scalar law, which is an example of (stationary) *Hamilton-Jacobi equation:* [19]

$$|\nabla S(x)| = n(x) \qquad \forall x \in \mathbf{R}^3. \tag{2.23}$$

Anisotropic medium. So far we dealt with a (possibly nonhomogeneous) isotropic medium. If instead the medium is anisotropic, then in general the light ray is not orthogonal to the wave front. In this case we may assume that there exists an invertible tensor-field $L = \{L_{ij}\}$ such that

$$\nabla S(\vec r) = L(\vec r) \cdot \frac{d\vec r}{|d\vec r|} \qquad \text{i.e.,} \qquad \frac{\partial}{\partial x_i} S(\vec r) = L_{ij}(\vec r)\, \frac{dr_j}{ds} \qquad \forall i \tag{2.24}$$

(still implying the sum over repeated indices). Setting

$$\widetilde n(\vec r, \vec v) := \sqrt{L_{ik}(\vec r)\, L_{jk}(\vec r)\, v_i\, v_j} \qquad \forall \vec r, \vec v \in \mathbf{R}^3, \tag{2.25}$$

here the optical length reads

$$A(\vec r) = \int_0^{\widehat L} \widetilde n(\vec r(s), \vec r\,'(s))\, ds, \tag{2.26}$$

similarly to the case of isotropic media, cf. (2.4). In this way the problem is completely geometric, in analogy with the Jacobi principle for mechanics, see Sect. II.1.

[19] For this reason we use the notation S, that we already introduced for the general H-J equation. The term *eikonal* is often used for this equation and for the function S.

Defining the optical momentum

$$\vec{p}(\vec{r},\vec{v}) := \nabla_{\vec{v}}\,\widetilde{n}(\vec{r},\vec{v}) \qquad \forall \vec{r},\vec{v}\in\mathbf{R}^3, \tag{2.27}$$

that is,

$$p_i(\vec{r},\vec{v}) := \frac{L_{ik}(\vec{r})\,L_{jk}(\vec{r})\,v_j}{\widetilde{n}(\vec{r},\vec{v})} \qquad \forall \vec{r},\vec{v}\in\mathbf{R}^3, \forall i, \tag{2.28}$$

the E-L (eikonal) equation reads

$$\frac{d}{ds}\vec{p}(\vec{r}(s),\vec{r}\,'(s)) = \nabla_{\vec{r}}\widetilde{n}(\vec{r}(s),\vec{r}\,'(s)) \qquad \forall s; \tag{2.29}$$

this generalizes (2.7). Note that by (2.28)

$$\vec{p}(\vec{r}(s),\vec{r}\,'(s))\cdot d\vec{r} = \widetilde{n}(\vec{r}(s),\vec{r}\,'(s))\,ds \qquad \forall s. \tag{2.30}$$

Riemannian metric. We may interpret the tensor field $L^\tau(x)L(x)$ as a Riemannian metric-tensor,[(20)] which corresponds to the squared element of length

$$\widetilde{n}(x)^2 ds^2 = L_{ik}(x)\,L_{jk}(x)\,dx_i(s)\,dx_j(s) \qquad \forall x\in\mathbf{R}^3. \tag{2.31}$$

We may then define the geodesic spheres as the sets of the points that are geodesically equidistant (in the sense of optical length) from a fixed point — the center of the geodesic sphere. Extending in a natural way this notion of distance, one defines families of geodesically equidistant surfaces in $\mathbf{R}^3$: these are envelopes of the geodesic spheres that have a fixed radius and center on a surface of the family. One may prove the following results:

(i) Let S be a solution of the equation (2.24), $C\in\mathbf{R}$ and $R>0$. The surface $S(x) = C+R$ is then the envelope of the geodesic spheres of radius R (in the sense of optical length) with center on the surface $S(x)=C$. We thus retrieve the Huygens's principle.

(ii) A family of level surfaces for the function S is geodesically equidistant with respect to metric $\{n_{ij}(x)\}$ if and only if S fulfills the corresponding H-J equation (2.19).

Optical axis and Hamiltonian. Let us assume that the light rays may be represented as graphs of functions $\mathbf{R}\to\mathbf{R}^2 : x\mapsto \vec{u} = (u_1,u_2)$. One then says that the x-axis is the *optical axis* of the optical system. In this case the optical length equals

$$A(\vec{u}) := \int_{x_0}^{x_1} \widetilde{n}(x,\vec{u}(x))\sqrt{1+|\vec{u}\,'(x)|^2}\,dx. \tag{2.32}$$

This corresponds to the Hamiltonian

$$H(\vec{p}) := -\int_{x_0}^{x_1} \sqrt{\widetilde{n}(x,\vec{u}(x))^2 - |\vec{p}(x)|^2}\,dx, \tag{2.33}$$

and to the eikonal equation

$$S_x(x,\vec{u})^2 + S_{\vec{u}}(x,\vec{u})^2 = \frac{\widetilde{n}(x)^2}{c^2} \qquad \forall (x,\vec{u})\in\mathbf{R}^3. \tag{2.34}$$

(20) to wit, a positive-definite bilinear symmetric form on each tangent space, that depends continuously on the point.

Synthesis. First we dealt with light propagation in an isotropic inhomogeneous medium. Following Fermat, we formulated an action-type variational principle, which represents light rays as extremals of the optical length. The corresponding E-L equation yields the eikonal equation, and by changing the integration parameter we derived alternative variational formulations; one of them provides a Newton-type dynamics and is associated to a Hamiltonian formulation. We saw that the eikonal equation accounts for the classical laws of refraction and reflexion.

We then stated the Huygens principle, which describes light propagation in terms of the advancement of wave fronts. These fronts consist of points that have the same optical distance S from the light source; for isotropic media they are orthogonal to light rays. We derived a H-J equation for the field $S = S(x)$. In the case of anisotropic media, this corresponds to assuming a Riemannian metric. In this framework we reformulated the eikonal equation in terms of the optical momentum.

II.3. From the Maxwell equations to geometrical optics

Asymptotics for infinitesimal wave length. In an isotropic medium a monochromatic wave of intensity $A(x)$ that is emitted by a point at the instant $t = 0$ may be described by a *wave field* of the form

$$\Psi(x,t) := A(x)e^{i[\varphi_0(x)-ct]} \qquad \forall x \in \mathbf{R}^3, \forall t > 0, \tag{3.1}$$

with c : velocity of light in vacuum. The surfaces

$$\Sigma_t := \{x \in \mathbf{R}^3 : \varphi_0(x) = ct\} \qquad \forall t > 0$$

represent the wave fronts, namely the set of points that are attained by the light ray at a same instant t.

If $x = x(t)$ is the parameteric equation of a light ray, then $\varphi_0(x(t)) = ct$ for any t; therefore

$$\nabla\varphi_0(x(t))\cdot x'(t) = c \qquad \forall t > 0. \tag{3.2}$$

Because of the Huygens principle, $x'(t)$ is orthogonal to the wave front, which then advances with scalar speed

$$v(x) = \frac{c}{|\nabla\varphi_0(x)|}. \tag{3.3}$$

For an isotropic medium we thus retrieve the scalar eikonal equation

$$|\nabla\varphi_0(x)| = \frac{c}{v(x)} =: n(x) \qquad \forall x \in \mathbf{R}^3. \tag{3.4}$$

From the wave equation to the eikonal equation. The field Ψ may represent for instance either the electric or the magnetic field. These fields fulfill the system of the Maxwell equations; it is easily checked that the function Ψ then fulfills the classical wave equation of D'Alembert:

$$D_t^2\Psi - v^2\Delta\Psi = 0 \qquad \forall x \in \mathbf{R}^3, \forall t > 0, \tag{3.5}$$

v being the wave velocity in the medium. [21] As we pointed out, in an isotropic homogeneous medium a monochromatic wave of intensity $A(x)$, that is emitted by a point at the instant $t = 0$, may be described by a field of the form

$$\Psi(x,t) := A(x)e^{i[\varphi_0(x)-ct]} \qquad \forall x \in \mathbf{R}^3, \forall t > 0. \tag{3.6}$$

[21] To tell the truth, the wave equation may be written in this simplified form only if v is uniform, that is for a homogeneous medium.

Replacing (3.6) into (3.5), via some calculations one gets

$$\Delta A - A\,|\nabla\phi_0|^2 + \frac{c^2}{v^2}\,A + i\big(2\nabla A\cdot\nabla\phi_0 + A\Delta\phi_0\big) = 0 \qquad \forall x \in \mathbf{R}^3, \forall t > 0, \tag{3.7}$$

that is, separating the real from the imaginary part,

$$\Delta A - A\,|\nabla\phi_0|^2 + \frac{c^2}{v^2}\,A = 0 \qquad \forall x \in \mathbf{R}^3, \forall t > 0, \tag{3.8}$$

$$2\nabla A\cdot\nabla\phi_0 + A\Delta\phi_0 = 0 \qquad \forall x \in \mathbf{R}^3, \forall t > 0. \tag{3.9}$$

As $A \neq 0$, we may rewrite (3.8) in the form

$$\frac{\Delta A}{A} - \Big(|\nabla\phi_0|^2 - \frac{c^2}{v^2}\Big) = 0 \qquad \forall x \in \mathbf{R}^3, \forall t > 0. \tag{3.10}$$

The coefficient c^2/v^2 is inversely proportional to the square of the wave length that the light has in the vacuum. Assuming that this wave length is short with respect to the distance at which the refraction index has sensible variations, one may show that the first addendum of (3.10) is negligible. Once more we then retrieve the eikonal equation

$$|\nabla\phi_0|^2 = \frac{c^2}{v^2} =: n^2. \qquad \forall x \in \mathbf{R}^3. \tag{3.11}$$

Maxwell's theory of electromagnetism thus yields geometrical optics, as the wave length vanishes. (Classical mechanics similarly stems from the Schrödinger equation, as the wave length vanishes.)

II.4. Hamilton-Jacobi equation

Derivation of the Hamilton-Jacobi equation from the Euler-Lagrange equation. Let $f \in C^0([a,b]\times\mathbf{R}\times\mathbf{R})$ be a given function. Letting the pair (x_1, y_1) vary, now we introduce what Hamilton named the *characteristic function:*

$$S(x_1, y_1) = \int_{x_0}^{x_1} f(x, u(x), u'(x))\,dx \qquad \forall (x_1, y_1) \in \mathbf{R}^2, \tag{4.1}$$

u being a curve of class C^1, whose graph joins (x_0, y_0) with (x_1, y_1). This function plays a relevant role in the canonical theory in physics, and more generally in the theory of linear first-order PDEs. (The analogy with the function (2.16) may be noticed.)

Let us write (x, y) in place of (x_1, y_1). After noticing that $u(x) = y$ (i.e., $u(x_1) = y_1$), let us introduce the momentum

$$p(x, y) := D_3 f(x, y, u'(x)) \qquad \forall (x, y) \in \mathbf{R}^2. \tag{4.2}$$

(Here p does not depend on $u'(x)$, since the function $u(\cdot)$ is kept fixed.) For a moment let us regard (4.1) as a functional of u, and assume that u is an extremal. The general variation formula (I.6.3) then yields

$$dS(x, y) = p(x, y)\,dy + [f(x, y, u'(x)) - p(x, y)\,u'(x)]\,dx \qquad \forall (x, y) \in \mathbf{R}^2, \tag{4.3}$$

that is,

$$\begin{aligned} &S_y(x,y) = p(x,y) \\ &S_x(x,y) = f(x,y,u'(x)) - p(x,y)\,u'(x) \end{aligned} \qquad \forall (x,y) \in \mathbf{R}^2. \tag{4.4}$$

Let us assume that $D_3^2 f(x,u,z) \neq 0$ for any (x,u,z), so that (4.2) may be made explicit with respect to $u'(x)$. Implying this relation, let us define the Hamiltonian function

$$H(x,y,p) := p\,u'(x) - f(x,y,u'(x)) \qquad \forall (x,y,p) \in \mathbf{R}^3, \tag{4.5}$$

still with y in place of $u(x)$. As we already remarked, because of (4.2) H does not depend on $u'(x)$. Eliminating p in the system (4.4) and using the definition (4.5), we get the *Hamilton-Jacobi equation* (more shortly, "H-J equation")

$$S_x(x,y) + H(x,y,S_y(x,y)) = 0 \qquad \forall (x,y) \in \mathbf{R}^2. \tag{4.6}$$

Derivation of the canonical equations from the Hamilton-Jacobi equation. We just derived the H-J equation from the E-L equation. Next we show that conversely a solution of the E-L equation (equivalently, of the canonical equations) may be retrieved from a solution of the H-J equation.

Let us assume that the function S is of class C^2 and solves the H-J equation (4.6); [(22)] let the function $x \mapsto y(x)$ be of classe C^1 and such that

$$y'(x) = H_p(x, y(x), S_y(x, y(x))) \qquad \forall x \in [x_0, x_1]. \tag{4.7}$$

Notice that this equation is uncoupled from that in p. This is a simple ODE of first order; under standard assumptions, it thus has a solution that also fulfills the condition $y(x_0) = y_0$.

Let us next set

$$p(x) := S_y(x, y(x)) \qquad \forall x \in [x_0, x_1], \tag{4.8}$$

so that (4.7) also reads

$$y'(x) = H_p(x, y(x), p(x)) \qquad \forall x \in [x_0, x_1]. \tag{4.8'}$$

This is one of the two canonical equations; more specifically, this is the one that directly follows from the Legendre transform. Next we derive the other canonical equation, too.

The definition (4.8) yields

$$p'(x) = S_{xy}(x, y(x)) + S_{yy}(x, y(x))\,y'(x) \qquad \forall x \in [x_0, x_1]. \tag{4.9}$$

On the other hand, differentiating the H-J equation (4.6) with respect to y we have

$$S_{xy}(x,y) + H_y(x,y,S_y(x,y)) + H_p(x,y,S_y(x,y))\,S_{yy}(x,y) = 0; \tag{4.10}$$

by evaluating this equation for $y = y(x)$ and reminding (4.7), we get

$$S_{xy}(x, y(x)) + H_y(x, y(x), S_y(x, y(x))) + y'(x)\,S_{yy}(x, y(x)) = 0. \tag{4.11}$$

By comparing (4.9) and (4.11), we obtain

$$p'(x) = -H_y(x, y(x), S_y(x, y(x))) = -H_y(x, y(x), p(x)) \qquad \forall x \in [x_0, x_1]; \tag{4.12}$$

the pair of functions $x \mapsto (y(x), p(x))$ thus also solves the second canonical equation in $[x_0, x_1]$, and we know that this is equivalent to the E-L equation.

Characterization of first integrals.

[(22)] For the solution of the H-J equation in general this regularity property is not obvious.

Theorem 4.1. *(Jacobi) Let $\{S(x,y,\alpha) : \alpha \in V\}$ be a family of solutions of class C^2 of the H-J equation, parameterized by a real α in a neighborhood V of $\alpha = 0$. Then:*

(i) If $y = y(x)$ fulfill the E-L equation for the functional F associated to the Lagrangian f, then

$$\frac{d}{dx} S_\alpha(x, y(x), \alpha) = 0 \qquad \forall x \in [x_0, x_1], \forall \alpha, \tag{4.13}$$

that is, for any α, $S_\alpha(x,y,\alpha)$ is a first integral.

(ii) On the other hand, if S fulfills the H-J equation and $y = y(x)$ satisfies (4.13), then

$$y'(x) = H_p(x, y(x), S_y(x, y(x), \alpha)) \qquad \forall x \in [x_0, x_1], \forall \alpha. \tag{4.14}$$

As we saw, setting $p(x) := S_y(x, y(x))$, (4.14) entails that the pair $x \mapsto (y(x), p(x))$ fulfills the canonical equations. Equivalently, the function $x \mapsto y(x)$ solves the E-L equation for the functional F associated to the Lagrangian f.

Proof. As we just saw, defining

$$p(x) := S_y(x, y(x), \alpha), \tag{4.15}$$

the pair of functions $x \mapsto (y(x), p(x))$ solves the canonical equations, for any α. [(23)] Denoting partial derivatives by indices (thus setting e.g. $S_\alpha := D_\alpha S$), we have

$$\frac{d}{dx} S_\alpha(x, y(x), \alpha) = S_{x\alpha}(x, y(x), \alpha) + S_{y\alpha}(x, y(x), \alpha)\, y'(x) \qquad \forall x, \alpha. \tag{4.16}$$

On the other hand by differentiating the H-J equation we get

$$S_{x\alpha}(x, y(x), \alpha) + H_p(x, y(x), S_y(x, y(x), \alpha))\, S_{y\alpha}(x, y(x), \alpha) = 0. \tag{4.17}$$

Moreover one of the canonical equations (the one that corresponds to the Legendre transform) yields

$$y'(x) = H_p(x, y(x), p(x)) \overset{(4.15')}{=} H_p(x, y(x), S_y(x, y(x), \alpha)). \tag{4.18}$$

By the three latter equations we get (4.13)

$$\frac{d}{dx} S_\alpha(x, y(x), \alpha) = 0 \qquad \forall x \in [x_0, x_1], \forall \alpha. \tag{4.19}$$

Let us now come to the converse statement. By (4.13)

$$\frac{d}{dx} S_\alpha(x, y(x), \alpha) = 0 \qquad \forall x \in [x_0, x_1], \forall \alpha, \tag{4.20}$$

then

$$y'(x) = H_p(x, y(x), S_y(x, y(x), \alpha)) \qquad \forall x \in [x_0, x_1], \forall \alpha. \tag{4.21}$$

As we already saw, setting $p(x) := S_y(x, y(x), \alpha)$ this entails that the pair of functions (y, p) fulfills the canonical equations.

[(23)] A priori p seems to depend on α. But, as p is determined by y and we assumed the extremal to be unique, p does not depend on α!

On the integration of the H-J equation. As we already saw, this equation may be used to define a solution of the canonical equations. But often it is more difficult to solve the H-J equation than the canonical system. Anyway next we show that, if the Hamiltonian function does not explicitly depend on x, then the H-J equation may be easily integrated.

Proposition 4.2. *Let $y \mapsto \widehat{S}(y,\alpha)$ be a family of functions parameterized by $\alpha \in \mathbf{R}$ such that*

$$H(y, \widehat{S}_y(y,\alpha)) = \alpha \qquad \forall (y,\alpha) \in \mathbf{R}^2. \tag{4.22}$$

The H-J equation is then solved by the family of functions

$$S(x,y,\alpha) := \widehat{S}(y,\alpha) - \alpha x \qquad \forall (x,y,\alpha) \in \mathbf{R}^2. \tag{4.23}$$

(For any y, obviously the equation (4.22) has a solution $S(y,\alpha)$ for any α if and only if the function $H(y,\cdot)$ is invertible.)

Proof. The definition (4.23) entails that

$$S_x(x,y,\alpha) = -\alpha, \qquad S_y(x,y,\alpha) = \widehat{S}_y(y,\alpha) \qquad \forall (y,\alpha) \in \mathbf{R}^2. \tag{4.24}$$

By replacing α and $\widehat{S}_y(y,\alpha)$ in (4.22), it follows that $S_x + H(y, S_y) = 0$. □

Synthesis. (i) Given a Lagrangian f, we introduced Hamilton's characteristic function

$$S(x,y) := \int_{x_0}^{x} f(s, u(s), u'(s))\, ds \qquad \forall (x,y) \in \mathbf{R}^2 \tag{4.25}$$

and the Hamiltonian function

$$H(x,y,p) := p\, z - f(x,y,z) \qquad \forall (x,y,p) \in \mathbf{R}^3. \tag{4.26}$$

We then derived the partial differential equation of Hamilton-Jacobi:

$$S_x(x,y) + H(x,y,S_y(x,y)) = 0 \qquad \forall (x,y) \in \mathbf{R}^2. \tag{4.27}$$

(ii) Given a solution S of the H-J equation, we saw that, if

$$y'(x) = H_p(x, y(x), S_y(x, y(x))), \qquad p(x) := S_y(x, y(x)) \qquad \forall x \in [x_0, x_1] \tag{4.28}$$

(the first one is an equation at the ordinary derivatives and is independent from the second equation, which is just a definition), then the pair of functions (y,p) solves the canonical system. The function $y = y(x)$ then solves the E-L equations.

(iii) We saw that, if $S(x,y,\alpha)$ (with α a real parameter) is a family of solutions of the H-J equation, then $y = y(x)$ is an extremal for the functional F if and only if $S_\alpha(x,y,\alpha)$ is a first integral for any α.

About the characteristics of partial differential equations (PDEs). The H-J equation is a PDE, whereas the E-L equation and the canonical equations are ODEs. These equations are strictly related: the E-L equation (that are equivalent to the canonical equations) determine the characteristic curves [24] of the H-J equation. The latter are the curves along which the signal propagates, in the sense that now we outline.

[24] This use of the term *characteristic* is not related to that of *characteristic function.*

For instance, let

$$\lambda \in \mathbf{R}, \qquad u^0 \in C^0(\mathbf{R}), \qquad a \in C^0([0,T]), \qquad a(t) > 0 \qquad \forall t \in [0,T],$$

and couple the following PDE

$$a(t)\, u_x(x,t) + u_t(x,t) = \lambda u(x,t) \qquad \forall (x,t) \in \mathbf{R} \times [0,T] \tag{4.29}$$

with the initial condition

$$u(x,0) = u^0(x) \qquad \forall x \in \mathbf{R}. \tag{4.30}$$

Let the function $x = x(t)$ be such that

$$x'(t) = a(t) \qquad \forall t \in [0,T]; \tag{4.31}$$

setting $A(t) := \int_0^t a(s)\, ds$ for any $t \in [0,T]$, this equation has a family of solutions:

$$x(t) = z + A(t) \qquad \forall t \in [0,T], \forall z \in \mathbf{R}.$$

Setting $w(t) := u(x(t),t)$ for any t, we have

$$w'(t) = a(t)\, u_x(x,t) + u_t(x,t) \overset{(4.29)}{=} \lambda w(t) \qquad \forall t \in [0,T], \tag{4.32}$$

Hence $w(t) = C_z\, e^{at}$ for any t, that is,

$$u(z + A(t), t) = C_z\, e^{\lambda t} \qquad \forall t \in [0,T],$$

whence it follows that $C_z = u(z,0) \overset{(4.30)}{=} u^0(z)$ for any $z \in \mathbf{R}$. Therefore

$$u(z + A(t), t) = u^0(z)\, e^{\lambda t} \qquad \forall (z,t) \in \mathbf{R} \times [0,T];$$

that is, setting $x = z + A(t)$,

$$u(x,t) = u^0(x - A(t))\, e^{\lambda t} \qquad \forall (x,t) \in \mathbf{R} \times [0,T]. \tag{4.33}$$

(This function clearly fulfills the partial differential equation and the initial condition.) For instance if $a(t) \equiv 1$ we have $u(x,t) = u^0(x-t)\, e^{\lambda t}$.

In this simple example the ODE $x'(t) = a(t)$ is the equation of the characteristics of the PDE $a(t)\, u_x + u_t = \lambda u$; moreover $w'(t) = \lambda w(t)$ is the equation along the characteristics.

These properties of characteristics apply to a much wider family of first-order partial differential equations. [25] For these equations the solution of the equation of characteristics allows one to construct the solution of the partial differential equation, and conversely; this is similar to what we saw for the canonical equations and the H-J equation.

The H-J equation of the harmonic oscillator. The equation of the harmonic oscillator is

$$m u''(t) + k u(t) = 0. \tag{4.34}$$

Defining the kinetic momentum $p := mu'$, this corresponds to the Hamiltonian

$$H(u,p) = \frac{p^2}{2m} + \frac{ku^2}{2}. \tag{4.35}$$

The H-J equation then reads

$$S_x(x,u) + \frac{S_u(x,u)^2}{2m} + \frac{ku^2}{2} = 0. \tag{4.36}$$

[25] These are either scalar or vector equations. As it is well known vector equations are equivalent to systems of scalar equations. Moreover any equation of any order n is equivalent to a system of n equations of first order.

III. MINIMIZATION, VARIATIONAL INEQUALITIES AND Γ-CONVERGENCE

III.1. Direct method of minimization (Tonelli theorem)

Topological properties. Let S be a topological space and $J : S \to]-\infty, +\infty]$; here we deal with the problem of finding $x_0 \in S$ such that $J(x_0) = \inf J$. By the classical Weierstrass theorem, if S is a compact topological space and J is continuous then this problem has a solution. However these assumptions also yield the existence of a maximum, and this suggests that they might be redundant. Indeed we shall see that, when dealing with minimization, continuity may be replaced by lower semicontinuity.

Let us denote the set of minimizers of J by $\mathcal{M}_J$; that is,

$$\mathcal{M}_J := \{x \in S : J(x) = \inf J\}.$$

We remind the reader that:

(i) J is said *lower semicontinuous* if and only if any sublevel set $S_{\tilde{a}} := \{x \in S : J(x) \le \tilde{a}\}$ is closed (or equivalently, any superlevel set $\{x \in S : J(x) > \tilde{a}\}$ is open);

(ii) J is said *sequentially lower semicontinuous* if and only if any sublevel set $S_{\tilde{a}}$ is sequentially closed (namely, the limit of every converging sequence in $S_{\tilde{a}}$ is also an element of $S_{\tilde{a}}$);

(iii) J is said *coercive* if and only if any sublevel set $S_{\tilde{a}}$ is relatively compact (namely, its closure is compact);

(iv) J is said *sequentially coercive* if and only if any sublevel set $S_{\tilde{a}}$ is relatively sequentially compact (namely, its closure is sequentially compact; namely, every sequence in $S_{\tilde{a}}$ has a subsesquence that converges to an element of the closure of $S_{\tilde{a}}$).

Notice that (i) entails (ii). The converse implication holds, and (iii) is equivalent to (iv), if the topological space S fulfills the *first countability axiom* (namely, any point has a countable basis of neighbourhoods). For instance, metric spaces have this property, at variance with any infinite-dimensional Banach space equipped with the weak topology.

*** About sequential properties.** In passing, we illustrate a topological construction. In any topological space (X, τ) the sequential topology, denoted by seq-τ, that is associated to τ is defined as follows. We say that a set $B \subset X$ is closed with respect to seq-τ if and only if it contains the limit of any sequence of elements of B that converges with respect to τ. The *sequential topology* seq-τ is finer than the original topology, i.e. $\tau \subset$ seq-τ. If the topology τ fulfills the first countability axiom (e.g., if it is metrizable), then it coincides with seq-τ; otherwise seq-τ is strictly finer than τ.

The sequential topology allows one to restate several sequential properties of the original topology. E.g., any set is sequentially compact with respect to τ if and only if it is compact with respect to seq-τ; a function $X \to \mathbf{R}$ is sequentially lower semicontinuous with respect to τ if and only if it is lower semicontinuous with respect to seq-τ; and so on. (One may thus conclude that the sequential lower semicontinuity and the sequential compactness are *topological properties).*

A Topological Result.

Theorem 1.1. *(Tonelli, 1920–23) Let S be a nonempty topological space and $J : S \to]-\infty, +\infty]$.*

(i) If J is lower semicontinuous and if for some $\tilde{a} \in \mathbf{R}$ the sublevel set $S_{\tilde{a}}$ is nonempty and compact [e.g., if J is coercive this holds], then $\mathcal{M}_J$ is nonempty and compact.

(ii) If J is sequentially lower semicontinuous and if for some $\tilde{a} \in \mathbf{R}$ the sublevel set $S_{\tilde{a}}$ is

nonempty and sequentially compact [e.g., this holds if J is sequentially coercive], then $\mathcal{M}_J$ is nonempty and sequentially compact.

Proof. (i) At first let us assume that J is lower semicontinuous, and that $S_{\tilde{a}}$ is a nonempty and compact sublevel set. The family $\mathcal{F} := \{S_a : a \le \tilde{a}, S_a \neq \emptyset\}$ consists of closed subsets of $S_{\tilde{a}}$, by the lower semicontinuity of J. These sublevel sets are nested, so that the intersection of any finite subfamily is nonempty. By the compactness of $S_{\tilde{a}}$, the intersection of the whole family, $\cap\mathcal{F}$, is then nonempty and compact. It then suffices to notice that $\cap\mathcal{F} = \mathcal{M}_J$. The compactness of $\mathcal{M}_J$ may similarly be checked.

(ii) Let us now assume that J is sequentially lower semicontinuous, and that $S_{\tilde{a}}$ is a nonempty and sequentially compact sublevel set. By definition of $\inf J$, there exists a sequence $\{x_n\}$ such that $J(x_n) \to \inf J$; possibly dropping a finite number of terms, we have $\{x_n\} \subset S_{\tilde{a}}$. By the sequential compactness of the latter set, there exist $x \in S$ and a subsequence $\{x_{n'}\}$ such that $x_{n'} \to x$. As J is sequentially lower semicontinuous, $\liminf J(x_{n'}) \ge J(x)$; therefore $J(x) = \inf J$. The sequential compactness of $\mathcal{M}_J$ may similarly be checked. [(26)] □

Well-Posedness. Many problems may be regarded as a transformation, $\mathcal{T}$, from a space of data, $\mathcal{D}$, to a space of solutions, $\mathcal{S}$. In several cases the solution exists and is unique, i.e., $\mathcal{T}$ is a single-valued mapping $\mathcal{D} \to \mathcal{S}$. If $\mathcal{D}$ and $\mathcal{S}$ are topological spaces, it is also of interest to see whether $\mathcal{T}$ is continuous; in this case the problem is said to be *well-posed in the sense of Hadamard*.

Dealing with minimization problems, another concept of well-posedness is also useful. Let S be a separated topological space. The problem of minimizing a proper functional $J : S \to]-\infty, +\infty]$ is said to be *well-posed in the sense of Tychonov* if and only if any minimizing sequence converges to a minimum point:

$$\forall\{u_n\} \subset S, \ \text{ if } \ J(u_n) \to \inf J \ \text{ then } \ \exists u \in S : u_n \to u, \ J(u) = \inf J. \tag{1.3}$$

It is easy to check that this holds if and only if

$$\begin{aligned} &\text{(i) a minimum point exists, and} \\ &\text{(ii) any minimizing sequence is convergent.} \end{aligned} \tag{1.4}$$

This entails the uniqueness of the minimum point.

For instance, let us consider the following functions $J_i : \mathbf{R} \to]-\infty, +\infty]$

$$J_1(u) := \begin{cases} u^2(u-1)^2 & \text{if } u \neq 1, \\ 1 & \text{if } u = 1, \end{cases} \qquad J_2(u) := u^2 e^{-u}, \qquad J_3(u) := u^2(u-1)^2. \tag{1.5}$$

None of the corresponding minimization problems is well-posed in the sense of Tychonov. Both J_1 and J_2 have one and only one minimum point, $u = 0$. However $\{u_n := 1 + 1/n\}$ is a minimizing sequence for J_1, but it converges to a point which is not of minimum; on the other hand $\{u_n := n\}$ is a minimizing sequence for J_2, but it does not converge.

[(26)] This type of argument based on the use of minimizing sequences is often referred to as the *direct method of the calculus of variations.* Dealing with a differentiable functional J defined on a topological vector space, the *indirect method* consists in studying the minimization problem via the *Euler equation* $J'(u) = 0$.

The direct method might also be used for part (i), just replacing sequences by nets and subsequence by cofinal subnets.

The function J_3 has two minimum points, at variance with J_1 and J_2. Therefore the minimization of J_3 is not well-posed in the sense of Tychonov, although this function looks less pathological than J_1 and J_2. A weaker concept of well-posedness was indeed proposed. A minimization problem is said *well-posed in the generalized sense of Tychonov* if and only if any minimizing sequence has a subsequence that converges to a minimum point:

$$\begin{aligned} &\forall \{u_n\} \subset S, \ \text{ if } \ J(u_n) \to \inf J, \ \text{ then} \\ &\exists u \in S, \exists \{u_{n'}\} \subset \{u_n\} : u_{n'} \to u, \ J(u) = \inf J. \end{aligned} \tag{1.6}$$

This holds if and only if

$$\begin{aligned} &\text{(i) any minimizing sequence has a convergent subsequence, and} \\ &\text{(ii) the limit of any convergent minimizing sequence is a minimum point.} \end{aligned} \tag{1.7}$$

J_3 is well-posed in this generalized sense, whereas J_1 and J_2 are not.

Exercises: (i) Check that (1.3) $\Leftrightarrow$ (1.4), and that (1.6) $\Leftrightarrow$ (1.7).

(ii) Discuss the possible generalized well-posedness of the minimization of the following functions:

$$J_4(u) := \begin{cases} u^2(u-1)^2(u+1)^2 & \text{if } u \neq -1, 0, 1, \\ -1 & \text{if } u = -1 \text{ or } u = 0 \text{ or } u = 1, \end{cases}$$

$$J_5(u) := \begin{cases} u^2 + u & \text{if } u^{-1} \in \mathbf{R} \setminus \mathbf{N}, \\ u^2 & \text{if } u^{-1} \notin \mathbf{N}, \end{cases}$$

$$J_6(u) := \begin{cases} u^2 + 1 & \text{if } u^{-1} \in \mathbf{R} \setminus \mathbf{N}, \\ u^2 & \text{if } u^{-1} \notin \mathbf{N}, \end{cases}$$

$$J_7(u) := \begin{cases} u^2 + 1 & \text{if } u \in \mathbf{R} \setminus \mathbf{Q}, \\ u^2 & \text{if } u \in \mathbf{Q}, \end{cases}$$

$$J_8(u) := \begin{cases} u^2 + 1 & \text{if } u \in \mathbf{Q}, \\ u^2 & \text{if } u \in \mathbf{R} \setminus \mathbf{Q}. \end{cases}$$

III.2. Variational inequalities (Lions-Stampacchia theorem)

In this section we illustrate the notion of variational inequality, and prove a celebrated theorem of Lions-Stampacchia.

Let H be a Hilbert space, with scalar product $(\cdot, \cdot)$ and norm $\| \cdot \|$. Let $f \in H$, set

$$J(v) := \frac{1}{2}\|v\|^2 - (f, v) \qquad \forall v \in H, \tag{2.1}$$

and consider the problem of minimizing J in a nonempty closed convex set $K \subset H$. Note that J is Fréchet-differentiable in H, with differential

$$J'(v) = v - f \qquad \forall v \in K. \tag{2.2}$$

Let us denote by I_K the indicator function of K:

$$I_K(v) := 0 \quad \forall v \in K, \qquad I_K(v) := +\infty \quad \forall v \in H \setminus K. \tag{2.3}$$

Lemma 2.1. *An element $u \in H$ minimizes J in K if and only if it fulfills either of the following variational inequalities*

$$u \in H \quad \text{and} \quad (u - f, u - v) + I_K(u) \le I_K(v) \qquad \forall v \in H, \tag{2.4}$$

$$u \in K \quad \text{and} \quad (u - f, u - v) \le 0 \quad \forall v \in K. \tag{2.5}$$

If K is a linear subspace of H, then (2.5) is equivalent to the variational equation

$$u \in K \quad \text{and} \quad (u - f, v) = 0 \quad \forall v \in K. \tag{2.5'}$$

Proof. As

$$(u, u - v) = \frac{1}{2}\|u\|^2 - \frac{1}{2}\|v\|^2 + \frac{1}{2}\|u - v\|^2 \ge \frac{1}{2}\|u\|^2 - \frac{1}{2}\|v\|^2;$$

if u fulfills (2.4) then it minimizes J. Conversely, if u minimizes J then

$$J(u + \lambda(v - u)) - J(u) \ge 0 \qquad \forall v \in K, \forall \lambda \in]0, 1[.$$

Dividing by λ and passing to the limit as λ vanishes, we get

$$\big(u - f, v - u\big) = \big(J'(u), v - u\big) \ge 0 \qquad \forall v \in K,$$

namely (2.5). The equivalence between (2.4) and (2.5) is easily checked.

Finally, if K is a subspace of H, by selecting $v = u \pm \widetilde{v}$ in (2.5) for any $\widetilde{v} \in H$ (and then dropping the tilde), (2.5') follows. □

Lemma 2.2. *(i) An element $u \in H$ minimizes J in K if and only if u is the projection of f onto K: $u = P_K(f)$.*

(ii) The projection operator $P_K : H \to K$ is nonexpansive, i.e.,

$$\|P_K(f_1) - P_K(f_2)\| \le \|f_1 - f_2\| \qquad \forall f_1, f_2 \in H. \tag{2.6}$$

Proof. (i) As $J(v) = \frac{1}{2}\|v - f\|^2 - \frac{1}{2}\|f\|^2$ for any $v \in V$, part (i) holds.

(ii) Let us write (2.5) for f_1 and $u_1 := P_K(f_1)$ (f_2 and $u_2 := P_K(f_2)$, resp.), take $v = u_2$ ($v = u_1$, resp.), and then sum the two inequalities. This yields

$$\|u_1 - u_2\|^2 \le (f_1 - f_2, u_1 - u_2) \le \|f_1 - f_2\| \, \|u_1 - u_2\|,$$

whence $\|u_1 - u_2\| \le \|f_1 - f_2\|$, namely (2.6). □

More generally, one may consider a linear bounded operator $A : H \to H$, fix any $f \in H$, and minimize the function

$$\widehat{J}(v) := \frac{1}{2}(Av, v) - (f, v) \qquad \forall v \in H. \tag{2.7}$$

This problem is equivalent to the variational inequality

$$u \in K \quad \text{and} \quad (A_s u - f, u - v) \le 0 \quad \forall v \in K, \tag{2.8}$$

where $A_s := (A + A^*)/2$ is the *symmetric part* of A. [27]

Even more generally, one may consider a nonlinear operator $\mathcal{A} : H \to H$ and the variational inequality

$$u \in K \quad \text{and} \quad (\mathcal{A}(u) - f, u - v) \le 0 \quad \forall v \in K. \tag{2.9}$$

In general this variational inequality need not be equivalent to the minimization of any function. This fails even if $\mathcal{A}$ is linear, whenever this operator is not symmetric.

We shall assume that $\mathcal{A}$ is Lipschitz continuous and *strongly monotone* in K, that is,

$$\exists L > 0 : \forall u, v \in K,\ \|\mathcal{A}(u) - \mathcal{A}(v)\| \le L\|u - v\|, \tag{2.10}$$

$$\exists \alpha > 0 : \forall u, v \in K,\ (\mathcal{A}(u) - \mathcal{A}(v), u - v) \ge \alpha\|u - v\|^2. \tag{2.11}$$

Theorem 2.3. *(Lions-Stampacchia, 1965 about) Let K be a nonempty closed convex subset of a real Hilbert space H, and $\mathcal{A} : H \to H$ fulfill (2.10) and (2.11). Then for any $f \in H$ there exists one and only one solution of (2.9), and this depends Lipschitz continuously on f.*

Proof. For any $\rho > 0$, the inequality of (2.9) is equivalent to

$$\big(u - [u - \rho(\mathcal{A}(u) - f)], u - v\big) \le 0 \qquad \forall v \in K,$$

and this is tantamount to $u = P_K(u - \rho(\mathcal{A}(u) - f))$. We claim that for a suitable $\rho > 0$ the mapping $T = T_\rho : v \mapsto v - \rho(\mathcal{A}(v) - f)$ is a (strict) contraction, i.e.,

$$\exists a < 1 : \forall v_1, v_2 \in H \qquad \|T(v_1) - T(v_2)\| \le a\|v_1 - v_2\|.$$

Actually,

$$\begin{aligned} &\|T(v_1) - T(v_2)\|^2 \\ &= \|v_1 - v_2\|^2 + \rho^2\|\mathcal{A}(v_1) - \mathcal{A}(v_2)\|^2 - 2\rho(\mathcal{A}(v_1) - \mathcal{A}(v_2), v_1 - v_2) \\ &\overset{(2.10),(2.11)}{\le} \|v_1 - v_2\|^2 + \rho^2 L^2\|v_1 - v_2\|^2 - 2\rho\alpha\|v_1 - v_2\|^2 \\ &= (1 + \rho^2 L^2 - 2\rho\alpha)\|v_1 - v_2\|^2 \qquad \forall v_1, v_2 \in H. \end{aligned} \tag{2.12}$$

If $0 < \rho < 2\alpha/L^2$ then $1 + \rho^2 L^2 - 2\rho\alpha < 1$, so that T is a contraction; hence $P_K \circ T$ is also a contraction. By a classical fixed-point theorem due to Banach, [28] $P_K \circ T$ then has one and only one fixed point. Therefore the variational inequality (2.9) has one and only one solution.

The Lipschitz continuity of the *solution operator* $f \mapsto u$ is a straightforward consequence of (2.11). □

The latter theorem entails the following classical result.

---

[27] Any linear and bounded operator A acting in a Hilbert space may be written as the sum of its symmetric and anti-symmetric parts: $A = A_s + A_a$, where $A_s := (A + A^*)/2$ and $A_a := (A - A^*)/2$. Here A^* is the adjoint of A (which operates in $H^* = H$), that is $(A^*u, v) = (u, Av)$ for any $u, v \in H$. Note that $(Av, v) = (A_s v, v)$ for any $v \in H$.

[28] This classical result states that any (strict) contraction in a complete metric space has a fixed point, and this is unique.

Corollary 2.4. *(Lax-Milgram) Let H be a real Hilbert space, and $\mathcal{A} : H \to H$ be linear, bounded and strongly monotone. Then for any $f \in H$ there exists one and only one $u \in H$ such that $\mathcal{A}u = f$, and the mapping $f \mapsto u$ is linear and continuous.*

If the operator $\mathcal{A}$ is also symmetric, then the thesis follows from the classical Riesz-Fréchet theorem for the representation of the dual of a Hilbert space.

Generalizations. (i) Let an operator $\mathcal{A} : H \to H$ fulfill (2.10) and (2.11), and K be as above. Generalizing Lemma 2.1, one may easily check that

$$u \in K \quad \text{and} \quad (\mathcal{A}(u) - f, u - v) \leq 0 \quad \forall v \in K \tag{2.13}$$

is equivalent to

$$u \in H \quad \text{and} \quad (\mathcal{A}(u) - f, u - v) + I_K(u) \leq I_K(v) \quad \forall v \in H. \tag{2.14}$$

(ii) The latter existence theorem may be extended as follows.

Theorem 2.5. *Let H be a real Hilbert space, $\mathcal{A} : H \to H$ fulfill (2.10) and (2.11), and $J : H \to]-\infty, +\infty]$ be a convex and lower semicontinuous mapping. Then for any $f \in H$ there exists one and only one $u \in H$ such that*

$$u \in H \quad \text{and} \quad (\mathcal{A}(u) - f, u - v) + J(u) \leq J(v) \quad \forall v \in H, \tag{2.15}$$

and this depends Lipschitz continuously on f.

Theorem 2.3 actually corresponds to the case of $J = I_K$.

Proposition 2.6. *(Minty's trick) Under the above assumptions for the data, the variational inequality (2.9) is equivalent to*

$$u \in K \quad \text{and} \quad (\mathcal{A}(v) - f, u - v) \leq 0 \quad \forall v \in K. \tag{2.16}$$

A similar statement applies to (2.14).

Proof. As $(\mathcal{A}(u) - \mathcal{A}(v), u - v) \geq 0$ for any $u, v \in K$, (2.9) entails (2.16).

In order to prove the converse implication, let us select any $v \in K$, any $t \in]0, 1[$, and set $w = u + t(v - u)$, which is an element of K. By selecting $v = w$ in (2.16), we get

$$t(\mathcal{A}(u + t(v - u)) - f, u - v) = (\mathcal{A}(w) - f, u - w) \leq 0 \qquad \forall t \in]0, 1[.$$

Dividing both members by t, we have

$$(\mathcal{A}(u + t(v - u)) - f, u - v) \leq 0 \qquad \forall t \in]0, 1[.$$

Next let us pass to the limit as t vanishes. By (2.10) $\mathcal{A}(u + t(v - u)) \to \mathcal{A}(u)$ in V', and thus we get $(\mathcal{A}(u) - f, u - v) \leq 0$. □

III.3. The obstacle probem

Description of the phenomenon. Let us consider a (weightless) elastic string that is stretched between two points, that we represent by the pairs $(0,0)$ and $(a,0)$ in the (x,y)-plane. Let us assume that the string is constrained to stay below an obstacle, that we describe as the epigraph of a function

$$\psi : [0,a] \to \mathbf{R}, \qquad \text{with } \ \psi(0), \psi(a) > 0.$$

($\mathrm{epi}(\psi) := \{(x,y) \in X \times \mathbf{R}^2 : \psi(x) \le y\}$.) Representing the equilibrium configuration of the string as the graph of the (unknown) function $u : [0,a] \to \mathbf{R}$, *formally* [29] we thus have

$$u(0) = u(a) = 0, \tag{3.1}$$
$$u(x) \le \psi(x) \qquad \forall x \in [0,a], \tag{3.2}$$
$$u''(x) \ge 0 \qquad \forall x \in [0,a], \tag{3.3}$$
$$u(x) < \psi(x) \ \ \Rightarrow \ \ u''(x) = 0 \qquad \forall x \in [0,a]. \tag{3.4}$$

The latter condition expresses the evidence that the string attains a linear configuration where it does not touch the obstacle. The implication (3.4) may equivalently be replaced by the following *transversality condition:*

$$[u(x) - \psi(x)]\, u''(x) = 0 \qquad \forall x \in [0,a]. \tag{3.5}$$

As for the regularity prescriptions, here it is natural to assume that $\psi \in C^0([0,a])$, and to search for $u \in C^2([0,a])$.

Variational formulation. This problem may also be approached in a different way. The elastic energy of the string reads

$$F(u) := c \int_0^a [u'(x)]^2 \, dx \qquad (c : \text{constant } \ > 0). \tag{3.6}$$

The stationary configuration corresponds to the (unique) u that minimizes this strictly convex functional among the functions that fulfill the constraints (3.1) and (3.2).

For this variational problem it suffices to assume that $u \in C^1([0,a])$; however, for the analysis of the problem it is more appropriate to search for $u \in H^1(0,a)$. [30] Let us define the following nonempty closed convex subset of $H^1(0,a)$

$$K := \big\{ v \in H^1(0,a) : u(0) = u(a) = 0, \ u(x) \le \psi(x) \ \ \forall x \in [0,a] \big\}, \tag{3.7}$$

---

[29] *Formally* means without specifying the regularity assumptions.

A standard way of modelling phenomena in physical and nonphysical sciences proceeds as follows: first one or more equations or a variational formulation are displayed without specifying the functional framework. Afterwards the functional frame is fixed, and a precise analytical problem is formulated. Existence of the solution, possibly its uniqueness as well as other qualitative properties (especially regularity) are then proved. Finally, and effort is made to provide a precise meaning to the model that had initially been formulated.

[30] By $H^1(0,a)$ we denote the real Hilbert space of the absolutely continuous functions $v : [0,a] \to \mathbf{R}$ such that $\|v\|^2 := \int_0^a \{v(x)^2 + v'(x)^2\}\, dx < \infty$.

We denote by $H_0^1(0,a)$ the Hilbert subspace of $H^1(0,a)$ of the functions that vanish at the border. This Hilbert space may also be equipped with the equivalent norm $\|v\|_0 := \big(\int_0^a v'(x)^2 \, dx \big)^{1/2}$. Both $H^1(0,a)$ and $H_0^1(0,a)$ are Sobolev spaces.

and formulate the variational problem

$$\text{find } u \in K \text{ such that} \quad F(u) \le F(v) \ \ \forall v \in K. \tag{3.8}$$

This is obviously equivalent to minimizing $F + I_K$ over the whole $H^1(0,a)$.

For instance, if $\psi(x) = 2|x - a/2| - a/2$ for any $x \in [0,a]$, it is easy to see that F is minimized by $u(x) = |x-a/2| - a/2$ for any $x \in [0,a]$. In this case both ψ and u are elements of $H^1(0,a) \setminus C^1([0,a])$ (as both graphs have a corner).

By Lemma 2.1, (3.8) is equivalent to the following variational inequality

$$u \in K \quad \text{and} \quad (u', u' - v') \le 0 \quad \forall v \in K, \tag{3.9}$$

or equivalently

$$u \in H \quad \text{and} \quad (u', u' - v') + I_K(u) \le I_K(v) \quad \forall v \in H. \tag{3.10}$$

By Theorem 1.1, the variational problem (3.8) has one and only one minimizer. On the other hand, by Theorem 2.3 the equivalent variational inequality (3.9) has one and only one solution.

Exercise. (i) Formulate the obstacle problem in the case in which there are two obstacles. one below and one above the elastic string. Then extend the analysis of this section to this setting.

III.4. De Giorgi's Γ-convergence in metric spaces

Here we confine ourselves to metric spaces, although Γ-convergence may also be defined in the more general framework of topological spaces.

Definitions. Let (X,d) be a metric space, a sequence $\{f_n\}$ and f be functions $X \to [-\infty, +\infty]$, and $u \in X$. We say that f_n Γ-*converges to* f (in (X,d)) at u, and write

$$f(u) = \Gamma\text{-}\lim_{n\to\infty} f_n(u) \quad \text{or} \quad f_n(u) \xrightarrow{\Gamma} f(u),$$

if and only if

$$(i) \ \text{for any sequence } \{u_n\} \text{ in } X, \text{ if } u_n \to u \text{ then } \liminf_{n\to\infty} f_n(u_n) \ge f(u), \tag{4.1}$$

$$\begin{aligned} &\text{there exists a sequence } \{u_n\} \text{ in } X \\ (ii) \ &\text{such that } u_n \to u \text{ and } \limsup_{n\to\infty} f_n(u_n) \le f(u) \\ &\text{(equivalently, } f_n(u_n) \to f(u), \text{ by part (i)).} \end{aligned} \tag{4.2}$$

We say that f_n Γ-converges to f whenever this holds for any $u \in X$.

If in (4.2) we replace $\limsup_{n\to\infty} f_n(u_n)$ by $\liminf_{n\to\infty} f_n(u_n)$, then (i) and (ii) define the *inferior* Γ-*limit*

$$f(u) = \min\Big\{ \liminf_{n\to\infty} f_n(u_n) : u_n \to u \text{ in } X \Big\} =: \Gamma\text{-}\liminf_{n\to\infty} f_n(u) \qquad \forall u \in X.$$

On the other hand, if in (4.1) we replace $\liminf_{n\to\infty} f_n(u_n)$ by $\limsup_{n\to\infty} f_n(u_n)$, then (i) and (ii) define the *superior* Γ-*limit*

$$f(u) = \min\Big\{ \limsup_{n\to\infty} f_n(u_n) : u_n \to u \text{ in } X \Big\} =: \Gamma\text{-}\limsup_{n\to\infty} f_n(u) \qquad \forall u \in X.$$

The two latter limits exist for any sequence $\{f_n\}$, and of course

$$\Gamma\liminf_{n\to\infty} f_n(u) \le \Gamma\limsup_{n\to\infty} f_n(u) \qquad \forall u \in X.$$

Moreover, f_n Γ-converges if and only if the inferior and superior Γ-limits are equal; in this case their common value coincides with the Γ-limit.

There is no symmetry between the inferior and superior Γ-limits; actually, in general

$$\Gamma\liminf_{n\to\infty} (-f_n) \ne -\Gamma\limsup_{n\to\infty} f_n, \tag{4.3}$$

whence, in egeneral,

$$\Gamma\lim_{n\to\infty} (-f_n) \ne -\Gamma\lim_{n\to\infty} f_n, \quad \text{whenever both } \Gamma\text{-limits exist.}$$

The functions $\Gamma\liminf_{n\to\infty} f_n$ and $\Gamma\limsup_{n\to\infty} f_n$ indeed are both lower semicontinuous; the same then applies to the Γ-limit, whenever it exists. [31]

Proposition 4.1. *Let $\{f_n\}$ be a sequence of functions $X \to [-\infty, +\infty]$. Then:*
(i) If $f_n \xrightarrow{\Gamma} f$, then f is lower semicontinuous.
(ii) If $\{f_n\}$ is a nondecreasing sequence of functions $X \to [-\infty, +\infty]$, then

$$f_n \xrightarrow{\Gamma} \sup\{\bar f_n\}. \tag{4.4}$$

(iii) If $\{f_n\}$ is a nonincreasing sequence of functions $X \to [-\infty, +\infty]$, then

$$f_n \xrightarrow{\Gamma} \overline{\inf f_n}. \tag{4.4'}$$

Hence

$$\text{if} \quad f_n = f_0 \ \ \forall n, \quad \text{then} \quad f_n \xrightarrow{\Gamma} \bar f_0. \tag{4.5}$$

Here are some further properties of Γ-convergence.

Proposition 4.2. *(Comparison with the pointwise limit) Let $\{f_n\}$ be a sequence of functions $X \to [-\infty, +\infty]$, $f : X \to [-\infty, +\infty]$, and $f_n \xrightarrow{\Gamma} f$. Then:*

$$\Gamma\text{–}\liminf_{n\to\infty} f_n(u) \le \liminf_{n\to\infty} f_n(u) \quad \forall u \in X, \tag{4.6}$$

$$\Gamma\text{–}\limsup_{n\to\infty} f_n(u) \le \limsup_{n\to\infty} f_n(u) \quad \forall u \in X. \tag{4.7}$$

Equalities hold whenever (denoting by $\mathcal{U}(u)$ the family of neighborhoods of u)

$$\forall \varepsilon > 0, \exists U \in \mathcal{U}(u) : \forall v \in U, f_n(v) \ge f_n(u) - \varepsilon \quad \textit{(equi-lower-semicontinuity).} \tag{4.8}$$

Proposition 4.3. *(Compactness) If (X, d) is a separable metric space (i.e., X has a countable basis of open sets), [32] then any sequence $\{f_n\}$ of functions $X \to [-\infty, +\infty]$ has a Γ-convergent subsequence. [The limit need not be finite.]*

The next result entails a useful characterization of Γ-convergence as a variational convergence: whenever $f_n \xrightarrow{\Gamma} f$, the limit of any convergent sequence of minimizers of f_n minimizes f.

---

[31] For any function $f : X \to [-\infty, +\infty]$ we shall denote its lower semicontinuous regularized function by $\bar f$.

[32] Of course here we refer to the topology that is induced by the metric.

Proposition 4.4. *(Minimization) Let (X,d) be a metric space, and $\{f_n\}$ be a sequence of functions $X \to]-\infty,+\infty]$ such that $f_n \xrightarrow{\Gamma} f$. Assume that $\inf f_n > -\infty$ for any n, and that $\{\bar u_n\} \subset X$ and $\bar u \in X$ are such that*

$$f_n(\bar u_n) \le \inf f_n + \frac{1}{n} \quad \forall n, \tag{4.9}$$

$$\bar u_n \to \bar u \quad \text{in} \quad X. \tag{4.10}$$

Then

$$\inf f_n \to \inf f, \tag{4.11}$$

$$f_n(\bar u_n) \to f(\bar u) = \inf f. \tag{4.12}$$

Proof. (i) We have

$$f(\bar u) \overset{(4.1),(4.10)}{\le} \liminf_{n\to\infty} f_n(\bar u_n) \overset{(4.9)}{\le} \liminf_{n\to\infty} \inf f_n.$$

(ii) By (4.2), for any $v \in X$ there exists $\{v_n\} \subset X$ such that $v_n \to v$ in X and $f_n(v_n) \to f(v)$. By (4.9) then $f_n(\bar u_n) \le f_n(v_n) + \frac{1}{n}$ for any n. Therefore

$$f(\bar u) \overset{(4.1),(4.10)}{\le} \liminf_{n\to\infty} f_n(\bar u_n) \le \lim_{n\to\infty} f_n(v_n) = f(v) \qquad \forall v \in X.$$

Thus $f(\bar u) = \inf f$.

(iii) On the other hand, by (4.2) there exists $\{u_n\} \subset X$ such that $u_n \to \bar u$ in X and $f(\bar u) \ge \limsup_{n\to\infty} f_n(u_n)$. As obviously $f_n(u_n) \ge \inf f_n$ for any n, we then have

$$f(\bar u) \ge \limsup_{n\to\infty} f_n(u_n) \ge \limsup_{n\to\infty} \inf f_n.$$

Therefore $f(\bar u) = \inf f = \lim_{n\to\infty} \inf f_n$ and $f_n(\bar u_n) \to f(\bar u)$. □

Examples. Here X coincides with $\mathbf{R}$, equipped with the Euclidean metric.

(i) Let us set $f_n(x) := (-1)^n x$ for any $x \in \mathbf{R}$ and any n. Then

$$\Gamma\text{–}\liminf_{n\to\infty} f_n(x) = -|x|, \qquad \Gamma\text{–}\limsup_{n\to\infty} f_n(x) = |x| \qquad \forall x \in \mathbf{R}.$$

(ii) Let us set $f_n(x) := \cos(nx)$ for any $x \in [-\pi,\pi]$ and any n. Then $f_n \xrightarrow{\Gamma} -1$, although $f_n \to 0$ weakly in $L^1(-\pi,\pi)$, and the pointwise limit exists only for $x = 0$. Note that

$$(-1 =)\ \Gamma\text{–}\lim_{n\to\infty} f_n = \Gamma\text{–}\lim_{n\to\infty} (-f_n) \ne -\Gamma\text{–}\lim_{n\to\infty} f_n\ (= 1).$$

(iii) Let us set $f_n(x) := \sin(2^n x)$ for any $x \in [-\pi,\pi]$ and any n. Then $f_n \xrightarrow{\Gamma} -1$, although $f_n \to 0$ weakly in $L^1(-\pi,\pi)$, and $f_n(x) = 0$ for any x of the form $x = 2^{-n} m\pi$ for some $m \in \mathbf{N}$ (the set of the zeros of f_n thus tends to the whole interval $[-\pi,\pi]$).

(iv) Let us set $f_n(x) := x\cos(nx)$ and $f(x) := -|x|$ for any $x \in [-\pi,\pi]$ and any n. Then $f_n \xrightarrow{\Gamma} f$. Notice that f is even, although any f_n is odd.

(v) Let us set $f_n(x) := nx\exp(nx)$ for any $x \in \mathbf{R}$ and any n. Then $f_n \xrightarrow{\Gamma} f$, where

$$f(x) := \begin{cases} 0 & \text{if } x < 0, \\ -1/e & \text{if } x = 0, \\ +\infty & \text{if } x > 0, \end{cases}$$

although $f_n(0) \to 0$.

(vi) Let $\{q_n\}$ be an enumeration of $\mathbf{Q}$, and set

$$\begin{cases} f_n(x) := 0 & \text{if } \exists m \geq n \text{ such that } x = q_m \\ f_n(x) := 1 & \text{otherwise} \end{cases} \qquad \forall x \in \mathbf{R}.$$

Then $f_n(x) \nearrow 1$ for any $x \in \mathbf{R}$, but $f_n \xrightarrow{\Gamma} 0$, by part (ii) of Proposition 4.1.

The notion of Γ-convergence may be extended to topological spaces.

Convergence in the Sense of Kuratowski.

Definition. Let (X, d) be a metric space, $\{A_n\}$ be a sequence of subsets of X, and $A \subset X$. We say that A_n *converges to* A *in the sense of Kuratowski* if and only if:

(i) for any $u \in X$, for any sequence $\{u_n \in A_n\}$, if $u_n \to u$, then $u \in A$;

(ii) for any $u \in A$, there exists a sequence $\{u_n \in A_n\}$ such that $u_n \to u$.

For any function $X \to [-\infty, +\infty]$ we define its *epigraph:*

$$\text{epi}(f) := \{(u, z) \in X \times \mathbf{R} : f(u) \leq z\}.$$

On the basis of the next result, the Γ-convergence has also been named *epi-convergence.*

Proposition 4.5. *Let (X, d) be a metric space, and $\{f_n\}$ be a sequence of functions $X \to [-\infty, +\infty]$. Then $f_n \xrightarrow{\Gamma} f$ if and only if* $\text{epi}(f_n) \to \text{epi}(f)$ *in the sense of Kuratowski in $X \times \mathbf{R}$. Moreover, for any sequence of subsets $\{A_n\}$ of X, $I_{A_n} \xrightarrow{\Gamma} I_A$ if and only if $A_n \to A$ in the sense of Kuratowski in X.*

III.5. Ekeland's Minimization Principle

Theorem 5.1. *(Ekeland) Let (E, d) be a complete metric space, $F : E \to]-\infty, +\infty]$ be a lower semicontinuous mapping such that* $\inf F$ *is finite. Let $u_0 \in \text{Dom}(F)$, and c be a positive constant. Then there exists $\tilde{u}$ such that*

$$\begin{cases} F(\tilde{u}) + cd(\tilde{u}, u_0) \leq F(u_0), \\ F(\tilde{u}) < F(u) + cd(\tilde{u}, u) \qquad \forall u \in E \setminus \{\tilde{u}\}. \end{cases} \tag{5.1}$$

($\tilde{u} \in E$ need not be unique, and might coincide with u_0.)

*** Proof.** [32] For any $(u_i, a_i) \in E \times \mathbf{R}$ $(i = 1, 2)$, let us define the relation

$$(u_1, a_1) \preceq (u_2, a_2) \quad \Leftrightarrow \quad a_1 + cd(u_1, u_2) \leq a_2.$$

It is straightforward to check that this is an ordering. Let us set

$$G := \text{epi}(F) \cap \{(u, a) \in E \times \mathbf{R} : (u, a) \preceq (u_0, F(u_0))\}.$$

[32] Ahead we derive this result from a theorem of Brezis and Browder.

We claim that the ordering $\preceq$ is inferiorly inductive in G (see below for the definition). Let $\{(u_i, a_i)\}_{i\in I}$ be a totally ordered subset of G. The corresponding family $\{a_i\}_{i\in I}$ is nonincreasing and bounded, hence it converges to some $\bar a \in \mathbf{R}$. Moreover, as $\{(u_i, a_i)\}_{i\in I}$ is totally ordered,

$$cd(u_i, u_j) \le |a_i - a_j| \qquad \forall i, j \in I.$$

Hence $\{u_i\}_{i\in I}$ is a Cauchy net;[33] by the completeness of E, it then converges to some $\bar u \in E$. As F is lower semicontinuous epi(F) is closed; hence $(\bar u, \bar a) \in \mathrm{epi}(F)$. As $a_i + cd(u_i, u_j) \le a_j$ whenever $(u_i, a_i) \preceq (u_j, a_j)$, it follows that $\bar a + cd(\bar u, u_j) \le a_j$, that is $(\bar u, \bar a) \preceq (u_j, a_j)$, for any $j \in I$. Thus $(\bar u, \bar a)$ is a lower bound of $\{(u_i, a_i)\}_{i\in I}$.

By Zorn's lemma (see below), there exists then a minimal element $(\tilde u, \tilde a) \in G$. Note that $(\tilde u, \tilde a)$ necessarily stays on the graph of F, to wit $\tilde a = F(\tilde u)$. The inequality $(5.1)_1$ is thus fulfilled. The minimality means that

$$\forall (u, a) \in G, \qquad a + cd(\tilde u, u) \le F(\tilde u) \;\;\Rightarrow\;\; u = \tilde u, \;\; a = F(u).$$

Thus $(5.1)_2$ holds for any $(u, F(u)) \in G \setminus \{(\tilde u, F(\tilde u))\}$, hence for any $(u, F(u)) \in \mathrm{epi}(F) \setminus \{(\tilde u, F(\tilde u))\}$, that is, for any $u \in E \setminus \{\tilde u\}$. □

Remarks. (i) A remarkable aspect of Theorem 5.1 is that, at variance with Theorem III.1.1 (which however provides a stronger thesis), it does not assume any compactness property, and holds under the only hypothesis that E is complete.

***** (ii) If E were a reflexive Banach space and F were weakly lower semicontinuous, the function $J : E \to]-\infty, +\infty] : u \mapsto F(u) + c\|u - u_0\|$ would then be coercive and weakly lower semicontinuous. [34] The existence of an (absolute) minimizer $\tilde u$ of J would then follow from Theorem III.1.1; this $\tilde u$ would then fulfill (5.1). (However the weak topology is not metrizable, and the absolute minimization of J is stronger than (5.1).) □

Corollary 5.2. *Let B be a Banach space, with norm $\|\cdot\|$ and dual norm $\|\cdot\|_*$. Let $F : E \to]-\infty, +\infty]$ be lower semicontinuous and Gâteaux differentiable,* [35] *$u_0 \in \mathrm{Dom}(F)$, and $\varepsilon > 0$ be such that*

$$F(u_0) \le \inf F + \varepsilon. \tag{5.2}$$

Then there exists $\tilde u \in E$ such that

$$\begin{cases} F(\tilde u) \le F(u_0), \\ \|\tilde u - u_0\| \le \sqrt{\varepsilon}, \\ \|F'(\tilde u)\|_* \le \sqrt{\varepsilon}. \end{cases} \tag{5.3}$$

Proof. Let us apply Theorem 5.1 with $c := \sqrt{\varepsilon}$. $(5.1)_1$ yields $(5.3)_1$ and $(5.3)_2$, by (5.2).

By $(5.1)_2$, for any $v \in B$ an $\lambda > 0$ we have $F(\tilde u) - F(\tilde u + \lambda\sqrt{\varepsilon}v) \le \lambda\varepsilon\|v\|$. Dividing both sides by λ and letting it vanish, we get $-\langle F'(\tilde u), \sqrt{\varepsilon}v\rangle \le \varepsilon\|v\|$. As this also holds for $-v$, we get $\sqrt{\varepsilon}\|F'(\tilde u)\|\|v\| \le \varepsilon\|v\|$, whence $(5.3)_3$ follows. □

***** In the study of stationary points, compactness is often surrogated by the following *Palais-Smale condition*

$$\begin{aligned} &\text{for any bounded sequence } \{u_n\} \subset B, \text{ if } \; F'(u_n) \to 0 \text{ in } B', \\ &\text{then } \{u_n\} \text{ has a convergent subsequence.} \end{aligned} \tag{5.4}$$

[33] Here we must use the language of nets, since the index set I need not be countable. A reader might however think that $\{(u_i, a_i)\}_{i\in I}$ is just a sequence, without missing the flavor of the argument...

[34] This stems from a classical result of Functional Analysis.

[35] This holds e.g. if F is Fréchet differentiable.

*** Corollary 5.3.** *Let B be a Banach space, $F : B \to]-\infty, +\infty]$ be bounded from below, coercive (cf. (1.2)), lower semicontinuous and Gâteaux differentiable. If F fulfills (5.4), then it has a minimizer.*

Proof. By Corollary 5.2 there exists a minimizing sequence which fulfils the assumption of (5.4); hence it has a convergent subsequence. By the lower semicontinuity of F, its limit is a minimizer. □

The Ekeland principle has a number of interesting consequences. Here is an example.

Corollary 5.4. *(Caristi) Let (E, d) be a complete metric space and $f : E \to E$. Assume that there exists a lower semicontinuous function $F : E \to \mathbf{R}$ such that $\inf F > -\infty$ and*

$$d(x, f(x)) \le F(x) - F(f(x)) \qquad \forall x \in E. \tag{5.5}$$

Then f has a fixed point.

Proof. By Theorem 2.1, there exists $\tilde{u} \in E$ such that $F(\tilde{u}) - F(v) < (1/2)d(\tilde{u}, v)$ for any $v \in E$. By taking $v = f(\tilde{u})$ we get $F(\tilde{u}) - F(f(\tilde{u})) < (1/2)d(\tilde{u}, f(\tilde{u}))$. The hypothesis then yields $d(\tilde{u}, f(\tilde{u})) < (1/2)d(\tilde{u}, f(\tilde{u}))$, whence $f(\tilde{u}) = \tilde{u}$. □

The next theorem may be regarded as an abstract formulation of the Ekeland theorem in ordered sets.

*** Theorem 5.5.** *(Brezis-Browder) Let $(E, \preceq)$ be an ordered set and $F : E \to [-\infty, +\infty[$. Assume that:*

(i) any monotone decreasing sequence in E has a lower bound (countable inductivity);

(ii) F is bounded below;

(iii) F is isotone, *i.e., $F(u_1) \le F(u_2)$ whenever $u_1 \preceq u_2$.*

Then there exists $\tilde{u} \in E$ such that, for any $u \in E$, if $u \preceq \tilde{u}$ then $F(u) = F(\tilde{u})$.

Proof. Let us fix any $u_0 \in E$, and inductively construct a sequence $\{u_n\}$ as follows. Let u_n be given and set

$$M_n := \{v : v \preceq u_n\}, \qquad s_n := \inf_{M_n} F.$$

Let u_{n+1} be such that

$$u_{n+1} \preceq u_n, \qquad F(u_{n+1}) - s_n \le \frac{1}{2}[F(u_n) - s_n].$$

By (i) there exists a lower bound $\tilde{u}$ for $\{u_n\}$. By (ii) the sequence $\{s_n\}$ converges, and by (iii) $s_n \to F(\tilde{u})$.

Let $u \in E$ be such that $u \preceq \tilde{u}$. By (iii), $F(u) \le F(\tilde{u})$. On the other hand, for any n, $u \in M_n$ hence $s_n \le F(u)$; therefore $F(\tilde{u}) \le F(u)$, and we conclude that $F(u) = F(\tilde{u})$. □

The latter result entails the Ekeland theorem. This can be checked by setting

$$v \preceq u \quad \Leftrightarrow \quad F(v) + cd(v, u) \le F(u) \quad \forall u, v \in E.$$

*** Ordered Sets.** We review some notions about ordered sets and the Zorn lemma, that plays a role in analysis.

Let S be an ordered set (namely, a set equipped with a reflexive, antisymmetric and transitive relation). The order is said *total* if and only if, for any $x, y \in S$, either $x \leq y$ or $y \leq x$. A totally ordered subset is also called a *chain.*

Let $(S, \leq)$ be an ordered set, $A \subset S$ and $A \neq \emptyset$.

$x \in A$ is called a *maximal (minimal,* resp.) element of A if and only if $y \in A$ and $x \leq y$ ($y \leq x$, resp.) entail $x = y$.

$x \in A$ is called the *maximum (minimum,* resp.) of A if and only if $y \leq x$ ($x \leq y$, resp.) for any $y \in A$.

$x \in S$ is called an *upper (lower,* resp.) *bound* of A if and only if $y \leq x$ ($y \leq x$, resp.) for any $y \in A$.

$x \in S$ is called the *supremum* or *least upper bound (infimum* or *greatest lower bound,* resp.) of A if and only if it is the minimum (maximum, resp.) of the set of upper (lower, resp.) bounds of A; it will be denoted by $\sup A$ ($\inf A$, resp.).

S is said superiorly (inferiorly, resp.) *inductive* if and only if any totally ordered (nonempty) subset has an upper (lower, resp.) bound.

Theorem. *(Zorn's Lemma) Any nonempty superiorly (inferiorly, resp.) inductively ordered set has a maximal (minimal, resp.) element.*

This classical result is equivalent to Zermelo's *axiom of choice:* the Cartesian product of any nonempty family of nonempty sets is nonempty.

IV. OPTIMAL CONTROL

IV.1 Control problems

Loosely speaking, the control of processes consists in steering a state function $x = x(t)$ ($\in \mathbf{R}^n$) so as to achieve a certain goal, by using a control function $u = u(t)$ ($\in \mathbf{R}^m$). These problems are of paramount importance for economics and technology.

Let us fix a vector $x^0 \in \mathbf{R}^n$ and a (nonempty) set $\mathcal{U}_a$ of functions taking values in $\mathbf{R}^m$ ($\mathcal{U}_a$ might consist e.g. of the functions of t that range in some subset of $\mathbf{R}^m$), and a continuous function

$$f : \mathbf{R}^+ \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n,$$

that is uniformly Lipschitz-continuous w.r.t. the second argument. Let us then assume that the state variable depends on the control via a Cauchy problem of the form

$$\begin{cases} x'(t) = f(t, x(t), u(t)) & \forall t > 0 \\ x(0) = x^0. \end{cases} \tag{1.1}$$

This problem is well-posed, and we denote its solution by $x = x(t, x^0, u(\cdot))$.

Controllability. For any $\widetilde{t} > 0$, let us prescribe a (nonempty) target set $\mathcal{T}(\widetilde{t}) \subset \mathbf{R}^n$; $\mathcal{T}(\widetilde{t}) = \{0\}$ is a typical example. Let us assume that there exist at least an initial state x^0 and a control function $u \in \mathcal{U}_a$ such that $x(\widetilde{t}, x^0, u(\cdot)) \in \mathcal{T}(\widetilde{t})$. We shall denote by $\mathcal{C}(\widetilde{t})$ the set of these initial values, namely, the set of the states x^0 that can be steered to the target $\mathcal{T}(\widetilde{t})$ at time $\widetilde{t}$. In this framework the question arises of *synthesizing the control,* namely, of devising a control that drives the state x to the target.

We shall name *controllable set* the union of these sets as $\widetilde{t}$ varies, and denote it by $\mathcal{C}$; to wit $\mathcal{C} := \bigcup_{t>0} \mathcal{C}(t)$. A system is said *completely controllable* whenever $\mathcal{C} = \mathbf{R}^n$. The main questions concerning controllability consist in describing $\mathcal{C}$ and in studying how it depends on $\mathcal{U}_a$.

Let us fix any $t_1 > 0$ and set

$$z(t) := x(t_1 - t), \qquad w(t) := u(t_1 - t) \qquad \forall t \in [0, t_1]. \tag{1.2}$$

It is straightforward to see that the function x fulfills (1.1) in $[0, t_1]$ if and only if z fulfills the *time-reversed system*

$$\begin{cases} z'(t) = -f(t_1 - t, z(t), w(t)) & \forall t \in]0, t_1[, \\ z(0) = x(t_1), \qquad z(t_1) = x^0. \end{cases} \tag{1.3}$$

We thus conclude that x can be steered from x^0 to $x(t_1)$ by the dynamics $(1.1)_1$ if and only if $x^0 \in \mathcal{C}(t_1)$, that is, if and only if z can be steered from $x(t_1)$ to x^0 by the reversed dynamics $(1.3)_1$.

Bang-bang principle. Let us next assume that the *state equation* (1.1) reads

$$\begin{cases} x'(t) = Ax(t) + Bu(t) & \forall t > 0 \\ x(0) = x^0, \end{cases} \tag{1.4}$$

$A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times m}$ being prescribed constant matrices. The solution is then given by the classical *formula of variation of the parameters:*

$$x(t) = e^{At}x^0 + \int_0^t e^{A(t-s)}Bu(s)\,ds \qquad \forall t > 0. \tag{1.5}$$

Proposition 1.1. *([Ev], [MaSt]) Let the state equation be as in (1.4), $\mathcal{T} = \{0\}$, and assume that the controls are confined to a compact set $\mathcal{U}_a \subset \mathbf{R}^m$. Then the set of controllable states $\mathcal{C}$ is also compact (and convex).*

Moreover, if $\mathcal{U}_a$ and $\mathcal{U}_a^*$ are two compact subsets of $\mathbf{R}^m$ that have the same convex hull, then they correspond to the same controllable states. [(35)]

In this case the set of controllable states thus only depends on the convex hull of the control set. In particular this holds if $\mathcal{U}_a^*$ consists of the extremal points of $\mathcal{U}_a$. [(35)] For instance if $\mathcal{U}_a$ consists of the functions $u = u(t)$ that take values in a closed convex polygon, then the extremal points are the functions that take values at the corners of that polygon. In this case one speaks of *bang-bang* controls, because these controls may just jump from one corner to another. Among other things, this entails that it would not be reasonable to exclude discontinuous controls.

General control problems. Let F be a functional of the form

$$F(x,u) := \int_0^T g(t,x(t),u(t))\,dt + G(T,x(T)) \qquad \forall u \in \mathcal{U}, \forall x \in C^0([0,T]); \tag{1.6}$$

here we do not display the regularity assumptions of g and G.

The class of control problems in *Bolza form* consists in searching for a control $\bar{u} \in \mathcal{U}_a$ such that, defining the state $x = x(u)$ by an equation of the form (1.1),

$$J(\bar{u}) = \inf_{\mathcal{U}_a} J \qquad (J(u) := F(x(u),u) \text{ for any } u \in \mathcal{U}_a). \tag{1.7}$$

The functions g and G are respectively named *running cost* and *final cost.* If $G \equiv 0$ ($g \equiv 0$, respect.) this problem is said in *Lagrange form (Mayer form,* respect.) Via a suitable transformation (that we do not display),

$$\textit{any Bolza problem may be reduced to the Mayer form.} \tag{1.8}$$

If $G \equiv 0$ and $f(t,x,u) = u$ for any (t,x,u), then the state equation (1.1) reads $x'(t) = u(t)$, so that the cost functional reads $\int_0^T g(t,x(t),x'(t))\,dt$. In this case the control problem

---

[(35)] In passing we remind the reader that by the classical Krein-Milman theorem: any nonempty compact convex subset of a Banach space is the closed convex hull of its extremal points. The meaning of this result in $\mathbf{R}^M$ may easily be checked.

[(35)] A point ξ of a closed set $S \subset \mathbf{R}^M$ is an *extremal point* of S whenever

$$\forall \xi', \xi'' \in A, \forall \lambda \in]0,1[, \quad \xi = \lambda\xi' + (1-\lambda)\xi'' \;\Rightarrow\; \xi = \xi' = \xi'',$$

to wit ξ cannot be represented as a proper convex combination of points of S itself.

(This notion of extremality is completely unrelated from that of extremality of functionals, that we introduced in Part I of these note.)

is reduced to the minimization of this Lagrange functional. We may thus conclude that, from this point of view,

the control theory extends the basic Bolza problem of calculus of variations. (1.9)

(The calculus of variations in turn extends the study of stationary points of functions $\mathbf{R}^N \to \mathbf{R}$.)

IV.2. Linear optimal control problems (Lions's theory)

In this section we derive the optimality conditions for the optimal control of a linear system with a quadratic cost, that we formulate in Hilbert spaces.

Formulation of the problem. Let V and H be two real Hilbert spaces such that, identifying H with its dual H', [36]

$$V \subset H = H' \subset V' \quad \text{with dense and continuous injection.} \tag{2.1}$$

Let $f \in V'$, and A be a (possibly noninvertible) linear and continuous operator $V \to V'$. Let $\mathcal{U}$ be a real Hilbert space (the *space of controls)*, and

$$B : \mathcal{U} \to V' \quad \text{be linear and continuous.} \tag{2.2}$$

To any *control* $u \in \mathcal{U}$, let us associate a *state function* $y = y(u)$, which is a solution of the *state equation*

$$y \in V, \qquad Ay = f + Bu \qquad \forall u \in \mathcal{U}. \tag{2.3}$$

Let $\mathcal{H}$ be another real Hilbert space (the *space of observations)*,

$$C : V \to \mathcal{H} \quad \text{be linear and continuous;} \tag{2.3'}$$

to any state y let us then associate the *observation* $z := Cy$, so that

$$z(u) := Cy(u) \qquad \forall u \in \mathcal{U}. \tag{2.4}$$

(If one may directly observe the state, then $\mathcal{H} = V$ and C is the identity operator.)

Let $N : \mathcal{U} \to \mathcal{U}$ be another linear and continuous operator, such that

$$\exists \nu > 0 : \forall v \in \mathcal{U} \qquad (Nv, v) \geq \nu \|v\|^2, \tag{2.5}$$

and fix a

$$\text{closed, convex nonempty set } \ \mathcal{U}_a \subset \mathcal{U} \tag{2.6}$$

(the index "a" stays for "admissible"). Let us also fix any $z_d \in \mathcal{H}$ (the index "d" stays for "desired"), and define the *cost functional* as a cost of the observation plus a cost of the control:

$$J(v) := \|Cy(v) - z_d\|^2_{\mathcal{H}} + (Nv, v)_{\mathcal{U}} \qquad \forall v \in \mathcal{U}. \tag{2.7}$$

The optimal control problem reads

$$\text{find } \ u \in \mathcal{U}_a \ \text{ such that } \ J(u) = \inf_{\mathcal{U}_a} J; \tag{2.8}$$

[36] A similar setting often occurs in the analysis of PDEs. Notice that, having identified the dual space H' with H, we cannot identify V' with V !

any minimizer u of J is called an *optimal control.*

It is easy to see that the functional J is coercive, lower semicontinuous, and strictly convex. By Tonelli's Theorem III.1.1, we then infer that the control problem (2.8) has one and only one solution.

Characterization of the optimal control. In applications it is important to have a recipe to evaluate an optimal control; here we characterize optimal controls by a system of equations and variational inequalities *(optimality system).*

By a procedure analogous to that of Lemma III.1.2, it is easy to see that the problem (2.8) is equivalent to

$$\langle J'(u), v-u\rangle \geq 0 \qquad \forall v \in \mathcal{U}_a. \tag{2.9}$$

(In general a first-order condition like this is just necessary and need not be sufficient. Here it is also sufficient because the functional J is convex.) It is even easier to check that (2.8) is equivalent to the following *optimality condition:*

$$(Cy(u)-z_d, C[y(v)-y(u)])_{\mathcal{H}} + (Nu, v-u)_{\mathcal{U}} \geq 0 \qquad \forall v \in \mathcal{U}_a, \tag{2.9'}$$

which is thus a necessary and sufficient condition of optimality. We want to reformulate it reducing the two addenda to a unique scalar product.

Let us first denote by Λ_V the (Riesz) canonical isomorphism $V \to V'$: [37]

$$(u,v)_V = {}_{V'}\langle \Lambda_V u, v\rangle_V \qquad \forall u, v \in V.$$

Let us define $\Lambda_{\mathcal{U}}$ and $\Lambda_{\mathcal{H}}$ similarly, and denote by $C^* : \mathcal{H}' \to V'$ the Hilbert adjoint of $C : V \to \mathcal{H}$ (see the note at the end of this section). We have

$$\begin{aligned}
&(Cy(u)-z_d, C[y(v)-y(u)])_{\mathcal{H}} \\
&= {}_{\mathcal{H}'}\langle \Lambda_{\mathcal{H}}[Cy(u)-z_d], C[y(v)-y(u)]\rangle_{\mathcal{H}} \\
&= {}_{V'}\langle C^*\Lambda_{\mathcal{H}}[Cy(u)-z_d], y(v)-y(u)\rangle_V \\
&= (\Lambda_V^{-1}C^*\Lambda_{\mathcal{H}}[Cy(u)-z_d], y(v)-y(u))_V \qquad \forall v \in \mathcal{U}_a.
\end{aligned} \tag{2.10}$$

Let us distinguish two cases:

(i) A is invertible. In this case by the state equation (2.3) also reads $y(u) = A^{-1}(f+Bu)$, whence

$$y(v)-y(u) = A^{-1}B(v-u) \qquad \forall v \in \mathcal{U}_a. \tag{2.11}$$

Notice that $(CA^{-1}B)^* : \mathcal{H}' \to \mathcal{U}'$ is the adjoint of $CA^{-1}B : \mathcal{U} \to \mathcal{H}$, and $\Lambda_{\mathcal{U}}^{-1}(CA^{-1}B)^* : \mathcal{H}' \to \mathcal{U}$. We then have

$$\begin{aligned}
&(Cy(u)-z_d, C[y(v)-y(u)])_{\mathcal{H}} \\
&\overset{(2.11)}{=} (Cy(u)-z_d, CA^{-1}B(v-u))_{\mathcal{H}} \\
&= (\Lambda_{\mathcal{U}}^{-1}(CA^{-1}B)^*[Cy(u)-z_d], v-u)_{\mathcal{U}} \qquad \forall v \in \mathcal{U}_a.
\end{aligned} \tag{2.12}$$

By (2.10) and (2.12), the optimality condition (2.9') is equivalent to

$$(\Lambda_{\mathcal{U}}^{-1}(CA^{-1}B)^*[Cy(u)-z_d] + Nu, v-u)_{\mathcal{U}} \geq 0 \qquad \forall v \in \mathcal{U}_a. \tag{2.13}$$

[37] By ${}_{V'}\langle\cdot,\cdot\rangle_V$ obviously we denote the duality pairing between V' and V.

This reformulation of (2.10) is not very convenient, since it involves the inverse operator A^{-1}. Actually, even for a finite-dimensional linear system $Lw = g$ (L being an invertible square matrix), the numerical solution is never computed by inverting L. It is more reasonable to proceed along a different line, that however requires the solution of a further equation — the adjoint-state equation, that we are going to introduce. In this way we shall also be able to avoid the assumption of invertibility of the operator A.

(ii) Use of the adjoint state. Let us denote by $A^* : V \to V'$ the adjoint operator of $A : V \to V'$, and define the *adjoint-state variable* (also called *costate variable)* $p(u)$ as the solution of the *adjoint-state equation* [38]

$$p(u) \in V, \qquad A^*p(u) = C^*\Lambda_{\mathcal{H}}[Cy(u) - z_d] \ (\in V') \tag{2.14}$$

We get

$$\begin{aligned} &{}_{V'}\langle C^*\Lambda_{\mathcal{H}}[Cy(u) - z_d], y(v) - y(u)\rangle_V = {}_{V'}\langle A^*p(u), y(v) - y(u)\rangle_V \\ &= {}_V\langle p(u), A[y(v) - y(u)]\rangle_{V'} \overset{(2.3)}{=} {}_V\langle p(u), B(v - u)\rangle_{V'} \\ &= {}_{\mathcal{U}'}\langle B^*p(u), v - u\rangle_{\mathcal{U}} = (\Lambda_{\mathcal{U}}^{-1}B^*p(u), v - u)_{\mathcal{U}}. \end{aligned} \tag{2.15}$$

The optimality condition (2.9') then also reads

$$(\Lambda_{\mathcal{U}}^{-1}B^*p(u) + Nu, v - u)_{\mathcal{U}} \geq 0 \qquad \forall v \in \mathcal{U}_a. \tag{2.16}$$

We have thus proved the next statement.

Theorem 2.1. *(Necessary and sufficient optimality conditions) $u \in \mathcal{U}_a$ is an optimal control if and only the triplet $(u, p, y) \in \mathcal{U}_a \times V \times V$ fulfills the following* optimality system:

$$\begin{cases} Ay = f + Bu \quad \textit{(state equation: } \mathcal{U} \to V : u \mapsto y\textit{)} \\ A^*p = C^*\Lambda_{\mathcal{H}}[Cy - z_d] \quad \textit{(adjoint-state equation: } V \to V : y \mapsto p\textit{)} \\ (\Lambda_{\mathcal{U}}^{-1}B^*p + Nu, v - u)_{\mathcal{U}} \geq 0 \quad \forall v \in \mathcal{U}_a \quad \textit{(optimality condition).} \end{cases} \tag{2.17}$$

The equations $(2.17)_1$ and $(2.17)_2$ define a mapping $\mathcal{U} \to V : u \mapsto p = p(u)$; $(2.17)_3$ is thus a variational inequality for the unknown function u.

Setting $F_{p,u}(v) := (\Lambda_{\mathcal{U}}^{-1}B^*p + Nu, v)_{\mathcal{U}}$ for any $v \in \mathcal{U}$, the variational inequality $(2.17)_3$ is obviously equivalent to the following *minimum principle*

$$u \in \mathcal{U}_a, \qquad F_{p,u}(u) \leq F_{p,u}(v) \qquad \forall v \in \mathcal{U}_a. \tag{2.18}$$

*** Note on adjoint operators.** If B_1, B_2 are two Banach spaces and $L : B_1 \to B_2$ is a linear and continuous operator, then the *Banach adjoint* $L^* : B_2' \to B_1'$ is the linear and continuous operator defined as follows:

$${}_{B_1}\langle L^*f, v\rangle_{B_1} = {}_{B_2'}\langle f, Lv\rangle_{B_2} \qquad \forall v \in B_1, \forall f \in B_2'. \tag{2.19}$$

There is another notion of adjoint operator that applies to Hilbert spaces. If H_1, H_2 are two Hilbert spaces and $L : H_1 \to H_2$ is linear and continuous, then the *Hilbert adjoint* $L' : H_2 \to H_1$ is defined as follows:

$$(L'w, v)_{H_1} = (w, Lv)_{H_2} \qquad \forall v \in H_1, \forall w \in H_2. \tag{2.20}$$

[38] One writes $p = p(u)$ since p depends on y which in turn depends on u. One might thus also write $p = p(y(u))$ (but nobody write so).

To Hilbert spaces one may apply both notions of adjoint, and these are mutually related by the following identities

$$L' = \Lambda_{H_1}^{-1} L^* \Lambda_{H_2} \qquad \text{or equivalently} \qquad L^* = \Lambda_{H_1} L' \Lambda_{H_2}^{-1}. \tag{2.21}$$

This theory may be extended to noncontinuous (i.e., unbounded) operators.

IV.3 Dynamic programming and time-discrete Bellman's theory

In this section we illustrate the basis of the Bellman theory for the optimal control of time-discrete processes. Hereafter we deal with payoff maximization instead of cost minimization, as it is in the style of economists; the two problems may of course be addressed symmetrically.

The Bellman function. Let us consider a process that occurs through a finite number of stages, that we label by $j = 1, ..., n$. Let us assume that at the jth stage the state x_j ($\in \mathbf{R}^n$) and the control u_j ($\in U_j \subset \mathbf{R}^m$) determine the state

$$x_{j+1} = \varphi_j(x_j, u_j) \quad \text{with the payoff} \quad k_j(x_j, u_j)\ (\in \mathbf{R}) \qquad j = 1, ..., n, \tag{3.1}$$

for prescribed functions $\varphi_1, ..., \varphi_n$ and $k_1, ..., k_n$. If x_r is the state at the stage r, then the total payoff of the subprocess between the stages r and n under the controls $u_r, ..., u_n$ is

$$\begin{aligned} &K_r(x_r; u_r, ..., u_n) = \sum_{j=r}^{n} k_j(x_j, u_j) \qquad \text{for} \quad j = r, ..., n, \\ &\text{with} \quad x_{j+1} = \varphi_{j+1}(x_j, u_j) \quad \text{recursively for} \quad j = r, ..., n-1. \end{aligned} \tag{3.2}$$

Let us define the *Bellman function* S: [39]

$$\begin{aligned} &S_r(x) := \sup\left\{K_r(x; u_r, ..., u_n) : u_r \in U_r, ..., u_n \in U_n\right\} \quad \forall x, \forall r \in \{1, ..., n\} \\ &S_{n+1}(x) := 0 \qquad \forall x. \end{aligned} \tag{3.3}$$

This function plays a role analogous to that of the characteristic function $S = S(x, y)$ in the Hamilton-Jacobi theory of Sect. II.4.

Bellman's principle of dynamic programming.

Theorem 3.1. *(Bellman) The Bellman function S fulfills the (time-discrete)* principle of dynamic programming *(PDP)* [40]

$$\begin{aligned} S_r(x_r) = \sup\left\{k_r(x_r, u_r) + S_{r+1}(\varphi_r(x_r, u_r)) : u_r \in U_r\right\} \\ \forall x_r, \forall r \in \{1, ..., n\}. \end{aligned} \tag{3.4}$$

Proof. For any $r \in \{1, ..., n\}$, setting $K_{n+1} \equiv 0$, by (3.2)

$$\begin{aligned} K_r(x_r; u_r, ..., u_n) = k_r(x_r, u_r) + K_{r+1}(\varphi_r(x_r, u_r); u_{r+1}, ..., u_n) \\ \forall x_r, \forall u_r \in U_r, ..., u_n \in U_n. \end{aligned} \tag{3.5}$$

[39] Also called *value function,* because the supremum of the admissible payoffs is named the value of the optimization (or control) problem.

[40] Also called *principle of dynamic optimization*, or *principle of optimality.* (3.4) is a functional equation.

By taking the supremum w.r.t. $u_{r+1}, ..., u_n$, (3.4) follows. □

In a nutshell: according to the principle of dynamic programming, if an optimal trajectory is broken into two parts, then the second one is optimal — to wit, loosely speaking, the tail of any optimal trajectory is itself optimal.

Theorem 3.2. *(Characterization of the maximizers) Let us fix any $r \in \{1, ..., n\}$.*
For any x_r and any $(u_r, ..., u_n) \in U_r \times ... \times U_n$,

$$K_r(x_r; u_r, ..., u_n) = S_r(x_r) \tag{3.6}$$

if and only if the following optimality conditions *are fulfilled:*

$$S_j(x_j) = k_j(x_j, u_j) + S_{j+1}(\varphi_j(x_j, u_j)) \qquad \forall j \in \{r, ..., n\}. \tag{3.7}$$

Proof. For any x_r and any $(u_r, ..., u_n) \in U_r \times ... \times U_n$, let us set $x_{j+1} = \varphi_j(x_j, u_j)$ recursively for $j = r, ..., n-1$. The PDP (3.4) yields

$$\begin{aligned} S_r(x_r) &\geq k_r(x_r, u_r) + S_{r+1}(x_{r+1}) \\ &\geq k_r(x_r, u_r) + k_{r+1}(x_{r+1}, u_{r+1}) + S_{r+2}(x_{r+2}) \\ &\geq ... \geq \sum_{j=r}^{n} k_j(x_j, u_j) = K_r(x_r; u_r, ..., u_n). \end{aligned} \tag{3.8}$$

Therefore (3.6) holds if and only if these are all equalities, that is, if and only if (3.7) holds. □

Corollary 3.3. *(Bellman's optimality principle) Let $x_1 \in \mathbf{R}$ and $(u_r, ..., u_n) \in U_r \times ... \times U_n$. Then*

$$K_1(x_1; u_1, ..., u_n) = S_1(x_1) \tag{3.9}$$

if and only if, setting $x_{j+1} = \varphi_j(x_j, u_j)$ recursively for $j = 1, ..., n-1$,

$$K_r(x_r; u_r, ..., u_n) = S_r(x_r) \qquad \text{for } r = 1, ..., n. \tag{3.10}$$

Remark. It is not difficult to check the following generalizations of the Theorems 3.1 and 3.2:

(i) The PDP (3.4) admits the following obvious *multi-step* generalization:

$$\begin{aligned} S_r(x_r) = \sup\Big\{ &\sum_{j=r}^{s} k_j(x_j, u_j) + S_{s+1}(x_{s+1}) : u_r \in U_r, ..., u_s \in U_s \\ &\big(\text{with} \quad x_{\ell+1} := \varphi_\ell(x_\ell, u_\ell), \quad \ell = r, ..., s-1\big)\Big\} \\ &\forall x_r, \forall r, s \in \{1, ..., n-1\} (r \leq s). \end{aligned} \tag{3.11}$$

(ii) Dropping the supremum in (3.11) we get an obvious extension of the optimality condition (3.7). □

Implementation of dynamic programming. (3.4) is a functional equation for the unknown field S. Assuming that S in known, by means of Theorem 3.2 for any prescribed x_1

we can calculate an optimal solution $(\bar{x}_2, ..., \bar{x}_n)$, $(\bar{u}_1, ..., \bar{u}_n)$ of the maximization problems (3.10) by the following procedure, due to Bellman:

Step 1. Proceeding contrary to the direction of the process and setting $K_{n+1} \equiv 0$, for $r = n, ..., 1$, to any $x_r \in \mathbf{R}$ we associate a control $u_r \in U_r$ such that

$$k_r(x_r, u_r) + K_{r+1}(\varphi_r(x_r, u_r); u_{r+1}, ..., u_n) \text{ is maximized.} \tag{3.12}$$

This does not determine u_r, but rather defines a function $\widetilde{u}_r : x_r \mapsto u_r$.

Step 2. Proceeding in the direction of the process, we set

$$\begin{aligned} &\bar{x}_1 = x_1, \ \ \bar{u}_1 = \widetilde{u}_1(\bar{x}_1), \\ &\bar{x}_2 = \varphi_1(\bar{x}_1, \bar{u}_1), \ \ \bar{u}_2 = \widetilde{u}_2(\bar{x}_2), \ \ ..., \\ &\bar{x}_n = \varphi_{n-1}(\bar{x}_{n-1}, \bar{u}_{n-1}), \ \ \bar{u}_n = \widetilde{u}_n(\bar{x}_n). \end{aligned} \tag{3.13}$$

Notice that at the first step we evaluate functions (the $\widetilde{u}_r$s), whereas at the second one we evaluate states and controls.

Remarks. (i) The PDP is backward in time, at variance with the usual representation of physical processes. For the latter the present depends on the past; in dynamic programming of optimal control problems instead the present is ruled by the future. Notice that, if one takes a wrong step, then the whole plan must be revised. This applies to both the time-discrete and time-continuous cases (that we outline in the next section).

(ii) The PDP that we introduced in optimal control is backward. In the theory of *optimization* one also encounters forward dynamic-programming equations. For instance, let $h_r(x, y)$ be the payoff for moving from a point x to y at the rth step, and $\bar{S}_r(x)$ be the maximal payoff for reaching x at the rth step when starting from x^0. Then

$$\bar{S}_{r+1}(y) = \sup_x \{\bar{S}_r(x) + h_{r+1}(x, y)\} \qquad \forall r, \forall y, \tag{3.14}$$

which indeed is a forward equation. Actually here we have an initial datum, instead of a final one. □

Exercises. (i) Write a multi-step version of the forward equation (3.14).

(ii) Is the above strategy of optimization equivalent to selecting the highest payoff at every step (so-called *greedy strategy)?*

IV.4 Time-continuous Bellman's equation

In this section we deal with the Bellman theory for the optimal control of time-continuous processes.

Let $t_0 < t_1$, $X \subset \mathbf{R}^n$, $U \subset \mathbf{R}^m$, and fix a continuous real function

$$f :]t_0, t_1[\times X \times U \to \mathbf{R}$$

that is uniformly Lipschitz-continuous w.r.t. the second argument. For any $x_0 \in X$, we introduce the *side-conditions*

$$\begin{aligned} &x'(t) = f(t, x(t), u(t)) \qquad \forall t \in]t_0, t_1[\ \text{ (control equation)} \\ &x(t_0) = x_0 \qquad \text{(initial condition)} \\ &x(t_1) \in X \qquad \text{(final constraint)} \\ &u(t) \in U \qquad \forall t \in]t_0, t_1[\ \text{ (control constraint).} \end{aligned} \tag{4.1}$$

We shall say that a pair $(u,x) \in C^1([t_0,t_1];X) \times C^1([t_0,t_1];U)$ is *admissible* whenever it fulfills (4.1) for some $x_0 \in X$. We shall denote the solution of (4.1) by $x = x(t;t_0,x_0,u)$, and the set of these admissible pairs (u,x) by $\mathcal{C}(t_0,x_0)$. Next we prescribe another continuous function $F : X \to \mathbf{R}$, and complete the presentation of our time-continuous problem of optimal control:

$$\text{find} \quad (u,x) \in \mathcal{C}(t_0,x_0) \quad \text{such that} \quad F(x(t_1)) \quad \text{is maximized.} \tag{4.2}$$

(We write $F(x(t_1))$ in place of $F(x(t_1;t_0,x_0,u))$, and imply the side-conditions.)

Defining the *Bellman function* [40]

$$S(t_0,x_0) := \sup\big\{F(x(t_1)) : (u,x) \in \mathcal{C}(t_0,x_0)\big\}, \tag{4.3}$$

(4.2) also reads

$$\text{find } (u,x) \in \mathcal{C}(t_0,x_0) \quad \text{such that} \quad F(x(t_1)) = S(t_0,x_0). \tag{4.4}$$

Let us consider the following properties:

(a) $S(t_1,x_1) = F(x_1)$ for any $x_1 \in X$;
(b) $t \mapsto S(t,x(t))$ is nonincreasing on $[t_0,t_1]$ for any admissible pair (u,x);
(c) $t \mapsto S(t,x^*(t))$ is constant on $[t_0,t_1]$ for an admissible pair (u^*,x^*).

Theorem 4.1. *(Bellman) (i)* Necessity. *For any optimal pair (u^*,x^*), the Bellman function S defined in (4.3) fulfills the properties (a)–(c).*

(ii) Sufficiency. *If there exist a function S and an admissible pair (u^*,x^*) that fulfill the properties (a)–(c), then the pair (u^*,x^*) fulfills (4.4), namely it is optimal.*

Proof. (i) The property (a) is obvious.

Let $t_0 \le \tau_0 \le \tau_1 \le t_1$. We claim that

$$\begin{aligned} S(\tau_0,x(\tau_0)) &= \sup\big\{F(x(t_1)) : (u,x) \in \mathcal{C}(\tau_0,x_0)\big\} \\ &\ge \sup\big\{F(\bar{x}(t_1)) : (\bar{u},\bar{x}) \in \mathcal{C}(\tau_1,x(\tau_1))\big\} = S(\tau_1,x(\tau_1)). \end{aligned} \tag{4.5}$$

Indeed, if these two suprema are attained by $(u,x) \in \mathcal{C}(\tau_0,x_0)$ and $(\bar{u},\bar{x}) \in \mathcal{C}(\tau_1,x(\tau_1))$, then the pair that coincides with (u,x) in $[\tau_0,\tau_1]$ and with $(\bar{u},\bar{x})$ in $[\tau_1,t_1]$ is admissible in $[\tau_0,t_1]$. The inequality (4.5) then follows, and with it the property (b).

As (u^*,x^*) is optimal we have

$$S(t_0,x^*(t_0)) \overset{(4.1)_2}{=} S(t_0,x_0) \overset{(4.4)}{=} F(t_1^*,x^*(t_1^*)) \overset{(a)}{=} S(t_1^*,x^*(t_1^*)). \tag{4.6}$$

By (b) it then follows that $t \mapsto S(t,x^*(t))$ is constant on $[t_0,t_1^*]$. (c) is thus established.

(ii) For any admissible pair (u,x),

$$S(t_0,x_0) \overset{(b)}{\ge} S(t_1,x^*(t_1)) \overset{(a)}{=} F(x^*(t_1)). \tag{4.7}$$

[40] Obviously this formula defines the function S. By writing this, sometimes one also refers to the corresponding maximization problem, namely the twofold problem of: (i) determining whether a maximizer $(u,x) \in \mathcal{C}(t_0,x_0)$ exists; (ii) if a maximizer exists, evaluating it.

Moreover, if the pair (u^*, x^*) fulfills (c), then

$$S(t_0, x_0) \overset{(c)}{=} S(t_1, x^*(t_1)) \overset{(a)}{=} F(x^*(t_1)); \tag{4.8}$$

therefore the pair (u^*, x^*) is optimal. □

Theorem 4.2. *(Optimality condition) Let the Bellman function S be differentiable.* [41]
(i) If the pair (u, x) is admissible then

$$S_t(t, x(t)) + \nabla_x S(t, x(t)) \cdot f(t, x(t), u(t)) \leq 0 \qquad \forall t \in]t_0, t_1[. \tag{4.9}$$

(ii) The pair (u, x) is optimal if and only if

$$S_t(t, x(t)) + \nabla_x S(t, x(t)) \cdot f(t, x(t), u(t)) = 0 \qquad \forall t \in]t_0, t_1[. \tag{4.10}$$

Proof. As

$$\begin{aligned} \frac{d}{dt} S(t, x(t)) &= S_t(t, x(t)) + \nabla_x S(t, x(t)) \cdot x'(t) \\ &\overset{(4.1)_1}{=} S_t(t, x(t)) + \nabla_x S(t, x(t)) \cdot f(t, x(t), u(t)) \qquad \forall t \in]t_0, t_1[, \end{aligned}$$

the two statements above respectively follow from the properties (b) and (c) of S. □

Bellman's equation. At any instant the pair $(x(t), u(t))$ may attain any value of $X \times U$. By the latter theorem the Bellman function (if differentiable) then fulfills the following *Bellman equation:*

$$S_t(t, x) + \sup_{v \in U} \nabla_x S(t, x) \cdot f(t, x, v) = 0 \qquad \forall t \in]t_0, t_1[, \forall x \in X. \tag{4.11}$$

By the property (a) this equation is coupled with the final condition

$$S(t_1, x) = F(t_1, x) \qquad \forall x \in X. \tag{4.12}$$

Proposition 4.3. *If S is a differentiable function that fulfills (4.11) and (4.12), then the two properties of Theorem 4.2 hold.*

The above results provide the next statement.

Proposition 4.4. *The next three statements are equivalent:*
(i) S is the Bellman function associated to the optimal control problem (4.1) and (4.2);
(ii) the function S fulfills the properties (a)–(c);
(iii) if the function S is differentiable, then it fulfills the Bellman equation (4.11) coupled with the final condition (4.12). □

Exercise. Formulate and prove a time-discrete analogous of Theorem 4.1.

IV.5 The Pontryagin maximum principle

[41] Unfortunately this turns out to be a rather restrictive hypothesis, that excludes several cases of interest.

In this section we illustrate the Pontryagin maximum principle, which is essentially a reformulation of the Bellman equation.

The Hamilton-Jacobi-Bellman equation for a Mayer problem. Let us still consider the problem associated to the payoff (4.2). Setting

$$\begin{aligned} &\mathcal{H}(t,x,p,v) := p\cdot f(t,x,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n,\forall v\in U,\\ &H(t,x,p) := \sup_{v\in U}\ \mathcal{H}(t,x,p,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n, \end{aligned} \tag{5.1}$$

the Bellman equation (4.11) also reads

$$S_t + H(t,x,\nabla_x S) = 0 \qquad \forall t\in\,]t_0,t_1[,\forall x\in X. \tag{5.2}$$

This is an example of the class of Hamilton-Jacobi equations, which are quasilinear first-order PDEs. Some authors actually refer to this as the *Hamilton-Jacobi-Bellman equation* (or H-J-B equation).

Notice that:
(i) (5.2) is a PDE for the unknown function S;
(ii) once S is known, (4.10) is an algebraic equation for the unknown optimal pair $(x,u)\in\mathcal{C}(\tau_0,x_0)$ (this inclusion corresponds to the side-conditions (4.1)).

The Hamilton-Jacobi-Bellman equation for a Bolza problem. Let us now assume that there is a *running payoff* $-g$ besides the *final payoff* F, so that the total payoff reads

$$F(x(t_1)) - \int_{t_0}^{t_1} g(t,x(t),u(t))\,dt. \tag{5.3}$$

This corresponds to a Bolza problem. In this case, still assuming that S is differentiable, [41] one may easily extend Theorem 4.2, and thus derive the Bellman equation (5.2) with

$$\begin{aligned} &\mathcal{H}(t,x,p,v) := p\cdot f(t,x,v) - g(t,x,v) \quad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n,\forall v\in U,\\ &H(t,x,p) := \sup_{v\in U}\ \mathcal{H}(t,x,p,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n, \end{aligned} \tag{5.4}$$

in place of (5.1).

Pontryagin's maximum principle. We already pointed out that if $f(t,x,u)\equiv u$ then the state equation $(4.1)_1$ reads $y'(t)=u(t)$, so that

$$-\int_{t_0}^{t_1} g(t,x(t),u(t))\,dt = -\int_{t_0}^{t_1} g(t,x(t),x'(t))\,dt,$$

which is the standard Lagrange integral of the calculus of variations (sign apart). [42] In this case (5.4) reads

$$\begin{aligned} &\mathcal{H}(t,x,p,v) := p\cdot v - g(t,x,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n,\forall v\in U,\\ &H(t,x,p) := \sup_{v\in U}\ \mathcal{H}(t,x,p,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n, \end{aligned} \tag{5.5}$$

[41] This is an important issue for the analysis, since (as we already pointed out) in this context this hypothesis is rather restrictive.
[42] This change in sign is related to the fact that here we are dealing with a maximization problem.

which is the Legendre transform of the Lagrangian $g(t, x, \cdot)$. We have thus retrieved the classical Hamiltonian. For this reason in optimal control theory the function H that we defined in (5.4) is often referred to as the *(generalized) Hamiltonian.*

The maximization $(5.5)_2$ is sometimes called a maximum principle; [43] for this reason $(5.4)_2$ is often referred to as the *Pontryagin maximum principle.* This notion plays an important role in optimal control theory.

Feedback. The Bellman equation (4.10) establishes a relation between any state x and a (possibly nonunique) optimal control u. This defines $u = u^*(t, x)$ as a (possibly multivalued) mapping. Assuming that a solution S of the H-J equation (4.10) is known and setting

$$\Psi(t, x, v) := \nabla_x S(t, x) \cdot f(t, x, v), \tag{5.6}$$

(4.11) is thus equivalent to

$$\Psi(t, x, u) = \sup \Psi(t, x, \cdot) \quad \text{with the side-conditions (4.1).} \tag{5.7}$$

This relation between an optimal control and the corresponding state holds at any instant; this is an example of what in control theory is named a *feedback relation.* Relations of this type are of paramount applicative interest.

In optimal control theory one distinguishes between two classes of systems. In *open-loop* systems the optimal control is prescribed once for all, and is independent of the (past and) current state of the system. [44] In *closed-loop* (also named *Markovian)* systems instead at any instant t the optimal control is determined on the basis of the state at the instant t via a feedback relation. On the other hand the state is determined by the control up to the instant t, so that the optimal control is determined by the evolution of the control up to t: this (so to say) *closes the loop.*

In *open-loop* systems the optimal control is prescribed as a function of time: $u = u(t)$. In *closed-loop* systems the feedback relation provides the optimal control as a function of time and of the state variable x: $u(t) = u(t, x(t))$; of course this is possible only if at any time t the controller has access to the variable $x(t)$.

Closed-loop systems may be more demanding in terms of computation, but are to be preferred to the open-loop ones, also because they give the controller the possibility to react to deviations from the expected value of the state variable due to approximation errors and/or to exterior perturbations. One then says that the control is *robust.*

One speaks of the *synthesis problem.* For *open-loop* systems, this consists in selecting an optimal control. For *closed-loop* systems instead this problem consists in prescribing the feedback relation.

A linear example. Let us consider the so-called *linear-regulator problem:*

$$\begin{aligned} &\text{minimize} \quad \int_{t_0}^{t_1} \left(x(t)^2 + u(t)^2\right) dt \\ &x'(t) = ax(t) + bu(t) \qquad \forall t \in \,]t_0, t_1[\;\; \text{(state equation).} \end{aligned} \tag{5.8}$$

[43] In the theory of differential equations the term *maximum principle* has a different meaning.

[44] This is quite different from the concept of *open-loop* as a rhetorical device, that is used to instill curiosity by creating anticipation for what will come next. Example: "I just saw the most amazing thing!" Since the actor doesn't say what the amazing thing is, the observer is left to wonder what it is, thus creating anticipation for what the actor will say next.

Here the cost functional explicitly depends on the control. This dependence may be removed by setting $\vec{x} = (x_1, x_2) := (x, u)$, so that we may reformulate (5.8) as

$$\text{maximize} \quad -\int_{t_0}^{t_1} \left(x_1(t)^2 + x_2(t)^2\right) dt \tag{5.9}$$

$$\begin{cases} x_1'(t) = ax(t) + bx_2'(t) \\ x_2'(t) = u(t) \end{cases} \qquad \forall t \in]t_0, t_1[\ \text{ (state equation).} \tag{5.10}$$

*** Proposition 5.4.** *Let $c = c(t)$ be a solution of the following* Riccati equation

$$c'(t) + 2ac(t) + 1 = b^2 c(t)^2 \qquad \forall t \in]t_0, t_1[, \tag{5.11}$$

coupled with the final condition $c(t_1) = 0$. [45] *If the pair of functions $(u, \vec{x})$ fulfills the state equation (5.10) and is such that*

$$u'(t) = -c(t)bx_1(t) \qquad \forall t \in]t_0, t_1[\text{ (feedback equation)}, \tag{5.12}$$

then $(u, \vec{x})$ is an optimal pair.

Remark. In Sect. II.4, dealing with the Hamilton-Jacobi theory in mechanics, we studied a minimization problem and varied the right-side extremum. In this section instead we considered a maximization problem and varied the left-side extremum, because of the backward structure of optimal control problems.

If instead of maximizing a payoff one minimizes a cost, then a minimum principle is obtained, as it was the case in (2.18). □

Time-discrete vs. time-continuous. In the previous section, dealing with time-discrete optimization we proved the principle of dynamic programming (PDP); in the present section, dealing with time-continuous optimization we derived the Bellman equation. Here we derive the time-continuous PDP.

Let us now consider the Bolza problem that consists in maximizing the payoff (5.3) with the side-conditions (4.1). Denoting by $x(t; t_0, x_0, u(\cdot))$ any solution of (4.1), via the argument of Theorem 3.1 one may easily derive the time-continuous PDP in the following form:

$$S(t_0, x_0) = \sup_{u \in U} \left\{ \int_{t_0}^{\tau} g(t, x(t; t_0, x_0, u(\cdot)))\, dt + S(\tau, x(\tau; t_0, x_0, u(\cdot))) \right\} \qquad \forall x_0 \in X, \forall \tau \in]t_0, t_1[. \tag{5.13}$$

It is not difficult to see that the Bellman equation (5.2) (with the Hamiltonian H as in (5.5)) is the differential form of this integral condition. □

Exercise. Formulate a time-continuous PDP for a Mayer problem.

IV.6 The Pontryagin equations

In this section we show that:

[45] Riccati equations are ODEs of the form $y'(t) = \alpha y(t)^2 + \beta y(t) + \gamma$. They play an important role in the analysis of optimal control problems with a quadratic cost functional.

(i) the canonical equations for the generalized Hamiltonian (6.1) (below) are the *characteristic equations* [45] of the Hamilton-Jacobi-Bellman PDE, and

(ii) these canonical equations are equivalent to the Pontryagin equations (that we derive here, too).

Let us consider the Mayer-type control problem (4.1) and (4.2), and define the Bellman function S as in (4.3). As we saw, setting

$$H(t,x,p) := \sup_{v\in U} \; p\cdot f(t,x,v) \qquad \forall t\in\,]t_0,t_1[,\forall x\in X,\forall p\in\mathbf{R}^n, \tag{6.1}$$

the H-J-B equation reads

$$S_t(t,x)+H(t,x,\nabla_x S(t,x))=0 \qquad \forall t\in\,]t_0,t_1[,\forall x\in X. \tag{6.2}$$

The method of characteristics. Let us assume that the Bellman function is twice differentiable, and (as usual) set

$$p(t,x) := \nabla_x S(t,x) \qquad \forall (t,x)\in\mathbf{R}^{n+1}. \tag{6.3}$$

Differentiating the H-J-B equation (6.2) w.r.t x_i, we then have (summing over repeated indices)

$$\begin{aligned}\frac{\partial p_i}{\partial t}(t,x(t))+H_{x_i}(t,x,p(t,x))+\frac{\partial p_j}{\partial x_i}(t,x(t))\;H_{p_j}(t,x,p(t,x))=0\\ \forall t\in\,]t_0,t_1[,\forall x\in X, i=1,...,n.\end{aligned} \tag{6.4}$$

For a regular curve $t\mapsto x(t)$ in $\mathbf{R}^n$, to be specified ahead, we have

$$\frac{d}{dt}p_i(t,x(t))=\frac{\partial p_i}{\partial t}(t,x(t))+\dot{x}_j(t)\frac{\partial p_i}{\partial x_j}(t,x(t)) \qquad \forall t\in\,]t_0,t_1[,i=1,...,n. \tag{6.5}$$

Notice that by (6.3)

$$\frac{\partial p_i}{\partial x_j}=\frac{\partial^2 S}{\partial x_i\partial x_j}=\frac{\partial^2 S}{\partial x_j\partial x_i}=\frac{\partial p_j}{\partial x_i} \qquad \forall i,j=1,...,n. \tag{6.6}$$

By evaluating (6.4) along the curve $t\mapsto x(t)$ and using (6.5), we then get

$$\begin{aligned}\frac{d}{dt}p_i(t,x(t))=-H_{x_i}(t,x(t),p(t,x(t)))+\Big[\dot{x}_j(t)-H_{p_j}(t,x(t),p(t,x(t)))\Big]\frac{\partial p_j}{\partial x_i}\\ \forall t\in\,]t_0,t_1[,i=1,...,n.\end{aligned} \tag{6.7}$$

If $\dot{x}_j(t)=H_{p_j}(t,x(t),p(t,x(t)))$ for any j, we get rid of the term with the *uncomfortable* second-order factor $\partial p_j/\partial x_i$. In this way the equation (6.7) is reduced to

$$\frac{d}{dt}p_i(t,x(t))=-H_{x_i}(t,x(t),p(t,x(t))).$$

[45] These ODEs are a classical tool of the study of (first order) PDE. The basic idea is to solve the PDE in a domain by integrating suitable ODEs; these ODEs represent curves that join the interior points of the domain with the part of the boundary where the data are prescribed.

(46) We thus get the canonical system

$$\begin{cases} \dfrac{d}{dt}x_i(t) = H_{p_i}(t, x(t), \widetilde{p}(t)) \\ \dfrac{d}{dt}\widetilde{p}_i(t) = -H_{x_i}(t, x(t), \widetilde{p}(t)) \end{cases} \qquad \forall t \in]t_0, t_1[, i = 1, ..., n. \tag{6.8}$$

(Notice that we wrote $\widetilde{p}_i(t)$ in place of $p_i(t, x(t))$.) We thus derived the canonical equations for the generalized Hamiltonian (6.1) from the Bellman equation.

In terms of the theory of PDEs, the system (6.8) defines the so-called *projected characteristics* of the PDE (6.2).

The Pontryagin system. Let us set

$$u(t, x, p) := \text{arg-max } \{p \cdot f(t, x, \cdot)\} \qquad \forall t \in]t_0, t_1[, \forall x \in X, \forall p \in \mathbf{R}^n, \tag{6.9}$$

which defines the feedback relation. By (6.1) we then have

$$H(t, x, p) = p \cdot f(t, x, u(t, x, p)) \qquad \forall t \in]t_0, t_1[, \forall x \in X, \forall p \in \mathbf{R}^n. \tag{6.10}$$

Therefore

$$H_{x_i}(t, x, p) = p \cdot f_{x_i}(t, x, u(t, x, p)) + p \cdot f_{u_j}(t, x, u(t, x, p)) \, \frac{\partial u_j}{\partial x_i}(t, x, p) \qquad \forall t \in]t_0, t_1[, i = 1, ..., n, \tag{6.11}$$

$$H_{p_i}(t, x, p) = f(t, x, u(t, x, p)) + p \cdot f_{u_j}(t, x, u(t, x, p)) \, \frac{\partial u_j}{\partial p_i}(t, x, p) \qquad \forall t \in]t_0, t_1[, i = 1, ..., n. \tag{6.12}$$

If $U = \mathbf{R}^m$ then the maximum of $v \mapsto \mathcal{H}(t, x, p, v)$ (cf. (5.1)) is attained at an interior point; hence for $v = u(t, x, p)$

$$p \cdot f_{u_j}(t, x, u(t, x, p)) = p \cdot f_{v_j}(t, x, v)\big|_{v=u(t,x,p)} \overset{(5.1)}{=} \mathcal{H}_{v_j}(t, x, p, v)\big|_{v=u(t,x,p)} = 0 \qquad \forall t \in]t_0, t_1[, \forall x \in X, \forall p \in \mathbf{R}^n, j = 1, ..., n. \tag{6.13}$$

By (6.11)–(6.13), the canonical system (6.8) is then equivalent to the *Pontryagin system*

$$\begin{cases} \dfrac{d}{dt}x_i(t) = f(t, x(t), u(t)) \\ \dfrac{d}{dt}\widetilde{p}_i(t) = -\widetilde{p}(t) \cdot f_{x_i}(t, x(t), u(t)) \end{cases} \qquad \forall t \in]t_0, t_1[, i = 1, ..., n. \tag{6.14}$$

These are respectively referred to as the state equation and the adjoint equation.

The Pontryagin system (either in the form (6.8) or (6.14)) must be coupled with the initial condition $x(t_0) = x_0$, and with the final condition

$$\widetilde{p}_i(t_1) = F_{x_i}(x(t_1)) \quad i = 1, ..., n \text{ (so-called } \textit{transversality condition)}. \tag{6.15}$$

(46) This extends the adjoint-state equation to the case in which the state equation is non-linear.

Indeed, by the property (a) of Sect. IV.4,

$$p_i(t_1, x) := S_{x_i}(t_1, x) = F_{x_i}(x) \qquad \forall x \in X, i = 1, ..., n. \tag{6.16}$$

This analysis may be extended to the Bolza problem associated to the payoff (5.3). In that case we still have $(6.14)_1$ and $(6.14)_2$, but $(6.14)_3$ is replaced by

$$\frac{d}{dt}\widetilde{p}_i(t) = g(t, x(t)) - \widetilde{p}(t) \cdot f_{x_i}(t, x(t), u(t)) \qquad \forall t \in]t_0, t_1[, i = 1, ..., n. \tag{6.17}$$

Maximization conditions. We have seen that the Pontryagin system is equivalent to the canonical system, and that in turn this is equivalent to the extremality of the payoff functional, constrained by the side-conditions (4.1). Therefore the Pontryagin system is a necessary condition for maximality.

One can show that, if the Hamiltonian H is convex and F is concave, then the Pontryagin system is also a sufficient condition for maximality.

IV.7 Differential games (Isaacs's theory)

Game theory deals with conflicts among rational decision makers. Differential games represent systems that evolve in time according to a differential dynamics.

Zero-sum differential game. Let us fix a vector $x^0 \in X \subset \mathbf{R}^n$, two nonempty sets $U_A, U_B \subset \mathbf{R}^m$, and a continuous function

$$f : \mathbf{R}^+ \times \mathbf{R}^n \times U_A \times U_B \to \mathbf{R}^n \quad (U \subset \mathbf{R}^m), \tag{7.1}$$

that is uniformly Lipschitz-continuous w.r.t. the second argument. Let $\mathcal{A}$ ($\mathcal{B}$, resp.) be a set of measurable functions $\mathbf{R}^+ \to U_A$ ($\mathbf{R}^+ \to U_B$, resp.), let $0 < t_0 < t_1 < +\infty$, and let a variable $x = x(t)$ depend on two controls $a = a(t)$ and $b = b(t)$ via the Cauchy problem

$$\begin{cases} x'(t) = f(t, x(t), a(t), b(t)) & \forall t \in]t_0, t_1[, \\ x(t_0) = x_0. \end{cases} \tag{7.2}$$

We shall denote the solution by $x = x(t; t_0, x_0, a, b)$. We also prescribe two continuous functions

$$g : \mathbf{R}^+ \times \mathbf{R}^n \times U_A \times U_B \to \mathbf{R}, \qquad F : X \to \mathbf{R}, \tag{7.3}$$

and introduce a Bolza-type functional

$$J_{(t_0,x_0)}(a, b) := F(x(t_1; t_0, x_0, a, b)) - \int_{t_0}^{t_1} g(t, x(t; t_0, x_0, a, b), a(t), b(t))) \, dt \qquad \forall a \in \mathcal{A}, \forall b \in \mathcal{B}. \tag{7.4}$$

We display the initial point (t_0, x_0) as an index, since we shall let it vary, in a similar way to what we did in Bellman's theory of dynamic programming.

The model that we have in mind consists of two players A and B, that respectively select the controls $a \in \mathcal{A}$ and $b \in \mathcal{B}$, with the respective purpose of maximizing and minimizing the functional $J_{(t_0,x_0)}(a, b)$. This is a *zero-sum differential game.* [47]

[47] In economics and politics, differential games are also labelled as *dynamic games* (so that one may not use derivatives...).

If either $\mathcal{A}$ or $\mathcal{B}$ is a singleton, then the differential game is reduced to an optimal control problem. We may complement the statement (1.9) as follows:

(i) the theory of differential games extends the control theory,
(ii) the control theory extends the basic Bolza problem of calculus of variations. (7.4′)

Actually some developments of the present section mimic those of Sect. IV.4.

The static game. Let us define the following *(static) lower-* and *upper-value functions* of the game:

$$\begin{aligned} S_s^-(t_0,x_0) &:= \sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} J_{(t_0,x_0)}(a,b) \\ S_s^+(t_0,x_0) &:= \inf_{b\in\mathcal{B}} \sup_{a\in\mathcal{A}} J_{(t_0,x_0)}(a,b) \end{aligned} \qquad \forall t_0>0, \forall x_0\in X. \tag{7.5}$$

Proposition 7.1. *Under the above assumptions,*

$$S_s^-(t_0,x_0) \le S_s^+(t_0,x_0) \qquad \forall t_0>0, \forall x_0\in X. \tag{7.6}$$

Proof. For any fixed $\bar b\in\mathcal{B}$,

$$\inf_{b\in\mathcal{B}} J_{(t_0,x_0)}(a,b) \le J_{(t_0,x_0)}(a,\bar b) \qquad \forall a\in\mathcal{A}.$$

Taking first the $\sup_{a\in\mathcal{A}}$ and then the $\inf_{\bar b\in\mathcal{B}}$, (7.6) follows. □

The inf-sup inequality (7.6) is a basic property of the theory of saddle points, and of game theory. If $J_{(t_0,x_0)}$ is upper (lower, resp.) semicontinuous with respect to a (b, resp.), and $\mathcal{A},\mathcal{B}$ are compact, then the supremum is a maximum and the infimum is a minimum. As we saw for minimization problems, the compactness of the domain may be replaced by the coerciveness of the functional. More specifically, instead of assuming that $\mathcal{A}$ ($\mathcal{B}$, resp.) is compact, one may require that $-J$ is coercive with respect to a for any fixed b (J is coercive with respect to b for any fixed a, resp.).

Whenever in (7.6) the equality holds, one says that the game has a value, and the mapping

$$\mathbf{R}^+\times X\to\mathbf{R} : (t_0,x_0)\mapsto S_s(t_0,x_0) := S_s^-(t_0,x_0) = S_s^+(t_0,x_0)$$

is called the *value function of the (static) game.*

The static game corresponds to applying the open-loop approach to the differential game. The above rules do not take account of the obvious fact that at any instant the players ignore the future. The game is said static since in this way the peculiarity of the dynamics (i.e., causality) is lost. The differential game is thus reduced to a problem of saddle point, as in a standard game. [48]

In order to visualize this, let us assume that the game is discrete in time, and at each time-step (each *move,* if e.g. one thinks of chess), [49] one of the two players has the privilege of moving after the other one. For instance, $(7.5)_1$ corresponds to B choosing the move b after A has chosen a. Similarly, $(7.5)_2$ may be interpreted as A moving after B. If the inequality (7.6) is strict, then there is an advantage in moving as second. If, for instance, B (who strives to minimize J) moves after A, then the outcome is $S_s^-(t_0,x_0)$; by (7.6) this is a

[48] By contrast with differential games, sometimes nondifferential games are indeed labelled as static games.
[49] However, chess do not seem to fit the present framework.

smaller value (for B a better result) than $S_s^+(t_0, x_0)$, which is the outcome if A moves after B.

Game strategies and value functions. Next we introduce a closed-loop model, in which the control of each player depends on the state, and therefore on the past and present controls.

Each of the two players is assumed to have a *strategy,* [(50)] and to ignore the strategy of the opponent. The admissible strategies of the player A are the *nonanticipating* maps $\alpha : \mathcal{B} \to \mathcal{A}$; by this we mean that

$$\forall b, \tilde{b} \in \mathcal{B}, \forall t > 0, \quad \text{if} \quad b = \tilde{b} \ \text{ in }]0, t[, \ \text{ then } \ \alpha(b) = \alpha(\tilde{b}) \ \text{ in }]t, +\infty[. \tag{7.7}$$

Thus the strategies only depend on the past (here dependence on the present has little meaning, since a and b are defined a.e. in time), since the players (the strategists...) cannot foresee the future. The strategies of the player B are defined similarly. We shall denote by Γ_A (Γ_B, resp.) the set of all strategies of the player A (B, resp.).

Let us define the following *(dynamic) lower-* and *upper-value functions* of the game:

$$\begin{aligned} S^-(t_0, x_0) &:= \inf_{\beta \in \Gamma_B} \sup_{a \in \mathcal{A}} J_{(t_0, x_0)}(a, \beta(a)) \\ S^+(t_0, x_0) &:= \sup_{\alpha \in \Gamma_A} \inf_{b \in \mathcal{B}} J_{(t_0, x_0)}(\alpha(b), b) \end{aligned} \qquad \forall t_0 > 0, \forall x_0 \in X. \tag{7.8}$$

This reflects a conservative attitude: each player selects his/her strategy on the basis of the worst case (one may thus speak of a *worst-case design).*

Proposition 7.2. *Under the above assumptions,*

$$S_s^-(t_0, x_0) \le S^-(t_0, x_0) \le S^+(t_0, x_0) \le S_s^+(t_0, x_0) \qquad \forall t_0 > 0, \forall x_0 \in X. \tag{7.9}$$

Here one player selects a control, the other one replies according to his/her own strategy. We claim that, if the middle inequality of (7.9) is strict, then the player who plays as second has an advantage, similarly to what we saw for the static game. [(51)] For instance $(7.8)_1$ corresponds to B (who strives to minimize J) selecting his/her strategy after that A has moved; in this way the best outcome he/she may achieve is $S^-(t_0, x_0)$. Conversely $(7.8)_2$ corresponds to A (who strives to maximize J) selecting his/her strategy after that B has moved; in this way the best outcome he/she may achieve is $S^+(t_0, x_0)$. So the one who moves the second may achieve a better result (a smaller value for B, a larger value for A) than in the case in which he/she moves as first.

Notice that in (7.9) the inequality $S^-(t_0, x_0) \le S^+(t_0, x_0)$ reads "$\inf \sup \le \sup \inf$", which is the opposite of what occurs in (7.6)! How is this possible? or rather, doesn't the inequality "$\sup \inf \le \inf \sup$" stem from the argument of Proposition 7.1? The answer is no, because the two suprema are taken over different sets in $(7.8)_1$ and $(7.8)_2$, and the same holds for the two infima.

[(50)] In the theory of differential games, strategies play a similar role to that of feedback in the theory of optimal control.

However, there are also relevant differences. For instance, any closed-loop control problem may be transformed into an equivalent open-loop problem; that is, the optimal control may be represented in feedback form. On the other hand, closed-loop differential games are quite different from open-loop games, which cannot be reformulated in terms of feedback.

[(51)] This is at variance with what happens e.g. for chess, where however the two players do not move simultaneously.

Dynamic programming principle. The next statement may be proved along the lines of the analogous result for the control theory, see Sect. IV.4.

Proposition 7.3. *Under the above assumptions, for any $t_0 > 0$, any $x_0 \in X$ and any $\tau \in]t_0, t_1[$,*

$$\begin{aligned} S^-(t_0,x_0) &= \inf_{\beta\in\Gamma_B}\sup_{a\in\mathcal{A}}\Big\{-\int_{t_0}^{\tau} g(t,x(t;t_0,x_0,a(t),\beta(a(t))),a(t),\beta(a(t)))\,dt \\ &\qquad + S^-(\tau,x(\tau;t_0,x_0,a,\beta(a)))\Big\}, \\ S^+(t_0,x_0) &= \sup_{\alpha\in\Gamma_A}\inf_{b\in\mathcal{B}}\Big\{-\int_{t_0}^{\tau} g(t,x(t;t_0,x_0,\alpha(b(t)),b(t)),\alpha(b(t)),b(t))\,dt \\ &\qquad + S^+(\tau,x(\tau;t_0,x_0,\alpha(b),b))\Big\}. \end{aligned} \tag{7.10}$$

Coupled control problems. By (7.8), the differential game may be regarded as two coupled control problems. For instance, A faces a maximization problem with control $\alpha \in \Gamma_A$ and payoff

$$\Phi_{(t_0,x_0)}(\alpha) := \inf_{b\in\mathcal{B}} J_{(t_0,x_0)}(\alpha(b),b) \qquad \forall \alpha \in \Gamma_A, \tag{7.11}$$

for any admissible (t_0, x_0). The mapping S^+ is thus the value-function of this control problem. An analogous interpretation applies to S^- for the player B. The results of Sect. IV.4 then yield the next statement.

Proposition 7.4. *If $g \equiv 0$ then the following properties hold:*
(a) $S^+(t_1,x_1) = F(x_1)$ for any $x_1 \in X$;
(b) $t \mapsto \Phi_{(t,x(t;t_0,x_0,\alpha(b),b))}(\alpha)$ is nonincreasing in $[t_0,t_1]$ for any $\alpha \in \Gamma_A$;
(c) $t \mapsto \Phi_{(t,x(t;t_0,x_0,\alpha(b),b))}(\alpha)$ is constant in $[t_0,t_1]$ if and only if the supremum in (7.8)$_2$ is attained by α.

Symmetric properties hold for the function

$$\Psi_{(t_0,x_0)}(\beta) := \sup_{a\in\mathcal{A}} J_{(t_0,x_0)}(a,\beta(a)) \qquad \forall \beta \in \Gamma_B. \tag{7.12}$$

Hamilton-Jacobi-Isaacs equations. Let us next define the following *lower* and *upper Hamiltonian:*

$$\begin{aligned} H^-(t,x,p) &:= \sup_{a\in U_A}\inf_{b\in U_B}\{f(t,x,a,b)\cdot p - g(t,x,a,b)\} \\ H^+(t,x,p) &:= \inf_{b\in U_B}\sup_{a\in U_A}\{f(t,x,a,b)\cdot p - g(t,x,a,b)\} \end{aligned} \qquad \forall x\in X, \forall p\in\mathbf{R}^n, \forall t. \tag{7.13}$$

As in Proposition 7.1, a min-max inequality is easily established:

$$H^-(t,x,p) \le H^+(t,x,p) \qquad \forall x\in X, \forall p\in\mathbf{R}^n, \forall t. \tag{7.14}$$

Theorem 7.5. *(Isaacs) If*

$$\begin{aligned} &(H(t,x,p) :=)\ H^-(t,x,p) = H^+(t,x,p) \\ &\forall x\in X, \forall p\in\mathbf{R}^n, \forall t \ \text{ (Isaacs condition)}, \end{aligned} \tag{7.15}$$

then the common value is named the game Hamiltonian. *In this case the game has the value* $S := S^- = S^+$, *and this fulfills the following* Hamilton-Jacobi-Isaacs equation

$$S_t(t,x) + H(t,x,\nabla_x S(t,x)) = 0 \qquad \forall t \in]t_0,t_1[, \forall x \in X, \tag{7.16}$$

coupled with the final condition $S(t_1,x_1) = F(x_1)$ *for any* $x_1 \in X$.

S however need not coincide with either S_s^- or S_s^+.

Pontryagin-type equations. Let us assume that (7.15) holds, and set

$$p(t,x) := \nabla_x S(t,x) \qquad \forall t \in]t_0,t_1[, \forall x \in X. \tag{7.17}$$

If (7.15) is also fulfilled and $\widetilde{a} \in \mathcal{A}, \widetilde{b} \in \mathcal{B}$ are optimal controls, [52] then

$$H(t,x,p) = f(t,x,\widetilde{a},\widetilde{b}) \cdot p - g(t,x,\widetilde{a},\widetilde{b}) \qquad \forall x \in X, \forall p \in \mathbf{R}^n, \forall t. \tag{7.18}$$

Denoting by $\widetilde{x}$ the corresponding state and setting

$$\widetilde{p}(t) := p(t,x(t)) \qquad \forall t \in]t_0,t_1[, \tag{7.19}$$

one may show that [53]

$$\dot{\widetilde{p}}_i(t) = -H_{x_i}(t,x(t),\widetilde{p}(t)) \qquad \forall t \in]t_0,t_1[, i = 1,...,n. \tag{7.20}$$

On the other hand by (7.16) the state equation also reads

$$\dot{\widetilde{x}}_i(t) = H_{p_i}(t,\widetilde{x}(t),\widetilde{p}(t)) \qquad \forall t \in]t_0,t_1[, i = 1,...,n. \tag{7.21}$$

We thus retrieved the analogous of the canonical system (6.8) of control theory; a more explicit representation of (7.20) and (7.21) has a similar form to the Pontryagin system $(6.14)_2$ and $(6.14)_3$. In analogy with that setting, (7.20) and (7.21) must be coupled with the initial condition $x(t_0) = x_0$, and with the (final) transversality condition

$$\widetilde{p}_i(t_1) = F_{x_i}(x(t_1)) \qquad i = 1,...,n. \tag{7.22}$$

On mechanics, control theory, differential games, and PDE theory.

In analytical mechanics the canonical system of ODEs is associated to the H-J PDE.

In optimal control the Pontryagin system of ODEs is associated to the Bellman PDE.

In differential games, a Pontryagin-like system of ODEs is associated to the Isaacs PDE.

In the theory of PDEs, a characteristic system of ODEs (the *Charpit equations)* is associated to a quasi-linear first-order PDE. [54]

In each of these cases, a system of ODEs is equivalent to a PDE.

In mechanics $(6.8)_1$ stems from the Legendre transformation; $(6.8)_2$ from the Euler-Lagrange equation (typically, the motion law) and an initial-value is prescribed for p.

In optimal control theory $(6.14)_2$ is the state equation; $(6.14)_3$ is the adjoint-state equation, and a final-value is prescribed for p.

In differential games the setting is similar to that of optimal control.

Remarks. (i) One may provide a feedback-type formulation also for differential games, as for optimal control.

(ii) A time-discrete formulation with a time-discrete state equation may also be dealt with. □

[52] We are assuming that optimal controls exist. To prove this existence is a rather delicate problem, that has been addressed only recently.

[53] We do not display the argument, which is along the lines of Sect. IV.6.

[54] See e.g. Evans's book on PDEs p. 97.

V. ELEMENTS OF CONVEX CALCULUS

V.1. Convex lower semicontinuous functions

We deal with a real Banach space B, although several results of this chapter also hold for any real separated locally convex space. By $\langle \cdot, \cdot \rangle$ we denote the duality pairing between the topological dual B^* and B. As usual, if it is not otherwise specified, we refer to the *strong* topology.

Lemma 1.0. *(i) Any set $K \subset B$ is closed (convex, resp.) if and only if I_K is lower semicontinuous (convex, resp.).*

(ii) Any function $f : B \to]-\infty, +\infty]$ is lower semicontinuous (convex, resp.) if and only if epi(f) *is closed (convex, resp.).*

(iii) If $\{K_i\}_{i \in I}$ is a family of closed (convex, resp.) subsets of B, then $\bigcap_i K_i$ is closed (convex, resp.).

(iv) If $\{f_i\}_{i \in I}$ is a family of lower semicontinuous (convex, resp.) functions $B \to]-\infty, +\infty]$, then their upper hull $f(\cdot) := \sup_i f_i(\cdot)$ is lower semicontinuous (convex, resp.).

On account of a classical Mazur's theorem that is known as the *geometric Hahn-Banach theorem,* closed convex sets and convex lower semicontinuous functions play a key role.

• **Proposition 1.1.** *(i) Any convex set $K \subset B$ is closed if and only if it is weakly closed, and if and only if it is sequentially weakly closed.*

(ii) Any convex function $F : B \to]-\infty, +\infty]$ is lower semicontinuous if and only if it is weakly lower semicontinuous, and if and only if it is sequentially weakly lower semicontinuous.

Proof. Let K be a closed subset of B. If $u \notin K$ then, by Mazur's theorem $\{u\}$ (which obviously is a convex compact set!) can be strongly separated from K by means of a closed hyperplane. This means that there exists $u^* \in B^*$ and $C \in \mathbf{R}$ such that

$$\langle u^*, v \rangle \leq C \quad \forall v \in K, \qquad \langle u^*, u \rangle \geq C. \tag{2.}$$

Hence u does not belong to the weak closure of K, and this proves the first statement. The analogous statement for the sequential weak topology follows since this topology is intermediate between the strong and the weak topology. [(55)] The second part then follows from Lemma 1.0. □

Remark. Of course the same result also holds for B^*. However, if B is not reflexive, a subset of B^* may be convex and weakly (equivalently, strongly) closed without being weakly star closed. For instance, this occurs for the half-space $\{u^* \in B^* : \langle u^{**}, u^* \rangle \geq 0\}$ for any $u^{**} \in B^{**} \setminus B$. [(56)]

In order to avoid any use to bidual space B^{**}, we deal with the duality between B^* and B, rather than that between B^* and B^{**}. This may be expressed by saying that B and B^* are regarded as *spaces in duality*. We shall then often use the weak star topology instead of the weak one. Of course, for reflexive spaces this distinction is meaningless.

The argument of Proposition 1.1 also yields the following result, by which closed half-spaces and continuous affine functions also play an important role.

[(55)] On the other hand, for nonreflexive dual spaces this need not apply to the weak star topology, which is not intermediate between the strong and the weak topology.

[(56)] Here $\langle \cdot, \cdot \rangle$ represents the duality pairing between B^{**} and B^*.

• **Proposition 1.2.** *(i) Any set $K \subset B$ ($K \neq B$) is convex and closed if and only if it is the intersection of a (nonempty) family of closed half-spaces.*

(ii) Any function $F : B \to]-\infty, +\infty]$ is convex and lower semicontinuous if and only if it is the upper hull of a (nonempty) family of continuous affine functions. (57)

We shall denote by $\Gamma(B)$ the class of convex lower semicontinuous functions $B \to]-\infty, +\infty]$, and by $\Gamma_0(B)$ the subclass of functions not identically equal to $+\infty$.

For any set $K \subset B$, the smallest closed convex subset of B which contains K is called the *closed convex hull* of K. It coincides with the intersection of all the closed convex subsets of B which contain K. By Proposition 1.2(i), it also coincides with the intersection of all of the closed half-spaces which contain K. As the closure of any convex set is convex, that hull also coincides with $\overline{\text{co}}(K)$, namely, the closure of the convex hull of K. But it need not coincide with the convex hull of the closure of K, since this may not be closed. The set $K := \{(0,0)\} \cup \{(x,y) \in (\mathbf{R}^+)^2 : xy \geq 1\}$ is a counterexample.

Similarly, let us consider any function $F : B \to]-\infty, +\infty]$ which has a convex and lower semicontinuous lower bound. By Proposition 1.2(ii), F then has a continuous affine lower bound, and the upper hull of these lower bounds is the largest lower bound of F in $\Gamma(B)$. It is called the Γ-*regularized function* of F, and its epigraph coincides with the closed convex hull of the epigraph of F. []

Regularity Properties of Convex Functions. Convexity is a source of regularity, as it is shown by the following classical result.

*** Theorem 1.2'..** *(Alexandrov) If $F : \mathbf{R}^N \to \mathbf{R}$ is convex, then it is twice differentiable a.e. in $\mathbf{R}^N$. []*

*** Theorem 1.3.** *Let $F : B \to]-\infty, +\infty]$ be convex. If there exists a (nonempty) open set $A \subset B$ in which F is upperly bounded, then F is locally Lipschitz continuous in* $\text{int}(\text{Dom}(F))$. (57)

Proof. (i) We claim that F is continuous in A.

Let $u_0 \in A$. Without loss of generality, we may assume that $(u_0, F(u_0)) = (0,0)$. Let $M, R > 0$ be such that $F(v) \leq M$ for any v such that $\|v\| \leq R$. For any $\varepsilon > 0$, by the convexity we have $F(\varepsilon v) \leq (1-\varepsilon)F(0) + \varepsilon F(v) \leq \varepsilon M$; moreover $F(\varepsilon v) + F(-\varepsilon v) \geq F(0) = 0$, whence $F(\varepsilon v) \geq -F(-\varepsilon v) \geq -\varepsilon M$. Therefore F is continuous at 0.

(ii) We claim that F is continuous in $\text{int}(\text{Dom}(F))$.

We still assume that A is a neighbourhood of the origin. Let us fix any $w \in \text{int}(\text{Dom}(F))$, and let $\rho > 1$ be such that $z := \rho w \in \text{int}(\text{Dom}(F))$. Notice that $\tilde{A} := w + (1 - 1/\rho)A$ is a neighbourhood of w, and that $\sup_{\tilde{A}} F \leq \max\{F(z), \sup_A F\} < +\infty$ by the convexity of F. By part (i), F is then continuous at w.

(iii) Finally we prove that F is Lipschitz continuous in a neighbourhood of any point $u_0 \in \text{int}(\text{Dom}(F))$. Let $\delta > 0$ be such that $|F| \leq M$ in $B(u_0, 2\delta)$. Let us fix any $v, w \in B(u_0, \delta)$ and set $a := \|v - w\|$, $z := w + (\delta/a)(w - v)$. Thus $z \in B(w, \delta) \subset B(u_0, 2\delta)$. By convexity we have

$$F(w) \leq \frac{a}{a+\delta} F(z) + \frac{\delta}{a+\delta} F(v),$$

(57) In passing note that by convention the convex set B is the intersection of the empty family, and $F \equiv -\infty$ is the upper hull of the empty family of functions.

(57) By $\text{int}(A)$ we denote the interior of the set A.

whence

$$F(w) - F(v) \le \frac{a}{a+\delta}[F(z) - F(v)] \le \frac{a}{\delta} 2M = \frac{2M}{\delta}\|v - w\|. \qquad \square$$

Whenever F is convex and continuous at a point, the latter result applies.

Corollary 1.4. *If $F : \mathbf{R}^N \to]-\infty, +\infty]$ is convex, then it is locally Lipschitz continuous in* $\mathrm{int}(\mathrm{Dom}(F))$.

Proof. $\mathrm{int}(\mathrm{Dom}(F))$ contains an *N-simplex* S, i.e., a set of convex combinations of $N+1$ affinely independent points of $\mathrm{int}(\mathrm{Dom}(F))$. By the convexity, F is bounded in the interior of S, which is nonempty. It then suffices to apply the latter theorem. $\square$

*** Corollary 1.5.** *If $F \in \Gamma_0(B)$, then it is locally Lipschitz continuous. []*

V.2. The Fenchel transform

Let $F : B \to]-\infty, +\infty]$ be *proper* (i.e., with nonempty domain). We define the *conjugate function* [57]

$$F^*(u^*) := \sup_{u \in B}\{\langle u^*, u\rangle - F(u)\} \ \big(= \sup\{\langle u^*, u\rangle - r : (u, r) \in \mathrm{epi}(F)\}\big) \qquad \forall u^* \in B^*. \quad (2.1)$$

As B^* is a Banach space, for any proper function $G : B^* \to]-\infty, +\infty]$, the conjugate function $G^* : B^{**} \to]-\infty, +\infty]$ might be defined by using the duality pairing between B^* and its dual B^{**}. However, it seems more convenient to deal with the duality pairing between B^* and B, as we did above. We then restrict G^* to B: [58]

$$G^*(u) := \sup_{u^* \in B^*}\{\langle u^*, u\rangle - G(u^*)\} \qquad \forall u \in B. \quad (2.2)$$

If F is as above and F^* is also proper, we introduce the *biconjugate function* of F:

$$F^{**}(u) := \sup_{u^* \in B^*}\{\langle u^*, u\rangle - F^*(u^*)\} \ \big(= \sup\{\langle u^*, u\rangle - r : (u^*, r) \in \mathrm{epi}(F^*)\}\big) \quad \forall u \in B. \quad (2.3)$$

Similarly, if G^* is proper, we define the biconjugate function of a proper function $G : B^* \to]-\infty, +\infty]$:

$$G^{**}(u^*) := \sup_{u \in B}\{\langle u^*, u\rangle - G^*(u)\} \qquad \forall u^* \in B^*. \quad (2.4)$$

Note that

$$F^*(0) = -\inf_{u \in B} F(u), \qquad F^{**}(0) = -\inf_{u^* \in B^*} F^*(u). \quad (2.5)$$

For any proper function $F : B \to]-\infty, +\infty]$, F^* is proper if and only if F has a continuous affine lower bound.

We may then summarize the above definitions as follows:

(i) If $F : B \to]-\infty, +\infty]$ is proper, then $F^* : B^* \to]-\infty, +\infty]$ is defined as in (2.1).

(ii) If $G : B^* \to]-\infty, +\infty]$ is proper, then $G^* : B^{**} \to]-\infty, +\infty]$ is defined. Most often one restricts G^* to B.

---

[57] This has nothing to do with the transposed L^* of a linear operator L.

[58] No confusion should arise denoting this restriction by G^*.

(iii) If $F^* : B^* \to]-\infty, +\infty]$ is proper (i.e., if F has a continuous affine lower bound), then $(F^*)^* : B^{**} \to]-\infty, +\infty]$ is defined. Most often one restricts $(F^*)^*$ to B; we set

$$F^{**} := (F^*)^*\big|_B.$$

(iv) If $G^*\big|_B : B \to]-\infty, +\infty]$ is proper (i.e., if G has a continuous affine lower bound), then $(G^*\big|_B)^* : B^* \to]-\infty, +\infty]$ is defined.

Here is an application of the above setting to economics.

"If we interpret the vector space B as a *commodity space* and accordingly its dual B^* as a *price space,* and if we interpret $F : B \to]-\infty, +\infty]$ as a *cost function* that associates to every commodity $u \in B$ its cost $F(u) \in]-\infty, +\infty]$, then the conjugate function F^* may be interpreted as a *profit function* that associates to every price $u^* \in B^*$ the maximum profit $F^*(u^*) = \sup_{u \in B}\{\langle u^*, u\rangle - F(u)\}$ (since $\langle u^*, u\rangle$ is the value of u when the price system u^* prevails)." [59]

• **Theorem 2.1.** *Let $F : B \to]-\infty, +\infty]$ be proper and have a continuous affine lower bound. Then:*

(i) $F^* \in \Gamma_0(B^*)$ and $F^{**} \in \Gamma_0(B)$. F^* is even weakly star lower semicontinuous.

(ii) (Fenchel-Moreau theorem) *F^{**} is the Γ-regularized function of F; that is,*

$$F^{**}(u) = \sup\{\langle u^*, u\rangle + \alpha : u^* \in B^*, \alpha \in \mathbf{R}, \langle u^*, u\rangle + \alpha \le F(u)\} \qquad \forall u \in B, \tag{2.6}$$

or also $\mathrm{epi}(F^{**}) = \overline{\mathrm{co}}(\mathrm{epi}(F))$. *[Therefore $F^{**} = F$ whenever $F \in \Gamma_0(B)$.]*

(iii) (Fenchel's inequality and dual Fenchel's inequality)

$$\begin{aligned} &\langle u^*, u\rangle \le F(u) + F^*(u^*) \qquad \forall u \in \mathrm{Dom}(F), \forall u^* \in \mathrm{Dom}(F^*), \\ &\langle u^*, u\rangle \le F^{**}(u) + F^*(u^*) \qquad \forall u \in \mathrm{Dom}(F^{**}), \forall u^* \in \mathrm{Dom}(F^*). \end{aligned} \tag{2.7}$$

Proof. (i) is a consequence of part (ii) of Proposition 1.2.

Let us now come to the proof of (2.6). Although in this formula u is kept fixed and u^*, α are varied, here we fix any $u^* \in \mathrm{Dom}(F^*)$ and $\alpha \in \mathbf{R}$, and let u vary. By definition of $F^*(u^*)$ we have

$$\langle u^*, u\rangle + \alpha \le F(u) \quad \forall u \in B \quad \Leftrightarrow \quad \alpha \le -F^*(u^*).$$

That is, the function $L_{u^*} : u \mapsto \langle u^*, u\rangle - F^*(u^*)$ is a continuous and affine lower bound of F, and its constant term, $-F^*(u^*)$, is maximal in the family of these lower bounds. This family of functions is parameterized by u^*; its upper hull, F^{**}, is the Γ-regularized function of F, by definition of Γ-regularization. (ii) thus holds.

The inequalities (2.7) directly follows from the definitions of F^* and F^{**}. □

A function f is said *quasi-convex* [60] if and only if for any $a \in \mathbf{R}$ the *sublevel set* $\{v \in X : f(v) \le a\}$ is convex. It is easy to see that this holds if and only if

$$f(\lambda v_1 + (1-\lambda)v_2) \le \max\{f(v_1), f(v_2)\} \qquad \forall v_1, v_2 \in X, \forall \lambda \in]0,1[. \tag{2.8}$$

Obviously, any convex function is quasi-convex, but the converse may fail. For instance, the real function $x \mapsto \sqrt{|x|}$ is quasi-convex but not convex.

---

[59] From J.-P. Aubin: Applied Functional Analysis, 1979, p. 211.

[60] In the theory of PDEs the same term is also used with a completely different meaning, alas!

Proposition 2.2. *Let B be a Banach space equipped with the norm $\|\cdot\|$, and let B^* be equipped with the dual norm $\|\cdot\|_*$. Let $\varphi \in \Gamma_0(\mathbf{R})$ be even, and set $F(u) := \varphi(\|u\|)$ for any*

$$F(u) := \varphi(\|u\|) \quad \forall u \in B, \qquad G(u^*) := \varphi^*(\|u^*\|_*) \quad \forall u^* \in B^*. \tag{2.9}$$

Then F is convex and lower semicontinuous, and $F^* = G$.

Outline of the proof. By the continuity of the norm and the lower semicontinuity of φ, it is easily checked that F is lower semicontinuous, too. As φ is even, it is nondecreasing. By the convexity of the norm function, is then easily seen that F is convex. Moreover,

$$\begin{aligned} F^*(u^*) &= \sup_{u\in B}\{\langle u^*, u\rangle - F(u)\} = \sup_{u\in B}\{\langle u^*, u\rangle - \varphi(\|u\|)\} \\ &= \sup_{u\in B}\{\|u^*\|_*\,\|u\| - \varphi(\|u\|)\} = \sup_{r\in\mathbf{R}^+}\{r\|u^*\|_* - \varphi(r)\} \\ &= \varphi^*(\|u^*\|_*) = G(u^*) \qquad \forall u^* \in \mathrm{Dom}(F^*). \end{aligned} \tag{2.10}$$ □

Exercises. (i) Show that $I_{co(K)} = \mathrm{co}(I_K)$.

(ii) Exhibit a function $f : \mathbf{R} \to]-\infty, +\infty]$ which has no convex lower bound.

(iii) Set $f(u) := u^2$ for any $u \in \mathbf{R} \setminus \{0\}$, $f(0) := 1$. Check that $\mathrm{epi}(\mathrm{co}(f)) \neq \mathrm{co}(\mathrm{epi}(f))$. Notice that there exists no function $g : \mathbf{R} \to]-\infty, +\infty]$ such that $\mathrm{epi}(g) = \mathrm{co}(\mathrm{epi}(f))$.

(iv) Check that $f : \mathbf{R} \to]-\infty, +\infty]$ is convex if and only if $\mathrm{Dom}(f)$ is convex and $f\big|_{\mathrm{Dom}(f)}$ is also convex.

(v) Check that the pointwise limit of a sequence of convex functions is convex.

(vi) Prove the characterization (2.8) of quasi-convex functions.

(vii) The sum of two quasi-convex functions is necessarily quasi-convex?

(viii) Let K be a subset of a real vector space. Does $\mathrm{co}(K)$ coincide with the set of convex combinations of *pairs* of elements of K?

Further exercises. (i) Let $p, q \in]1, +\infty[$, $1/p + 1/q = 1$. Check that the function $\mathbf{R} \to \mathbf{R} : v^* \mapsto |v^*|^q/q$ is the conjugate of the function $v \mapsto |v|^p/p$. The Fenchel inequality $(2.7)_1$ then generalizes the classical *Young inequality:* $uu^* \leq |u|^p/p + |u^*|^{p'}/p'$ for any $u, u^* \in \mathbf{R}$.

Let Ω be an open subset of $\mathbf{R}^N$. Check that the functional $L^q(\Omega) \to \mathbf{R} : v^* \mapsto (1/q)\int_\Omega |v^*|^q dx$ is the conjugate of the functional $L^p(\Omega) \to \mathbf{R} : v \mapsto (1/p)\int_\Omega |v|^p dx$.

(ii) Check that (2.7) is equivalent to

$$\begin{aligned} &\langle u^*, u\rangle \leq r + s \qquad \forall (u, r) \in \mathrm{epi}(F), \forall (u^*, s) \in \mathrm{epi}(F^*), \\ &\langle u^*, u\rangle \leq r + s \qquad \forall (u, r) \in \mathrm{epi}(F^{**}), \forall (u^*, s) \in \mathrm{epi}(F^*); \end{aligned}$$

moreover, for instance,

$$\mathrm{epi}(F^*) = \{(u^*, s) \in B^* \times \mathbf{R} : \langle u^*, u\rangle \leq r + s,\ \forall (u, r) \in \mathrm{epi}(F)\}.$$

(iii) Under the assumptions of Theorem 2.1, prove that:

(a) $F \leq G$ entails $G^* \leq F^*$ (whenever G fulfils the conditions which we assumed for F);

(b) $F^{***} := (F^{**})^* = (F^*)^{**} = F^*$;

***** (iv) Let F be convex. Show that if F is lower semicontinuous at $u \in B$, then $F(u) = F^{**}(u)$.

***** (v) Let B be a Banach space, $u_0 \in B$, set $F(u) := \|u - u_0\|$ for any $u \in B$ and denote the unit ball of B^* by K^*. Check that $F^*(u^*) = I_{K^*}(u^*) + \langle u^*, u_0\rangle$ for any $u^* \in B^*$.

Let us then set $F_c(u) := F(cu)$ for any $u \in B$ and any $c \in \mathbf{R}$. How may F_c^* be represented? □

V.3. The subdifferential

Let $F : B \to]-\infty, +\infty]$ be proper. We define its *subdifferential*, ∂F, as follows:

$$\partial F(u) := \{u^* \in B^* : \langle u^*, u - v\rangle \geq F(u) - F(v),\ \forall v \in B\} \qquad \forall u \in \mathrm{Dom}(F). \tag{3.1}$$

$\partial F(u) = \emptyset$ is not excluded, and we set $\partial F(u) = \emptyset$ for any $u \in B \setminus \mathrm{Dom}(F)$. We also define the *(effective) domain* of ∂F by $\mathrm{Dom}(\partial F) := \{u \in B : \partial F(u) \neq \emptyset\}$, and say that F is subdifferentiable at u if and only if $\partial F(u) \neq \emptyset$. The elements of $\partial F(u)$ are called *subgradients* of F at u.

The condition (3.1) means that the continuous and affine function $L : v \mapsto \langle u^*, v - u\rangle + F(u)$ is a lower bound of F. Notice that L is *exact* at u, that is, $L(u) = F(u)$.

Dually, if $F^* : B^* \to]-\infty, +\infty]$ is proper, we set

$$\partial F^*(u^*) := \{u \in B : \langle u, u^* - v^*\rangle \geq F^*(u^*) - F^*(v^*),\ \forall v^* \in B^*\} \qquad \forall u^* \in \mathrm{Dom}(F^*), \tag{3.2}$$

and $\partial F^*(u^*) = \emptyset$ for any $u^* \in B^* \setminus \mathrm{Dom}(F^*)$. As above, the duality pairing between B^* and B (rather than between B^* and B^{**}) is here used. This is tantamount to defining $\partial F^*(u^*)$ as a subset of B^{**}, and then to restrict it to B.

Examples. (i) Let H be a (real) Hilbert space, $1 \leq p < +\infty$ and set $F_p(u) := \|u\|^p/p$ for any $u \in H$. It is convenient to identify H with its dual space, and then to define the subdifferential as a subset of H, simply by replacing the duality pairing by the scalar product in the definition. As usual, this turns out to dropping the Riesz isomorphism between H and H^*.

If $p > 1$, then $\partial F_p(u) = \{\|u\|^{p-2}u\}$ for any $u \in H$; in particular, $\partial F_1(u) = \{\|u\|^{-1}u\}$ for any $u \in H \setminus \{0\}$, and $\partial F_1(0) = \{v \in H : \|v\| \leq 1\}$.

In particular, if $H := \mathbf{R}$ then $\partial F_1 = \mathrm{sign}$, where we set

$$\mathrm{sign}(x) := \{-1\} \text{ if } x < 0, \quad \mathrm{sign}(0) := [-1, 1], \quad \mathrm{sign}(x) := \{1\} \text{ if } x > 0. \tag{3.3}$$

(ii) Let $(A, \mathcal{A}, \mu)$ be a measure space, $1 \leq p < +\infty$, $F_p : \mathbf{R} \to \mathbf{R} : v \mapsto |v|^p/p$, and set $\Phi_p(u) := \int_A F_p(u)d\mu$ for any $u \in B := L^p(A, \mathcal{A}, \mu)$. Then

$$\partial\Phi_p(u) = \{u^* \in L^{p'}(A, \mathcal{A}, \mu) : u^* \in \partial F_p(u), \mu\text{-a.e. in } A\} \qquad \forall u \in B.$$

(Here $p' := p/(p-1)$ if $p \neq 1$, $1' := +\infty$, as usual.)

(iii) Let us set $B := \mathbf{R}$ and $F_+(u) = F_-(u) := |u|$ for any $u \in \mathbf{R} \setminus \{0\}$, $F_+(0) := 1$, $F_-(0) := -1$. Then $\partial F_+(u) = \mathrm{sign}(u)$ for any $\mathbf{R} \setminus \{0\}$ and $\partial F_+(0) = \emptyset$. On the other hand, $\partial F_-(u) = \emptyset$ for any $\mathbf{R} \setminus \{0\}$ and $\partial F_-(0) = [-1, 1]$.

(iv) Let Ω be an open subset of $\mathbf{R}^N$ $(N \geq 1)$, $1 < p < +\infty$, either $B := W_0^{1,p}(\Omega)$ or $B := W^{1,p}(\Omega)$, and set

$$F(u) := \frac{1}{p}\int_\Omega |\nabla u|^p dx \qquad \forall u \in B. \tag{3.4}$$

This functional is convex and continuous on the whole B. By examples (i) and (ii), for any $u \in B$, $\xi := |\nabla u|^{p-2}\nabla u$ is the only element of $L^{p'}(\Omega)^N$ such that

$$\int_\Omega \xi \cdot \nabla(u - v)dx \geq \frac{1}{p}\int_\Omega \left(|\nabla u|^p - |\nabla v|^p\right) dx \qquad \forall v \in B. \tag{3.5}$$

Thus, setting $L_u : v \mapsto \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v dx$, $L_u \in B^*$ and $L_u \in \partial F(u)$. If $B := W_0^{1,p}(\Omega)$ then $\partial F(u) = -\nabla \cdot \left(|\nabla u|^{p-2}\nabla u\right)$ in $\mathcal{D}'(\Omega)$. ∂F is thus single-valued.

This holds also for $p = 1$, provided that we replace $|\nabla u|^{p-2}\nabla u$ by any element of $\partial g(\nabla u)$, where $g(\xi) := |\xi|$ for any $\xi \in \mathbf{R}^N$. In this case ∂F is multi-valued.

(v) Let Ω be as above, $1 < p < +\infty$, set $B := L^p(\Omega)$ and

$$\tilde{F}(u) := \begin{cases} \dfrac{1}{p}\displaystyle\int_\Omega |\nabla u|^p dx & \forall u \in W_0^{1,p}(\Omega), \\ +\infty & \forall u \in L^p(\Omega) \setminus W_0^{1,p}(\Omega). \end{cases} \tag{3.6}$$

By (3.5), $\mathrm{Dom}(\partial\tilde{F}) = \{u \in W_0^{1,p}(\Omega) : \nabla \cdot \left(|\nabla u|^{p-2}\nabla u\right) \in L^{p'}(\Omega)\}$ (otherwise the first integral of (3.5) is meaningless) and $\partial\tilde{F}(u) = -\nabla \cdot \left(|\nabla u|^{p-2}\nabla u\right)$ for any $u \in \mathrm{Dom}(\partial\tilde{F})$. For instance, for $p = 2$, $\mathrm{Dom}(\partial\tilde{F}) = H_0^1(\Omega) \cap H^2(\Omega)$ and $\partial\tilde{F}(u) = -\Delta u$ for any $u \in \mathrm{Dom}(\partial\tilde{F})$.

If in (3.6) we replace $W_0^{1,p}(\Omega)$ by $\tilde{B} := W^{1,p}(\Omega)$, the representation of $\partial\tilde{F}$ is more delicate.

(vi) These examples may easily be extended if $|\nabla u|^p$ is replaced by $\varphi(\nabla u)$, where $\varphi : \mathbf{R}^N \to \mathbf{R}$ is convex and $\varphi(\xi)$ grows at infinity at most like $|\xi|^p$ (that is, there exist $C, M > 0$ such that $\varphi(\xi) \le C|\xi|^p + M$ for any $\xi \in \mathbf{R}^N$). □

• **Theorem 3.1.** *Let $F : B \to]-\infty, +\infty]$ and assume that any function of which here we consider either the conjugate or the subdifferential is proper. Then for any $u \in B$ and any $u^* \in B^*$ we have:*

(i) $u^* \in \partial F(u) \Leftrightarrow F(u) + F^*(u^*) = \langle u^*, u\rangle$ (Fenchel's equality), or equivalently:

$u^* \in \partial F(u) \Leftrightarrow$ there exist $(u, r) \in \mathrm{epi}(F)$ and $(u^*, s) \in \mathrm{epi}(F^*)$ such that $r + s = \langle u^*, u\rangle$.

(ii) $u \in \partial F^*(u^*) \Leftrightarrow F^{**}(u) + F^*(u^*) = \langle u^*, u\rangle$ (dual Fenchel's equality), or equivalently:

$u \in \partial F^*(u^*) \Leftrightarrow$ there exist $(u, r) \in \mathrm{epi}(F^{**})$ and $(u^*, s) \in \mathrm{epi}(F^*)$ such that $r + s = \langle u^*, u\rangle$.

(iii) $u^* \in \partial F(u) \Rightarrow u \in \partial F^*(u^*)$. The converse holds if $F(u) = F^{**}(u)$.

(iv) $\partial F(u)$ is convex and weakly star closed (hence strongly and weakly closed); $\partial F^*(u^*)$ is convex and closed.

(v) The operator ∂F is monotone, that is,

$$\langle u_1^* - u_2^*, u_1 - u_2\rangle \ge 0 \qquad \forall u_i \in \mathrm{Dom}(\partial F), \forall u_i^* \in \partial F(u_i) (i = 1, 2). \tag{3.7}$$

(vi) $F(u) = \inf F \Leftrightarrow \partial F(u) \ni 0 \Leftrightarrow \left[F(u) = F^{**}(u), \partial F^{**}(u) \ni 0\right] \Rightarrow u \in \partial F^*(0)$. If $F \in \Gamma_0(B)$ then the latter implication may be inverted.

Proof. By (3.1), $u^* \in \partial F(u)$ if and only if $\langle u^*, v\rangle - F(v) \le \langle u^*, u\rangle - F(u)$ for any $v \in B$, that is,

$$F^*(u^*) := \sup_{v \in B}\{\langle u^*, v\rangle - F(v)\} = \langle u^*, u\rangle - F(u).$$

(i) thus holds. (ii) may be derived similarly.

As $F^{**} \le F$, (iii) follows from (i), (ii) and (2.7).

By (3.1),

$$\partial F(u) = \bigcap_{v \in B} \{u^* \in B^* : \langle u^*, u - v\rangle \ge F(u) - F(v)\};$$

thus $\partial F(u)$ is the intersection of a (nonempty) family of weakly star closed half-spaces. This yields the first part of (iv). The remainder may be proved similarly.

(v) and (vi) easily follow from the definition of subdifferential (cf. Proposition 3.2 below). □

By the Fenchel's inequality, $F(u) + F^*(u^*) \geq \langle u^*, u\rangle$ (which holds for any $(u, u^*) \in B \times B^*$), the Fenchel's equality is equivalent to the opposite Fenchel's inequality, $F(u)+F^*(u^*) \leq \langle u^*, u\rangle$. By the same token, the dual Fenchel's equality is equivalent to the opposite dual Fenchel's inequality: $F^{**}(u) + F^*(u^*) \geq \langle u^*, u\rangle$.

By part (iii) of the latter result, for any $F \in \Gamma_0(B)$, $\partial F^* = (\partial F)^{-1}$. (61)

Corollary 3.1'. *Let $F : B \to]-\infty, +\infty]$ and F^* be proper, and set*

$$\Phi(u, u^*) := F(u) + F^*(u^*) - \langle u^*, u\rangle \qquad \forall u \in \mathrm{Dom}(F), \forall u^* \in \mathrm{Dom}(F^*).$$

Then

$$u^* \in \partial F(u) \quad \textit{if and only if} \quad \Phi(u, u^*) = \inf_{\mathrm{Dom}(F)\times\mathrm{Dom}(F^*)} \Phi. \tag{3.7}'$$

Proof. The Fenchel equality and inequality respectively read

$$u^* \in \partial F(u) \quad \text{if and only if} \quad \Phi(u, u^*) = 0; \qquad 0 \leq \inf_{\mathrm{Dom}(F)\times\mathrm{Dom}(F^*)} \Phi.$$

This yields the "only if" part. Conversely, if $\Phi(u, u^*) = \inf_{\mathrm{Dom}(F)\times\mathrm{Dom}(F^*)} \Phi$, then $\Phi(u, u^*) \leq \Phi(v, u^*)$ for any $v \in B$, whence $u^* \in \partial F(u)$. The "if" part thus holds. (62) □

Note that (3.7)' yields

$$u^* \in \partial F(u) \quad \text{if and only if} \quad \partial_u \Phi(u, u^*) \ni 0, \quad \partial_{u^*} \Phi(u, u^*) \ni 0. \tag{3.7}'$$

The following result may easily be proved via the Fenchel equality and inequality.

Proposition 3.2. *Under the assumptions of Theorem 3.1 we have:*
(i) If $\partial F(u) \neq \emptyset$, then $F(u) = F^{**}(u)$.
(ii) If $F(u) = F^{**}(u)$, then $\partial F(u) = \partial F^{**}(u)$ (possibly $= \emptyset$). [Ex]

The two latter implications cannot be inverted in general. As counterexamples take, e.g., $B = \mathbf{R}$ and respectively $F_1(x) := +\infty$ for any $x < 0$, $F_1(x) := -\sqrt{x}$ for any $x \geq 0$, $F_2(x) := F_1(x)$ for any $x \neq 0$, $F_2(0) := 1$. In either case consider the point $x = 0$.

• **Theorem 3.3.** *(Rockafellar) Let $F_1, F_2 : B \to]-\infty, +\infty]$. Then*

$$\partial F_1(u) + \partial F_2(u) \subset \partial(F_1 + F_2)(u) \qquad \forall u \in \mathrm{Dom}(F_1) \cap \mathrm{Dom}(F_2). \tag{3.8}$$

The opposite inclusion holds if $F_1, F_2 \in \Gamma_0(B)$, and either F_1 or F_2 is continuous at some point $u_0 \in \mathrm{Dom}(F_1) \cap \mathrm{Dom}(F_2)$. (63)

Proof. To check (3.8) it suffices to write the definition of subdifferential for F_1 and F_2, and sum the two inequalities.

---

(61) The inverse of any multi-valued function $f : A \to 2^B$ is defined as follows: for any $(a, b) \in A \times B$, $a \in f^{-1}(b)$ if and only if $b \in f(a)$. For multi-valued functions, this is not always equivalent to the property $f^{-1} \circ f = f \circ f^{-1} = Id$.

(62) This is just another way of proving the Fenchel equality.

(63) By Corollary 1.5, this hypothesis is equivalent to $[\mathrm{int}(\mathrm{Dom}(F_1)) \cap \mathrm{Dom}(F_2)] \cup [\mathrm{Dom}(F_1) \cap \mathrm{int}(\mathrm{Dom}(F_2))] \neq \emptyset$.

To prove the opposite inclusion, let us assume that, e.g., F_1 fulfils the continuity hypothesis. Let $u^* \in \partial(F_1 + F_2)(u)$, that is,

$$\langle u^*, u - v\rangle \geq F_1(u) + F_2(u) - F_1(v) - F_2(v) \qquad \forall v \in \mathrm{Dom}(F_1) \cap \mathrm{Dom}(F_2), \qquad (3.8)'$$

and set

$$A_1 := \{(v, r) \in B \times \mathbf{R} : F_1(u) - F_1(v) > r\},$$
$$A_2 := \{(v, r) \in B \times \mathbf{R} : F_2(v) - F_2(u) + \langle u^*, u - v\rangle \leq r\}.$$

These sets are nonempty, convex and disjoint, by (3.8)'. A_1 is open, by the continuity hypothesis and by Theorem 1.3 (applied to the function $(v, r) \mapsto F_1(v) + r$). Hence, by Theorem II.2.10, A_1 and A_2 may be separated by a closed affine hyperplane. The latter is nonvertical, i.e., it is not of the form $\langle w^*, v\rangle = 0$, since otherwise $\mathrm{Dom}(F_1) \cap \mathrm{Dom}(F_2) = \emptyset$. Thus there exists $(w^*, s) \in B^* \times \mathbf{R}$ such that

$$F_1(u) - F_1(v) \leq \langle w^*, v\rangle - s \leq F_2(v) - F_2(u) + \langle u^*, u - v\rangle \qquad \forall v \in \mathrm{Dom}(F_1) \cap \mathrm{Dom}(F_2).$$

By choosing $v := u$ we get $s = \langle w^*, u\rangle$. Hence $-w^* \in \partial F_1(u)$ and $u^* + w^* \in \partial F_2(u)$. □

The continuity assumption cannot be dropped. As a counterexample take $B = \mathbf{R}$, $F_1(x) := +\infty$ for any $x < 0$, $F_1(x) := -\sqrt{x}$ for any $x \geq 0$, $F_2 := I_{]-\infty,0]}$. Hence $(F_1 + F_2)(0) = 0$, $(F_1 + F_2)(x) = +\infty$ for any $x \neq 0$. Therefore $\partial(F_1 + F_2)(0) = \mathbf{R}$, whereas $\partial F_1(0) + \partial F_2(0) = \emptyset + [0, +\infty[= \emptyset$.

Proposition 3.4. *Let B_1, B_2 be Banach spaces over $\mathbf{R}$, $L : B_1 \to B_2$ be linear and continuous, and $F \in \Gamma(B_2)$. Then $F \circ L \in \Gamma(B_1)$.*

Moreover, if F is continuous at some point $\bar{p} \in B_2$, then

$$\partial(F \circ L)(u) = (L^* \circ \partial F)(Lu) \qquad \forall u \in B_1. \qquad [] \qquad (3.9)$$

• **Proposition 3.5.** *Let $F : B \to]-\infty, +\infty]$ be lower semicontinuous at some $u \in B$, and $\{(u_n, u_n^*)\}$ be a sequence in $B \times B^*$. If*

$$u_n^* \in \partial F(u_n) \ \ \forall n \in \mathbf{N}, \quad u_n \to u \ \text{ weakly in } B, \quad u_n^* \to u^* \ \text{ weakly star in } B^*, \qquad (3.10)$$

$$\liminf_{n\to\infty} \langle u_n^*, u_n\rangle \leq \langle u^*, u\rangle, \qquad (3.11)$$

then $u^* \in \partial F(u)$.

Proof. For any n, $u_n^* \in \partial F(u_n)$ if and only if

$$\langle u_n^*, u_n - v\rangle \geq F(u_n) - F(v) \qquad \forall v \in \mathrm{Dom}(F).$$

It then suffices to pass to the inferior limit as $n \to \infty$. □

The latter result entails that the operator ∂F is strongly-weakly star and weakly-strongly sequentially closed in $B \times B^*$.

Proposition 3.6. *Let $F : B \to]-\infty, +\infty]$ be convex. If F has a point of continuity, then* $\mathrm{int}(\mathrm{Dom}(F)) \subset \mathrm{Dom}(\partial F) (\subset \mathrm{Dom}(F))$.

Proof. If F is continuous at some point, then it is continuous at any interior point of $\mathrm{Dom}(F)$, by Theorem 1.2. Let us fix any $u_0 \in \mathrm{int}(\mathrm{Dom}(F))$. As $\mathrm{int}(\mathrm{epi}(F))$ is nonempty

and $(u_0, F(u_0)) \in \mathrm{int}(\mathrm{epi}(F))$, by Theorem II.2.10 there exists a closed hyperplane through $(u_0, F(u_0))$ which is tangent to $\mathrm{epi}(F)$. Thus there exists $(u^*, r) \in (B^* \times \mathbf{R}) \setminus \{(0,0)\}$ such that

$$ {}_{B^*\times\mathbf{R}}\langle (u^*, r), (u - u_0, a - F(u_0))\rangle_{B\times\mathbf{R}} = {}_{B^*}\langle u^*, u - u_0\rangle_B + r[a - F(u_0)] \geq 0 \quad \forall (u, a) \in \mathrm{int}(\mathrm{epi}(F)). $$

We have $r \neq 0$, since otherwise ${}_{B^*}\langle u^*, u - u_0\rangle_B \geq 0$ for any $u \in \mathrm{int}(\mathrm{Dom}(F))$, whence $u^* = 0$ as $u_0 \in \mathrm{int}(\mathrm{Dom}(F))$. Moreover $r > 0$, as a may be arbitrarily large. Taking $a := F(u)$, we get ${}_{B^*}\langle -u^*/r, u_0 - u\rangle_B \geq F(u_0) - F(u)$ for any $u \in \mathrm{Dom}(F)$, that is, $-u^*/r \in \partial F(u_0)$. □

Proposition 3.7. *If $F \in \Gamma_0(B)$ then $\mathrm{Dom}(\partial F)$ is dense in $\mathrm{Dom}(F)$.*

This result will be proved in Sect. XIII.4.

Exercises. (i) Let $F \in \Gamma_0(B)$ and $u^* \in B^*$. Check that $\partial F^*(u^*)$ coincides with the set of points which minimize the function $u \mapsto F(u) - \langle u^*, u\rangle$.

(ii) Show by a counterexample that in general $u \in \partial F^*(u^*)$ does not entail $u^* \in \partial F(u)$.

(iii) Evaluate F_i^*, F_i^{**}, ∂F_i, ∂F_i^*, ∂F_i^{**}, for any F_i defined as follows:

(a) $F_1(x) := |x|^2$ for any $x \in \mathbf{R} \setminus \{0\}$, $F_1(0) := -1$.

(b) $F_2(x) := |x|^2$ for any $x \in \mathbf{R} \setminus \{0\}$, $F_2(0) := 1$.

(c) $F_3(x) := -|x|^2$ for any $x \in \mathbf{R}$.

(d) $F_4(x) := \arctan x$ for any $x \in \mathbf{R}$.

(e) $F_5(x) := -x$ for any $x < 0$, $F_5(x) := -\sqrt{x}$ for any $x \geq 0$.

Hint: In order to evaluate the conjugate, one may use part (iii) of Theorem 3.1.

(iv) Let $F : B \to]-\infty, +\infty]$ $(F \not\equiv +\infty)$ and $u \in K \subset B$. Check that the two following properties are equivalent:

(a) $F(u) = \inf_K F$,

(b) $0 \in \partial(F + I_K)(u)$.

If $F \in \Gamma_0(B)$, K is closed and convex, and $[\mathrm{int}(\mathrm{Dom}(F)) \cap K] \cup [\mathrm{Dom}(F) \cap \mathrm{int}(K)] \neq \emptyset$, then (a) and (b) are also equivalent to

(c) $0 \in \partial F(u) + \partial I_K(u)$, i.e., $-\partial I_K(u) \cap \partial F(u) \neq \emptyset$.

(v) Consider the elementary statements

$$ \frac{1}{2}x^2 + \frac{1}{2}y^2 \geq xy \qquad \forall x, y \in \mathbf{R}, $$

$$ \frac{1}{2}x^2 + \frac{1}{2}y^2 = xy \quad \text{if and only if} \quad x = y \qquad \forall x, y \in \mathbf{R}. $$

Show that they respectively express the Fenchel inequality and the Fenchel equality for the function $x \mapsto \frac{1}{2}x^2$ in $\mathbf{R}$. Apply then the Fenchel inequality and the Fenchel equality to the function $x \mapsto \frac{1}{2}x^p$ in $\mathbf{R}$ for any $p \in]1, +\infty[$.

(vi) Show that the Fenchel inequality and equality entail the monotonicity of the subdifferential.

Inverse Fenchel equality and inequality: for any $u^* \in B^*$ and any function $F : B \to \mathbf{R} \cup \{+\infty\}$,

$$ \exists \xi \in \mathbf{R} : \begin{cases} \exists u \in B : & \langle u^*, u\rangle = F(u) + \xi \\ \forall v \in B, & \langle u^*, v\rangle \leq F(v) + \xi \end{cases} \quad \Rightarrow \quad u^* \in \partial F(u), \xi = F^*(u^*). \tag{3.12} $$

(The opposite implication coincides with the usual Fenchel equality and inequality.) Obviously the equality on the left may equivalently be replaced by $\langle u^*, u\rangle \geq F(u) + \xi$.

In particular, this entails the following statement:

for any $(u, u^*) \in B \times B^*$ and any functions $F : B \to \mathbf{R} \cup \{+\infty\}$ and $G : B^* \to \mathbf{R} \cup \{+\infty\}$,

$$\begin{cases} \langle u^*, u\rangle = F(u) + G(u^*) \\ \forall (v, v^*) \in B \times B^*,\ \langle v^*, v\rangle \leq F(v) + G(v^*) \end{cases} \quad \Rightarrow \quad \begin{cases} u^* \in \partial F(u), G(u^*) = F^*(u^*) \\ u \in \partial G(u^*), F(u) = G^*(u). \end{cases} \tag{3.13}$$

References

[Br] B. van Brunt: *The Calculus of Variations.* Springer 2004

[BuGiHi] G. Buttazzo, M. Giaquinta, S. Hildebrandt: *One-dimensional variational problems. An introduction.* The Clarendon Press and Oxford University Press, New York, 1998

[De] A. Defranceschi: *Note del corso di Calcolo delle Variazioni.*
http://latemar.science.unitn.it/segue/index.php? &site=2011CalcoloVariazioni&
section=149&action=site

[Ev] L.C. Evans: *Introduction to optimal control theory.*
http://math.berkeley.edu/ evans/

[FaMa] A. Fasano, S, Marmi: *Meccanica Analitica.* Boringhieri, Torino

[GeFo] I.M. Gelfand, S.V. Fomin: *Calculus of Variations.* Dover Publications Inc.

[GiMo] M. Giaquinta, G. Modica: *Analisi Matematica. V. Funzioni of più variabili: ulteriori sviluppi.* Pitagora, Bologna 2005

[HiTr] S. Hildebrandt, A. Tromba: *Principi di minimo. Forme ottimali in natura.* Edizioni della Normale, Pisa 2007 [euristico e discorsivo, assolutamente consigliato]

[Ho] D.D. Holm: *Geometric Mechanics.* Imperial College Press, London 2011

[MaSt] J. Macki, A. Strauss: *Introduction to optimal control theory.* Springer 1982

[Mo] V. Moretti: *Fisica matematica I: Elementi di meccanica razionale, meccanica analitica e teoria della stabilità.* Dipartimento of Matematica dell'Università of Trento

[Ro] A. Romano: *Meccanica razionale.* (2 volumi, il primo con G. Starita) Liguori, Napoli 1995 [il volume 1 raccoglie un'ampia gamma di strumenti analitici e soprattutto geometrici]

[Tr] J. Troutman: *Variational calculus and optimal control.* Springer 1996

[Wi] Wikipedia, the free encyclopedia [da avvicinare con le dovute cautele].