

## Chapter 2

# The Pontryagin Maximum Principle: From Necessary Conditions to the Construction of an Optimal Solution

We now proceed to the study of a finite-dimensional optimal control problem, i.e., a dynamic optimization problem in which the state of the system,  $x = x(t)$ , is linked in time to the application of a control function,  $u = u(t)$ , by means of the solution to an ordinary differential equation whose right-hand side is shaped by the control. We now consider multidimensional systems in which both the state and the control variables no longer need to be scalar. In particular, the results presented here also provide high-dimensional generalizations for the classical theorems of the calculus of variations developed in Chap. 1. So far, we have considered only the simplest problem in the calculus of variations in which the functional is minimized over all curves that satisfy prescribed boundary conditions. Much more than in the calculus of variations, an optimal control problem is determined by its *constraints*. Of these, the most important one is represented by the *dynamics*, which in this text will always be given by an ordinary differential equation,

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

and the optimization is carried out over a subset of solutions to this differential equation, so-called *admissible controlled trajectories*, not just simply over all differentiable curves. In most optimal control problems, the controls are required to satisfy *control constraints* in the form

$$u \in U$$

requiring that the control function  $u(t)$  take values in a prescribed set  $U$  at (almost) all times  $t$ . This set  $U$  is called the control set and in our formulations will always be taken as a subset of  $\mathbb{R}^m$ , but otherwise arbitrary. For example, the choice  $U = \{u_1, \dots, u_r\}$  would define a control system that is allowed to switch between  $r$  possible settings. We also consider *terminal constraints* of the form

$$(T, x(T)) \in N,$$

where  $T$  denotes the final time on the trajectory and  $N$  is a subset of the combined time–state space  $\mathbb{R} \times \mathbb{R}^n$ . Restrictions on the final time  $T$ , for example a fixed terminal time, will be included in this constraint. We shall impose assumptions that make  $N$  a “nice” geometric object. Many more types of constraints are conceivable and occur in real systems. For example, state-space constraints restrict the state of the system from entering prohibited regions. Mixed control-state constraints are simultaneous requirements on the state and control in the sense that if the state of the system has a specific value, then only a limited choice of control actions is available. These clearly are realistic and important scenarios. However, the inclusion of constraints of this type leads to a more complex theory, and in this text we restrict our treatment to what is a *finite-dimensional optimal control problem with control and terminal constraints*. Given these constraints, we then consider an *objective* of the form

$$\mathcal{J}(u) = \int_{t_0}^T L(s, x(s), u(s)) ds + \varphi(T, x(T))$$

with the integral representing the running cost along the controlled trajectory and the function  $\varphi$  defined on  $N$  defining a penalty term on the final state. A precise problem formulation including all assumptions will be given in Sect. 2.2, which also contains a statement of the main necessary conditions for optimality, the Pontryagin maximum principle [193].

The rest of the chapter will then be devoted to illustrating the use of this result, with the proof deferred until Chap. 4. Among the illustrations we provide, we include a statement of the necessary conditions for optimality for the calculus of variations problem in  $\mathbb{R}^n$  (Sect. 2.3), the classical linear-quadratic regulator (Sects. 2.1 and 2.4), several examples of optimal solutions for the time-optimal control problem to the origin in  $\mathbb{R}^2$  for time-invariant linear systems (Sects. 2.5 and 2.6) and some classical examples of optimal control problems with a time-varying or nonlinear dynamics (Sect. 2.7). General properties of optimal solutions for the time-optimal control problem for nonlinear systems that are affine functions of the control(s) will be developed in Sect. 2.8, which provides an introduction to some of the Lie derivative-based techniques that form the basis for geometric methods in optimal control. This section also includes a discussion of singular controls and additional necessary conditions for optimality of the corresponding controlled trajectories, such as the Legendre–Clebsch condition. We then use the developed theory to analyze some generic cases for the time-optimal control problem in the plane (Sects. 2.9 and 2.10). These results, due to H. Sussmann [230, 236], serve as a first illustration of the power of geometric methods in the solution of optimal control problems. We close this chapter (Sect. 2.11) with a derivation of the optimal controls for the Fuller problem, a classical optimal control problem whose solutions are given by chattering arcs, i.e., the associated controls switch infinitely often on a finite interval and are no longer piecewise continuous.

In this chapter, the emphasis is on illustrating the use of the necessary conditions for optimality of the maximum principle. We simplify the presentation by making the mathematically unjustified, but in practical problems often satisfied, assumption

that optimal controls are piecewise continuous. With only minor modifications, all the results presented in this chapter remain valid for the more general class of locally bounded Lebesgue measurable functions, and in subsequent chapters we then shall work with this, for our purpose, adequately general class of controls.

We close these introductory comments with establishing our notation. The equations of the maximum principle and many of the involved computations can be written in a concise and elegant form that avoids the use of matrix transpositions so common in the classical textbooks if a proper notation is established. In this chapter, our state space will always be  $\mathbb{R}^n$  or some open subset of it, and we write the state  $x$  as a column vector. However, in our notation, we already here distinguish between what in the formulation on manifolds will be tangent vectors, which we write as column vectors, and cotangent vectors (or covectors for short), which we write as row vectors. For example,  $\dot{x}$  is a tangent vector, and thus the right-hand side of the dynamics,  $f(t, x, u)$  in the formulations above, is a column vector. On the other hand, geometrically, multipliers  $\lambda$  represent linear functionals and thus are covectors. We denote the space of  $n$ -dimensional covectors or row vectors by  $(\mathbb{R}^n)^*$ , but do not distinguish between  $\mathbb{R}$  and  $\mathbb{R}^*$ . For a scalar continuously differentiable function  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $x \mapsto h(x)$ , we consistently write the gradient with respect to  $x$  as a row vector and denote it by  $\nabla h(x)$  or  $\frac{\partial h}{\partial x}(x)$ , i.e.,

$$\nabla h(x) = \frac{\partial h}{\partial x}(x) = \left( \frac{\partial h}{\partial x_1}(x), \dots, \frac{\partial h}{\partial x_n}(x) \right).$$

For a vector-valued continuously differentiable map  $H$ ,

$$H : \mathbb{R}^k \rightarrow \mathbb{R}^\ell, \quad x \mapsto H(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_k(x) \end{pmatrix},$$

we denote the Jacobian matrix of the partial derivatives of the components  $h_i(x)$  with respect to the variables  $x_j$  by

$$DH(x) = \frac{\partial H}{\partial x}(x) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1}(x) & \dots & \frac{\partial h_1}{\partial x_k}(x) \\ \vdots & \frac{\partial h_i}{\partial x_j}(x) & \vdots \\ \frac{\partial h_k}{\partial x_1}(x) & \dots & \frac{\partial h_k}{\partial x_k}(x) \end{pmatrix}_{1 \leq i, j \leq k},$$

with  $i$  as row index and  $j$  as column index. Thus, the Jacobian matrix is the matrix whose  $i$ th row is given by the gradient of the component  $h_i$ . The Hessian matrix of a twice continuously differentiable function  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $x \mapsto h(x)$ , is the matrix of the second-order partial derivatives of  $h$  and will be denoted by  $D^2h(x) = \frac{\partial^2 h}{\partial x^2}(x)$ . With the convention above, the Hessian of  $h$  is the Jacobian matrix of the transpose of the gradient of  $h$ ,

$$D^2h(x) = \frac{\partial^2 h}{\partial x^2}(x) = \frac{\partial (\nabla h)^T}{\partial x}(x).$$

If  $\Lambda = (\lambda_1, \dots, \lambda_n)$  is a row vector of continuously differentiable functions  $\lambda_j: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $x \mapsto \lambda_j(x)$ ,  $j = 1, \dots, n$ , then, and consistent with the notation just introduced, we denote the matrix of the partial derivatives  $\left( \frac{\partial \lambda_j}{\partial x_i} \right)_{1 \leq i, j \leq n}$  with row index  $i$  and column index  $j$  by  $\frac{\partial \Lambda}{\partial x}$ , that is,

$$\frac{\partial \Lambda}{\partial x}(x) = \left( \frac{\partial \Lambda^T}{\partial x}(x) \right)^T \quad \text{or} \quad D\Lambda(x) = (D(\Lambda^T(x)))^T.$$

Not only does this formalism properly distinguish the different geometric meanings of the variables involved, but it also allows us to write almost all formulas without having to use transposes and simplifies the notation considerably.

Finally, we denote the space of all  $k \times \ell$  matrices of real numbers by  $\mathbb{R}^{k \times \ell}$ . We assume that the reader is familiar with the basic concepts of matrix algebra and recall that a matrix  $P \in \mathbb{R}^{n \times n}$  is *positive semidefinite* if it is symmetric and if  $v^T P v \geq 0$  for any vector  $v \in \mathbb{R}^n$ ;  $P$  is said to be *positive definite* if  $P$  is positive semidefinite and if in addition,  $v^T P v = 0$  holds only for  $v = 0$ . It is well-known from linear algebra that a matrix  $P$  is positive definite/semidefinite if and only if all eigenvalues are positive/nonnegative. Note that as a symmetric matrix,  $P$  has a full set of  $n$  real eigenvalues.

## 2.1 Linear-Quadratic Optimal Control

Before formulating the general optimal control problem, we first fully solve by elementary means what, from an applications point of view, justifiably may be considered the single most important optimal control problem, the so-called *linear-quadratic regulator*. Mathematically, this is but a small extension of the simplest problem in the calculus of variations—neither control nor terminal constraints are imposed—in the sense that the trivial dynamics  $\dot{x} = u$  is replaced by a linear differential equation  $\dot{x} = Ax + Bu$  and the objective to be minimized is a positive definite quadratic form in  $x$  and  $u$ . Standard calculus of variations techniques suffice to solve this problem. In fact, Legendre's idea of “completing the square” presented in Sect. 1.4 works to perfection here and in this section we give an elementary and self-contained derivation of the optimal solution based on Legendre's argument.

The importance of the problem lies in its practical applications. Essentially, this is the problem to regulate a typically nonlinear system around some reference trajectory. In the mathematical formulation below, the reference trajectory and control are normalized to be  $x \equiv 0$  and  $u \equiv 0$ . As such, but also due to the simplicity of its solution and the fact that this solution easily allows the inclusion of stochastic effects (e.g., noisy measurements and estimation of the states from an

incomplete set of measurements by means of the Kalman filter), the linear-quadratic regulator is the theoretical basis for many practical control schemes whose aim is to regulate a system around some set point. Real systems based on this principle range from autopilots in commercial aircraft to advanced stability control systems in cars to standard chemical process control. Naturally, this problem, and its manifold extensions, are the subject of numerous textbooks on automatic control, one of the best still being the classical text by Kwakernaak and Sivan [144]. For this reason, this topic is not in the focus of our presentation in this text, and we shall limit ourselves to its connection with conjugate points and perturbation feedback control for nonlinear optimal control problems. These will be discussed in the context of sufficient conditions for a strong local minimum in Sect. 5.3.

Let  $[0, T]$  be a finite and fixed time horizon and suppose

$$\begin{aligned} A : [0, T] &\rightarrow \mathbb{R}^{n \times n}, \quad t \mapsto A(t), & B : [0, T] &\rightarrow \mathbb{R}^{n \times m} \quad t \mapsto B(t), \\ Q : [0, T] &\rightarrow \mathbb{R}^{n \times n}, \quad t \mapsto Q(t), & R : [0, T] &\rightarrow \mathbb{R}^{m \times m} \quad t \mapsto R(t), \end{aligned}$$

are continuous matrix-valued functions defined on  $[0, T]$ . We assume that the matrices  $Q(t)$  and  $R(t)$  are symmetric and in addition that  $Q(t)$  is positive semidefinite and  $R(t)$  is positive definite for all  $t \in [0, T]$ . Furthermore, let  $S_T \in \mathbb{R}^{n \times n}$  be a constant, symmetric, and positive semidefinite matrix. The *linear-quadratic regulator* then is the following optimal control problem:

[LQ] Find a continuous function  $u : [0, T] \rightarrow \mathbb{R}^m$ , the control, that minimizes a quadratic objective of the form

$$J(u) = \frac{1}{2} \int_0^T [x^T(t)Q(t)x(t) + u^T(t)R(t)u(t)] dt + \frac{1}{2} x^T(T)S_T x(T) \quad (2.1)$$

subject to the linear dynamics

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0. \quad (2.2)$$

It follows from well-known results about ordinary differential equations (see Appendix B) that the initial value problem for the homogeneous linear matrix differential equation

$$\dot{X}(t) = A(t)X(t) \quad \text{and} \quad X(s) = \text{Id}$$

has a unique solution  $\Phi(t, s)$ , called its fundamental solution. For any initial time  $s \in [0, T]$ , this solution exists over the full interval  $[0, T]$ . The unique solution  $x(t; x_0)$  to the homogeneous vector equation  $\dot{x}(t) = A(t)x(t)$  with initial condition  $x(0) = x_0$  is then given by  $x(t; x_0) = \Phi(t, 0)x_0$ , and as is easily verified, the solution to the inhomogeneous equation (2.2) is obtained by variation of constants as

$$x(t; x_0) = \Phi(t, 0) \left( x_0 + \int_0^t \Phi(0, s)B(s)u(s)ds \right).$$

This solution is called the *trajectory corresponding to the control*  $u$ . If the matrix  $A$  is time-invariant, then  $\Phi(t, s)$  is simply given by the absolutely convergent matrix exponential,

$$\Phi(t, s) = \exp(A(t - s)) = \sum_{k=0}^{\infty} \frac{A^k}{k!} (t - s)^k.$$

In the time-varying case, for scalar problems, it is still possible to write down an explicit formula as

$$\Phi(t, s) = \exp\left(\int_s^t A(r) dr\right),$$

but in dimensions  $n \geq 2$  this formula no longer is valid, since generally  $A(t)$  and  $\exp\left(\int_s^t A(r) dr\right)$  do not commute. Series expansions of the solution can still be given in higher dimensions and are related to Lie-algebraic formulas in connection with the Baker–Campbell–Hausdorff formula involving commutators [256] (see also Sect. 4.5), but will not be needed here. The important fact simply is that the fundamental matrix  $\Phi$  exists and is unique.

**Theorem 2.1.1.** *The solution to the linear-quadratic optimal control problem [LQ] is given by the linear feedback control*

$$u_*(t, x) = -R(t)^{-1}B(t)^T S(t)x,$$

where  $S$  is the solution to the Riccati terminal value problem

$$\dot{S} + SA(t) + A^T(t)S - SB(t)R(t)^{-1}B^T(t)S + Q(t) \equiv 0, \quad S(T) = S_T. \quad (2.3)$$

This solution  $S$  exists on the full interval  $[0, T]$  and is positive semidefinite. The minimal value of the objective is given by  $\frac{1}{2}x_0^T S(0)x_0$ .

*Proof.* This is Legendre's argument from the calculus of variations adjusted to this setting. Let  $u : [0, T] \rightarrow \mathbb{R}^m$  be any continuous control and let  $x : [0, T] \rightarrow \mathbb{R}^n$  denote the corresponding trajectory. Dropping the argument  $t$  from the notation, we have for any differentiable matrix function  $S \in \mathbb{R}^{n \times n}$  that

$$\begin{aligned} \frac{d}{dt}(x^T S x) &= \dot{x}^T S x + x^T \dot{S} x + x^T S \dot{x} \\ &= (Ax + Bu)^T S x + x^T \dot{S} x + x^T S(Ax + Bu), \end{aligned}$$

and thus, by adjoining this quantity to the Lagrangian in the objective, we can express the cost equivalently as

$$\begin{aligned} J(u) &= \frac{1}{2} \int_0^T [x^T (Q + A^T S + \dot{S} + SA)x + x^T S B u + u^T B^T S^T x + u^T R u] dt \\ &\quad + \frac{1}{2} x^T(T) [S_T - S(T)] x(T) + \frac{1}{2} x_0^T S(0) x_0. \end{aligned}$$

Take  $S$  as a symmetric matrix and complete the square to get

$$\begin{aligned} J(u) = & \frac{1}{2} \int_0^T [x^T (\dot{S} + SA + A^T S - SBR^{-1}B^T S + Q)x \\ & + (u + R^{-1}B^T Sx)^T R(u + R^{-1}B^T Sx)] dt \\ & + \frac{1}{2} x^T(T) [S_T - S(T)] x(T) + \frac{1}{2} x_0^T S(0) x_0. \end{aligned}$$

For the moment, let us assume that there exists a solution  $S$  to the matrix Riccati equation (2.3) over the full interval  $[0, T]$ . Then the objective simplifies to

$$J(u) = \frac{1}{2} \int_0^T (u + R^{-1}B^T Sx)^T R(u + R^{-1}B^T Sx) dt + \frac{1}{2} x_0^T S(0) x_0.$$

Since the matrix  $R$  is continuous and positive definite over  $[0, T]$ , the minimum is realized if and only if

$$u(t) = -R^{-1}(t)B^T(t)S(t)x(t),$$

and the minimum value is given by

$$\frac{1}{2} x_0^T S(0) x_0.$$

Thus the optimal solution to the linear-quadratic control problem is given as a linear *feedback* function, i.e., a function  $u_* : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^m$  defined in the time–state space, given by

$$u_*(t, x) = -R^{-1}(t)B^T(t)S(t)x.$$

For this argument to be valid, it remains to argue that such a solution  $S$  to the initial value problem (2.3) indeed does exist on all of  $[0, T]$ . It follows from general results about the existence of solutions to ordinary differential equations that such a solution exists on some maximal interval  $(\tau, T]$  and that as  $t \searrow \tau$  (i.e.,  $t \rightarrow \tau$  and  $t > \tau$ ), at least one of the components of the solution  $S(t)$  needs to diverge to  $+\infty$  or  $-\infty$ . For if this were not the case, then by the local existence theorem on ODEs, the solution could be extended further onto some small interval  $(\tau - \varepsilon, \tau + \varepsilon)$ , contradicting the maximality of the interval  $(\tau, T]$ . In general, however, this explosion time  $\tau$  could be nonnegative, invalidating the argument above. That this is not the case for the linear-quadratic regulator problem is a consequence of the positivity assumptions on the objective, specifically, the definiteness assumptions on the matrices  $R$ ,  $Q$ , and  $S_T$ .

In order to see this, suppose the explosion time  $\tau$  of the solution to the Riccati equation is nonnegative,  $\tau \geq 0$ , and consider the linear-quadratic regulator problem for variable initial conditions  $(t_0, x_0) \in [0, T] \times \mathbb{R}^n$ . If  $t_0 > \tau$ , then the reasoning above is valid; thus the solution to the minimization problem [LQ] is given by the feedback control  $u_*(t, x)$ , and the minimal value is  $J(u_*) = \frac{1}{2} x_0^T S(t_0) x_0$ .

This holds for arbitrary initial conditions  $x_0$ . Since  $J(u)$  is always nonnegative by our assumptions on the matrices in the objective, the matrix  $S(t_0)$  must be positive semidefinite. But we can choose  $t_0$  arbitrarily in the interval  $(\tau, T]$ , and thus it follows that the matrix  $S(t)$  is positive semidefinite on this interval. Furthermore, since for any other control  $u$  defined on  $[t_0, T]$  we have that

$$J(u) \geq \frac{1}{2} x_0^T S(t_0) x_0,$$

using the control  $u \equiv 0$ , we obtain an upper bound in the form

$$0 \leq \frac{1}{2} x_0^T S(t_0) x_0 \leq \frac{1}{2} x_0^T \left( \int_{t_0}^T \Phi(t, t_0)^T Q(t) \Phi(t, t_0) dt + \Phi(T, t_0)^T S_T \Phi(T, t_0) \right) x_0 \quad (2.4)$$

for every  $x_0 \in \mathbb{R}^n$ . Choosing for the initial condition  $x_0$  the  $i$ th coordinate vectors,  $e_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ , with the 1 in the  $i$ th position, the lower estimate in Eq. (2.4) gives  $S_{ii}(t_0) \geq 0$ . The upper estimate is continuous in  $t_0$  on the full interval  $[0, T]$  and thus remains bounded over the full interval. Hence there exists a positive constant  $C$  such that

$$0 \leq S_{ii}(t_0) \leq C \quad \text{for all } t_0 \in (\tau, T].$$

Choosing  $x_0 = e_i \pm \theta e_j$ , we furthermore obtain

$$0 \leq (e_i \pm \theta e_j)^T S(t_0) (e_i \pm \theta e_j) = S_{ii}(t_0) + 2\theta S_{ij}(t_0) + \theta^2 S_{jj}(t_0)$$

for all  $\theta \in \mathbb{R}$ , which is equivalent to

$$S_{ij}^2(t_0) \leq S_{ii}(t_0) S_{jj}(t_0).$$

But then all entries  $S_{ij}(t_0)$  of the matrix  $S(t_0)$  take values in the interval  $[-C, C]$  for all times  $t_0$  from the interval  $(\tau, T]$ . Hence there cannot be an explosion of the solution as  $t_0 \searrow \tau$ . This contradicts the fact that  $(\tau, T]$  is the maximal interval of existence for the solution  $S$  of Eq. (2.3). Thus we must have  $\tau < 0$ , and the solution to the Riccati equation exists over the full interval  $[0, T]$ .  $\square$

## 2.2 Optimal Control Problems

We now formulate the optimal control problem to be considered in this text and introduce the main necessary conditions for optimality, the Pontryagin maximum principle [193].



### 2.2.1 Control Systems

We think of a control system as a collection of time-dependent vector fields on a differentiable manifold parameterized by controls that by means of the solutions of the corresponding ordinary differential equations, give rise to a family of controlled trajectories. An optimal control problem then is the task to minimize some functional over these controlled trajectories subject to additional constraints. We shall postpone a precise definition along these lines until Chap. 4, where we actually prove the maximum principle. Here, in view of the still introductory character of this chapter, we retain the more elementary formulation of optimal control problems with state space  $\mathbb{R}^n$ . However, we already arrange the material according to this framework.

**Definition 2.2.1.** A **control system** is a 4-tuple  $\Sigma = (M, U, f, \mathcal{U})$  consisting of a state space  $M$ , a control set  $U$ , a dynamics  $f$ , and a class  $\mathcal{U}$  of admissible controls.

Throughout this chapter, we make the following assumptions about the data defining the control system:

1. The *state space*  $M$  is an open and connected subset of  $\mathbb{R}^n$ .
2. The *control set*  $U$  is a subset of  $\mathbb{R}^m$ . No further regularity conditions on the structure of  $U$  need to be imposed, although in many practical situations  $U$  is compact and convex.
3. The *dynamics*  $\dot{x} = f(t, x, u)$  is defined by a family of time-varying vector fields  $f$  parameterized by the control values  $u \in U$ ,

$$f : \mathbb{R} \times M \times U \rightarrow \mathbb{R}^n, \quad (t, x, u) \mapsto f(t, x, u),$$

i.e.,  $f$  assigns to every point  $(t, x, u) \in \mathbb{R} \times M \times U$  a (tangent) vector  $f(t, x, u) \in \mathbb{R}^n$ . We assume that the time-varying vector fields are continuous in  $(t, x, u)$ , differentiable in  $x$  for fixed  $(t, u) \in \mathbb{R} \times U$ , and that the partial derivatives  $\frac{\partial f}{\partial x}(t, x, u)$  are continuous as a function of all variables; no differentiability assumptions in the control variable  $u$  are made.

4. The class  $\mathcal{U}$  of *admissible controls* is taken to be piecewise continuous functions  $u$  defined on a compact interval  $I \subset \mathbb{R}$  with values in the control set  $U$ . Without loss of generality, we assume that controls are continuous from the left.

These specifications are simplifications of the setting considered in Chap. 4. Here our aim is to formulate the fundamental necessary conditions for optimality and then to illustrate how these conditions can be put to work. For this, the simpler framework formulated above that requires only some knowledge of advanced calculus and ordinary differential equations is adequate, and it simplifies the technical aspects of the theory. In the more general framework considered in Chap. 4, the state space  $M$  will be a  $C^r$ -manifold, and the class  $\mathcal{U}$  of admissible controls will consist of all locally bounded Lebesgue measurable functions  $u$  that take values in the control set  $U$ , i.e., given a compact interval  $I \subset \mathbb{R}$ , there exists a compact subset

$V$  of  $U$  such that  $u$  takes values in  $V$  almost everywhere on  $I$ . In particular, if the control set  $U$  already is compact, then admissible controls are simply Lebesgue measurable functions that take values in  $U$  almost everywhere. The need for taking as admissible controls the class of Lebesgue measurable functions lies in the fact that the class of piecewise continuous controls simply is too small, and this will already be seen in Sect. 2.11 of this chapter, to guarantee the existence of optimal solutions. Greater generality is required for several important and fundamental results to be valid. Locally bounded Lebesgue measurable functions are pointwise limits of piecewise continuous functions and provide the required closure properties needed for many arguments. A brief exposition of Lebesgue measurable functions is given in Appendix D, but this will be needed only in Chaps. 3, 4, and some of 6. Similarly, many control systems, especially those connected with mechanical systems (e.g., robotic manipulators) have natural state-space descriptions that are manifolds. Clearly, the circle  $S^1$  is a far superior model for the state space of a fixed-amplitude oscillation than  $\mathbb{R}^2$ . The sphere  $S^2$  is the only reasonable model to calculate the shortest air route from Paris to Sydney. But these generalizations will be considered only in Chap. 4.

In the same spirit, we always impose conditions on the dynamics that for a given admissible control, guarantee not only the existence of solutions to the differential equation, but also its uniqueness. From an engineering perspective, this is as important<sup>1</sup> a condition as existence of solutions, and we will insist on it being satisfied. Using the practical class of piecewise continuous controls in this chapter suffices for our arguments and simplifies the reasoning. Given any piecewise continuous control  $u \in \mathcal{U}$  defined over some open interval  $J$ , it follows from standard local existence and uniqueness results for ordinary differential equations (see Appendix B) that for any initial condition  $x(t_0) = x_0$  with  $t_0 \in J$ , there exists a unique solution  $x$  to the initial value problem

$$\dot{x}(t) = f(t, x, u(t)), \quad x(t_0) = x_0, \quad (2.5)$$

defined over some maximal interval  $(\tau_-, \tau_+) \subset J$  that contains  $t_0$ .

**Definition 2.2.2 (Admissible controlled trajectory).** Given an admissible control  $u \in \mathcal{U}$  defined over an interval  $J$ , let  $x$  be the unique solution to the initial value problem (2.5) with maximal interval of definition  $I = (\tau_-, \tau_+)$ . We call this solution  $x$  the trajectory corresponding to the control  $u$  and call the pair  $(x, u)$  an admissible controlled trajectory over the interval  $I$ .

An optimal control problem then consists in finding, among all admissible controlled trajectories, one that minimizes an objective, possibly subject to additional constraints. In this text, in addition to the control constraints that are implicit in the definition of the control set, we consider only **terminal constraints** in the form of a target set into which the controls need to steer the system. However, we restrict the

---

<sup>1</sup>From our point of view, uniqueness may be the more important of the two conditions.

terminal set to have the regular geometric structure of a  $k$ -dimensional embedded submanifold  $N$  in  $\mathbb{R} \times M$  (see Appendix C). More specifically, we assume that

$$N = \{(t, x) \in \mathbb{R} \times M : \Psi(t, x) = 0\},$$

where  $\Psi : \mathbb{R} \times M \rightarrow \mathbb{R}^{n+1-k}$ ,  $(t, x) \mapsto \Psi(t, x) = (\psi_0(t, x), \dots, \psi_{n-k}(t, x))^T$ , is a continuously differentiable mapping and the matrix  $D\Psi$  of the partial derivatives with respect to  $(t, x)$  is of full rank  $n+1-k$  everywhere on  $N$ , i.e., the gradients of the functions  $\psi_0(t, x), \dots, \psi_{n-k}(t, x)$  are linearly independent on  $N$ .

Finally, the **objective** is given in so-called Bolza form as the integral of a Lagrangian  $L$  plus a penalty term  $\varphi$ . For the Lagrangian we make the same regularity assumptions as on the dynamics  $f$ , i.e., the function

$$L : \mathbb{R} \times M \times U \rightarrow \mathbb{R}, \quad (t, x, u) \mapsto L(t, x, u),$$

is continuous in  $(t, x, u)$ , differentiable in  $x$  for fixed  $(t, u) \in \mathbb{R} \times U$ , and the derivative  $\frac{\partial L}{\partial x}(t, x, u)$  is continuous as a function of all variables. The penalty term  $\varphi$  is given by a continuously differentiable function

$$\varphi : \mathbb{R} \times M \rightarrow \mathbb{R}, \quad (t, x) \mapsto \varphi(t, x).$$

Clearly, this function needs to be defined only on  $N$ . Since we assume that  $N$  is an embedded submanifold of  $\mathbb{R}^{n+1}$ , if necessary, we can always extend  $\varphi$  to a differentiable function  $\varphi : \mathbb{R} \times M \rightarrow \mathbb{R}$  locally, and thus for simplicity we assume that  $\varphi$  is defined in the ambient state space. The objective or cost functional is then given as

$$\mathcal{J}(u) = \int_{t_0}^T L(s, x(s), u(s)) ds + \varphi(T, x(T)), \quad (2.6)$$

where  $x$  is the unique trajectory corresponding to the control  $u$ . The terminal time  $T$  can be fixed or free. A fixed terminal time simply will be modeled as the equation  $\varphi_0(t, x) = t - T$  in the mapping  $\Psi$  defining the constraint in  $N$ . The initial time  $t_0$  and initial condition  $x_0$  are fixed, but arbitrary. Then the optimal control problem is the following one:

[OC] Minimize the objective  $\mathcal{J}(u)$  over all admissible controlled trajectories  $(x, u)$  defined over an interval  $[t_0, T]$  that satisfy the terminal constraint  $(T, x(T)) \in N$ .

### 2.2.2 The Pontryagin Maximum Principle

The maximum principle of optimal control gives the fundamental necessary conditions for a controlled trajectory  $(x, u)$  to be optimal. It was developed in the mid 1950s in the Soviet Union by a group of mathematicians under the leadership

of L.S. Pontryagin, also including V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko, and is known as the Pontryagin maximum principle [41, 193]. Below, and consistent with our choice of admissible controls, we give its formulation under the additional assumption that the optimal control is piecewise continuous. Recall that we write tangent vectors as column vectors and cotangent vectors (i.e., multipliers) as row vectors.

**Definition 2.2.3 (Hamiltonian).** The (control) *Hamiltonian* function  $H$  of the optimal control problem [OC] is defined as

$$H : \mathbb{R} \times [0, \infty) \times (\mathbb{R}^n)^* \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$$

with

$$H(t, \lambda_0, \lambda, x, u) = \lambda_0 L(t, x, u) + \lambda f(t, x, u). \quad (2.7)$$

**Theorem 2.2.1 (Pontryagin maximum principle).** [193] *Let  $(x_*, u_*)$  be a controlled trajectory defined over the interval  $[t_0, T]$  with the control  $u_*$  piecewise continuous. If  $(x_*, u_*)$  is optimal, then there exist a constant  $\lambda_0 \geq 0$  and a covector  $\lambda : [t_0, T] \rightarrow (\mathbb{R}^n)^*$ , the so-called adjoint variable, such that the following conditions are satisfied:*

1. Nontriviality of the multipliers:  $(\lambda_0, \lambda(t)) \neq 0$  for all  $t \in [t_0, T]$ .
2. Adjoint equation: the adjoint variable  $\lambda$  is a solution to the time-varying linear differential equation

$$\dot{\lambda}(t) = -\lambda_0 L_x(t, x_*(t), u_*(t)) - \lambda(t) f_x(t, x_*(t), u_*(t)). \quad (2.8)$$

3. Minimum condition: everywhere in  $[t_0, T]$  we have that

$$H(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) = \min_{v \in U} H(t, \lambda_0, \lambda(t), x_*(t), v). \quad (2.9)$$

*If the Lagrangian  $L$  and the dynamics  $f$  are continuously differentiable in  $t$ , then the function*

$$h : t \mapsto H(t, \lambda_0, \lambda(t), x_*(t), u_*(t))$$

*is continuously differentiable with derivative given by*

$$\dot{h}(t) = \frac{dh}{dt}(t) = \frac{\partial H}{\partial t}(t, \lambda_0, \lambda(t), x_*(t), u_*(t)). \quad (2.10)$$

4. Transversality condition: at the endpoint of the controlled trajectory, the covector

$$(H + \lambda_0 \varphi_t, -\lambda + \lambda_0 \varphi_x)$$

*is orthogonal to the terminal constraint  $N$ , i.e., there exists a multiplier  $v \in (\mathbb{R}^{n+1-k})^*$  such that*

$$H + \lambda_0 \varphi_t + v D_t \Psi = 0, \quad \lambda = \lambda_0 \varphi_x + v D_x \Psi \quad \text{at } (T, x_*(T)). \quad (2.11)$$

The following statement is an immediate special case.

**Corollary 2.2.1.** *If the Lagrangian  $L$  and the dynamics  $f$  are time-invariant (do not depend on  $t$ ), then the function  $h : t \mapsto H(t, \lambda_0, \lambda(t), x_*(t), u_*(t))$  is constant. If  $\varphi$  and  $\Psi$  also do not depend on  $t$  (and in this case the terminal time  $T$  necessarily is free), then for any multiplier  $(\lambda_0, \lambda)$  that satisfies the conditions of the maximum principle, the Hamiltonian  $H$  vanishes identically along the optimal controlled trajectory  $(x_*, u_*)$ :*

$$H(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) \equiv 0. \quad \square$$

We start our discussions of the maximum principle by introducing some useful terminology and give a brief and somewhat informal description of the significance of each condition.

**Definition 2.2.4 (Extremals; normal and abnormal).** We call controlled trajectories  $(x, u)$  for which there exist multipliers  $\lambda_0$  and  $\lambda$  such that the conditions of the maximum principle are satisfied *extremals*, and the triples  $(x, u, (\lambda_0, \lambda))$  including the multipliers are called *extremal lifts* (to the cotangent bundle in case of manifolds). If  $\lambda_0 > 0$ , then the extremal lift is called *normal* while it is called *abnormal* if  $\lambda_0 = 0$ .

**1. Normal and abnormal extremal lifts.** The maximum principle takes the form of a multiplier rule with multiplier  $(\lambda_0, \lambda(t))$ . The nontriviality condition precludes a trivial solution of these conditions with  $(\lambda_0, \lambda(t)) = (0, 0)$ . Since the conditions are linear in the multipliers  $(\lambda_0, \lambda)$ , it is always possible to normalize this vector. For example, if  $\lambda_0 > 0$ , then the conditions do not change if we divide by  $\lambda_0$  and instead consider as the new multiplier  $(1, \tilde{\lambda}(t))$ , where  $\tilde{\lambda}(t) = \lambda(t)/\lambda_0$ . Thus, without loss of generality, we may always assume that  $\lambda_0 = 1$  if the extremal lift is normal. Note that it is a property of the extremal lift, not the controlled trajectory, to be normal or abnormal. It is possible that both normal and abnormal extremal lifts exist for a given controlled trajectory  $(x, u)$ . For this reason, controlled trajectories for which only abnormal extremal lifts exist are sometimes called *strictly abnormal*. We shall see in Sect. 2.3 that all extremals for the simplest problem in the calculus of variations are normal, and this fact actually is the source of the terminology, which goes back to Carathéodory [67]. In spite of their name, abnormal extremals are by no means pathological situations, and if they exist, they often play an important role in determining the structure of optimal solutions. We shall see in Sect. 2.6 that the synthesis of optimal trajectories for the problem of steering points to the origin time-optimally for the harmonic oscillator with bounded controls, a simple and standard text book example, contains optimal, strictly abnormal extremals and that these play a crucial role in determining the overall structure of the solutions.

**2. Adjoint system.** First note that as a solution to a linear time-varying ordinary differential equation with piecewise continuous entries, the adjoint variable  $\lambda(\cdot)$

exists over the full interval  $[t_0, T]$ . We shall see in the proof of the maximum principle in Sect. 4.2 that  $(\lambda_0, \lambda(t))$  arises as a normal vector to a hyperplane in  $(t, x)$ -space (hence also the nontriviality condition) that evolves in time according to the adjoint equation. This equation arises as the adjoint in the sense of linear ordinary differential equations of the so-called *variational equation*

$$\dot{y} = f_x(t, x_*(t), u_*(t))y, \quad (2.12)$$

which transports tangent vectors (that will be generated by means of variations) along a reference controlled trajectory  $t \mapsto (x_*(t), u_*(t))$ . Solutions of the adjoint system provide the corresponding transport for covectors along this curve. In terms of the Hamiltonian  $H$ , the coupled system consisting of the dynamics and the adjoint equation can be written as

$$\dot{x}_*(t) = \frac{\partial H}{\partial \lambda}(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) \quad \text{and} \quad \dot{\lambda}(t) = -\frac{\partial H}{\partial x}(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) \quad (2.13)$$

and thus forms a *Hamiltonian system* that is coupled with the control  $u_*$  through the minimization condition (2.9).

- 3. Minimum condition.** In the original formulation of the theorem by Pontryagin et al. [193], this condition was formulated as a maximum condition and gave the result its name. In fact, depending on the choice of the signs associated with the multipliers  $\lambda_0$  and  $\lambda$ , the maximum principle can be stated in four equivalent versions. Here, since most of the problems we will be considering are cast as minimization problems, we prefer this more natural formulation, but retain the classical name. The minimum condition (2.9) states that in order to solve the minimization problem on the function space of controls, the control  $u_*$  needs to be chosen so that for some extremal lift, it minimizes the Hamiltonian  $H$  pointwise over the control set  $U$ , i.e., for every  $t \in [t_0, T]$ , the control  $u_*(t)$  is a minimizer of the function  $v \mapsto H(t, \lambda_0, \lambda(t), x_*(t), v)$  over the control set  $U$ . Note that it is not required just that the control satisfy the necessary conditions for minimality—and this is how a weak version of the maximum principle is formulated—but that the control  $u_*(t)$  be a true minimizer over the control set  $U$ . This condition typically is the starting point for any analysis of an optimal control problem. Formally, we first try to “solve” the minimization condition (2.9) for the control  $u$  as a function of the other variables,  $u = u(t, x_*, \lambda_0, \lambda)$ , and then substitute the “result” into the differential equations for dynamics and adjoint variable to get

$$\begin{aligned} \dot{x} &= f(t, x, u(t, x_*; \lambda_0, \lambda)), & x(t_0) &= x_0, \\ \dot{\lambda}(t) &= -\lambda_0 L_x(t, x_*(t), u(t, x_*; \lambda_0, \lambda)) - \lambda(t) f_x(t, x_*(t), u(t, x_*; \lambda_0, \lambda)). \end{aligned}$$

Since multiple solutions to the minimization problem can exist, this is not in general a unique specification of the control. Even if the minimization problem

does have a unique solution, this solution depends on the multiplier, i.e., lives in the cotangent bundle, and thus need not give rise to unique controlled trajectories.

- 4. Transversality conditions.** Equations (2.5) and (2.8) form a system in  $2n + 1$  variables (the state  $x$ , the multiplier  $\lambda$ , and the terminal time  $T$ ) with the initial condition  $x_0$  specified for the state at time  $t_0$ . Information about the remaining  $n + 1$  conditions is contained in the transversality conditions at the endpoint. The requirement that the terminal state lie on the manifold  $N$ ,  $(T, x(T)) \in N$ , imposes  $n + 1 - k$  conditions and thus leaves  $k$  degrees of freedom. The adjoint variable  $\lambda(T) \in (\mathbb{R}^n)^*$  at the terminal time  $T$  is determined on the  $k$ -dimensional tangent space to  $N$  at  $(T, x_*(T))$  by the relation

$$\lambda(T) = \lambda_0 \phi_x(T, x_*(T)) + v D_x \Psi(T, x_*(T))$$

and the multiplier  $v \in (\mathbb{R}^{n+1-k})^*$  in this equation accounts for  $n - (n + 1 - k) = k - 1$  degrees of freedom, with the last degree of freedom taken up by the equation

$$H(T, \lambda_0, \lambda(T), x_*(T), u_*(T)) + \lambda_0 \phi_t(T, x_*(T)) + v D_t \Psi(T, x_*(T)) = 0$$

which gives information about the terminal time  $T$ . Overall, there thus are  $2n + 1$  equations for the boundary values  $x(T)$ ,  $\lambda(T)$ , and  $T$ . Hence, at least in nondegenerate situations, the transversality conditions provide the required information about the missing boundary conditions for both the adjoint variable and the terminal time  $T$ .

The geometric statement that the vector  $(H + \lambda_0 \phi_t, -\lambda + \lambda_0 \phi_x)$  is orthogonal to the terminal constraint  $N$  at the endpoint of the controlled trajectory is valid for any embedded submanifold  $N$ . For since the condition is local, it is always possible to choose a collection of functions  $\psi_i$ ,  $i = 0, \dots, n - k$ , so that  $N = \{(t, x) : \Psi(t, x) = 0\}$  and the gradients of the functions  $\psi_i$  are linearly independent at  $(T, x_*(T))$ . The gradients  $\nabla \psi_i$  are all orthogonal to  $N$ , and since they are linearly independent, they span the space normal to  $N$ . Thus any covector normal to  $N$  at  $(T, x_*(T))$  is a linear combination of these covectors. Since the gradients are the rows of the matrix  $D\Psi(T, x_*(T))$ , there exists a row vector  $v = (v_0, \dots, v_{n-k})$  such that

$$(H + \lambda_0 \phi_t, -\lambda + \lambda_0 \phi_x) = -v (D_t \Psi, D_x \Psi).$$

This is equivalent to the formulation given in the theorem.

**Summarizing,** in order to solve an optimal control problem, in principle, we need to find all solutions to a boundary value problem on state and costate, coupled by a minimization condition, and then compare the costs that the projections of these solutions onto the controlled trajectories give. Clearly, this is not an easy problem, and thus the rest of this chapter will be spent on illustrating how one may go about doing this for some classes of optimal control problems, namely (i) once more the simplest problem in the calculus of variations, but now in dimension  $n$ , (ii) the linear-quadratic regulator, but now deriving its solution using the maximum principle,

(iii) the time-optimal control problem for linear time-invariant systems, and (iv) time-optimal control for general single-input, nonlinear, control-affine systems in the plane.

### 2.3 The Simplest Problem in the Calculus of Variations in $\mathbb{R}^n$

We once more consider the simplest problem in the calculus of variations, but now in arbitrary dimension  $n$ . This is a special case of an optimal control problem, and we illustrate how far-reaching the conditions of the maximum principle are by briefly deriving the highdimensional versions of the necessary conditions for optimality developed in Chap. 1.

Let  $L : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $(t, x, u) \mapsto L(t, x, u)$ , be a continuous function that for fixed  $t \in [a, b]$ , is differentiable in  $(x, u)$  with the partial derivatives  $\frac{\partial L}{\partial x}(t, x, u)$  and  $\frac{\partial L}{\partial u}(t, x, u)$  continuous in all variables. Also, let  $A$  and  $B$  be two given points in  $\mathbb{R}^n$ . We then consider the following problem:

[CV] Find, among all continuously differentiable curves  $x : [a, b] \mapsto \mathbb{R}^n$  that satisfy the boundary conditions  $x(a) = A$  and  $x(b) = B$ , one that minimizes the functional

$$I(x) = \int_a^b L(t, x(t), \dot{x}(t)) dt.$$

Calculus of variations problems are optimal control problems with a *trivial dynamics*,  $\dot{x} = u$ , and *no restrictions on the control set*: the state space is given by  $M = \mathbb{R}^n$ , the control set  $U$  is all of  $\mathbb{R}^n$ , and within our framework, the class  $\mathcal{U}$  of admissible controls is given by all piecewise continuous functions; the terminal manifold  $N$  is zero-dimensional given by the point  $B$ . If  $x_* : [a, b] \mapsto \mathbb{R}^n$  is an optimal solution, then with  $u_*(t) = \dot{x}_*(t)$ , the conditions of the maximum principle state that there exist a constant  $\lambda_0 \geq 0$  and an adjoint variable  $\lambda : [a, b] \rightarrow (\mathbb{R}^n)^*$  satisfying

$$\dot{\lambda}(t) = -\lambda_0 \frac{\partial L}{\partial x}(t, x_*(t), u_*(t))$$

such that  $(\lambda_0, \lambda(t)) \neq 0$  for all  $t \in [a, b]$  and

$$\lambda_0 L(t, x_*(t), u_*(t)) + \lambda(t) u_*(t) = \min_{v \in \mathbb{R}^n} [\lambda_0 L(t, x_*(t), v) + \lambda(t) v] = \text{const.} \quad (2.14)$$

Since the interval  $[a, b]$  and the endpoint are fixed, no transversality conditions apply: the vector  $v$  can be any vector in  $(\mathbb{R}^{n+1})^*$  leaving the terminal values of  $\lambda$  and  $H(b, \lambda_0, \lambda(b), x_*(b), u_*(b))$  free. But *extremals for the simplest problem in the calculus of variations are always normal*: If  $\lambda_0 = 0$ , then the minimum condition (2.14) implies that  $u_*(t)$  minimizes the linear function  $v \mapsto \lambda(t)v$  over  $\mathbb{R}^n$ . But such a minimum exists only if  $\lambda(t) = 0$ , and this then contradicts the nontriviality of the multipliers. Thus  $\lambda_0$  cannot vanish, and without loss of generality we may normalize it as  $\lambda_0 = 1$ .



The first-order necessary conditions for minimizing the function

$$v \mapsto L(t, x_*(t), v) + \lambda(t)v$$

over  $\mathbb{R}^n$  then imply that

$$\frac{\partial L}{\partial u}(t, x_*(t), u_*(t)) + \lambda(t) = 0.$$

Combining this relation with the adjoint equation, while identifying  $\dot{x}_*$  with the control  $u$ , gives the standard form of the *Euler–Lagrange equation*, now valid for the coordinates of the respective gradients of the Lagrangian

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}}(t, x_*(t), \dot{x}_*(t)) \right) = \frac{\partial L}{\partial x}(t, x_*(t), \dot{x}_*(t)).$$

The actual minimum condition (2.14) of the maximum principle is the *Weierstrass condition* of the calculus of variations: recall that the Weierstrass excess function  $E$  was defined as

$$E(t, x, y, u) = L(t, x, u) - L(t, x, y) - \frac{\partial L}{\partial \dot{x}}(t, x, y)(u - y);$$

thus condition (2.14) states that

$$E(t, x_*(t), \dot{x}_*(t), u) \geq 0 \quad \text{for all } u \in \mathbb{R}^n.$$

As shown in Sect. 1.6, this is a necessary condition for a *strong* local minimum of a very different character from that of the Euler–Lagrange equation. Recall that the piecewise continuous variations used in its proof allowed the derivatives to diverge, and thus this no longer is a necessary condition for a weak minimum. We shall see in Sect. 4.2 that Weierstrass’s variations pointed the path to the variations used in the proof of the maximum principle.

If the Lagrangian  $L$  is twice continuously differentiable, additional regularity statements about extremals easily follow from the maximum principle. For example, the second-order necessary condition for the function  $v \mapsto L(t, x_*(t), v) + \lambda(t)v$  to have a minimum over  $\mathbb{R}^n$  at  $\dot{x}_*(t)$  implies that the Hessian matrix

$$\frac{\partial^2 L}{\partial \dot{x}^2}(t, x_*(t), \dot{x}_*(t))$$

is positive semidefinite for  $t \in [a, b]$ . This is the multi-dimensional version of the *Legendre condition*. The *strengthened Legendre condition* holds over the interval  $[a, b]$  if this matrix is positive definite for  $t \in [a, b]$ . In this case, as in the scalar case, the *Hilbert differentiability theorem* is valid, and the extremal  $x_*$  is twice continuously differentiable. The argument is the same as in the scalar case: for some

constant  $c$  the extremal  $x_*$  is a solution to the Euler–Lagrange equation in integrated form,

$$\frac{\partial L}{\partial \dot{x}}(t, x_*(t), \dot{x}_*(t)) - \int_a^t \frac{\partial L}{\partial x}(t, x_*(s), \dot{x}_*(s)) ds - c = 0,$$

and defining a function  $F(t, w)$  as

$$F(t, w) = \frac{\partial L}{\partial \dot{x}}(t, x_*(t), w) - \int_a^t \frac{\partial L}{\partial x}(t, x_*(s), \dot{x}_*(s)) ds - c,$$

the equation  $F(t, w) = 0$  has the solution  $w(t) = \dot{x}_*(t)$ . By the implicit function theorem, this solution is continuously differentiable if the partial derivative

$$\frac{\partial F}{\partial w}(t, \dot{x}_*(t)) = \frac{\partial^2 L}{\partial \dot{x}^2}(t, x_*(t), \dot{x}_*(t))$$

is nonsingular. Hence  $x_*$  is twice continuously differentiable at all points where the strengthened Legendre condition holds.

The connections between optimal control and problems in the calculus of variations can be carried further including generalizations of the Jacobi condition and field theory. These aspects will be developed in Chap. 5.

## 2.4 The Linear-Quadratic Regulator Revisited

We briefly return to the linear-quadratic regulator and give a derivation of the optimal feedback control law from the conditions of the maximum principle. This argument is instructive and will be expanded further in Sect. 5.3 in connection with conjugate points for the optimal control problem. Also, in low dimensions, explicit solutions of the Riccati equation for the feedback gain  $S$  can be computed using these constructions, and we illustrate this with two scalar examples. As in the calculus of variations, there are no restrictions on the control set, i.e.,  $U = \mathbb{R}^m$ , but now a dynamics (albeit a simple linear one) is involved. In fact, since the Hamiltonian  $H$  is strictly convex in the control  $u$ , for this case, variational arguments as they were developed in Chap. 1 would still be sufficient to characterize the minimum.

### 2.4.1 A Derivation of the Optimal Control from the Maximum Principle

Recall that the linear quadratic regulator [LQ] is the problem of minimizing a quadratic objective of the form

$$J(u) = \frac{1}{2} \int_0^T [x^T(t)Q(t)x(t) + u^T(t)R(t)u(t)] dt + \frac{1}{2} x^T(T)S_T x(T)$$

over all (piecewise) continuous functions  $u : [0, T] \rightarrow \mathbb{R}^m$  defined over a fixed interval  $[0, T]$  subject to a linear dynamics

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0.$$

The entries of the matrices  $A(\cdot)$ ,  $B(\cdot)$ ,  $R(\cdot)$ , and  $Q(\cdot)$  are continuous functions on the interval  $[0, T]$ , and the matrices  $R(\cdot)$  and  $Q(\cdot)$  are symmetric;  $R(\cdot)$  is positive definite, and  $Q(\cdot)$  positive semidefinite;  $S_T$  is a constant positive definite matrix.

As in the simplest problem of the calculus of variations, all *extremals are normal*: formally, since the problem is a minimization over a fixed interval  $[0, T]$  without terminal constraints, the submanifold  $N$  is described by a single function  $\Psi : [0, \infty) \times M \rightarrow \mathbb{R}^1$ ,  $(t, x) \mapsto \Psi(t, x) = t - T$ , defining the final time  $T$ , and the transversality condition (2.11) reduces to  $\lambda(T) = \lambda_0 x^T(T)S_T$ . Thus, the adjoint equation with terminal condition is given by

$$\dot{\lambda} = -\lambda_0 x^T Q(t) - \lambda A(t), \quad \lambda(T) = \lambda_0 x^T(T)S_T.$$

If  $\lambda_0 = 0$ , then  $\lambda$  is a solution to a homogeneous linear equation with 0 boundary conditions, hence identically zero. But this contradicts the nontriviality statement of the maximum principle. Thus, without loss of generality, we set  $\lambda_0 = 1$ . The Hamiltonian function  $H$  then takes the form

$$H = \frac{1}{2} x^T Q(t)x + \frac{1}{2} u^T R(t)u + \lambda(A(t)x + B(t)u),$$

and since the matrix  $R$  is positive definite, is strictly convex with a unique minimum given by the stationary point of the gradient in  $u$ ,

$$\frac{\partial H}{\partial u} = u^T R(t) + \lambda B(t) = 0,$$

i.e.,

$$u = -R^{-1}(t)B^T(t)\lambda^T. \quad (2.15)$$

For the subsequent calculation it is more convenient to write the equations in terms of  $\lambda^T$ , and we therefore define  $\mu = \lambda^T$ . Substituting Eq. (2.15) into the system and adjoint equation gives the following classical linear two-point boundary value problem for  $x$  and  $\mu$ :

$$\begin{pmatrix} \dot{x} \\ \dot{\mu} \end{pmatrix} = \begin{pmatrix} A(t) & -B(t)R(t)^{-1}B(t)^T \\ -Q(t) & -A(t)^T \end{pmatrix} \begin{pmatrix} x \\ \mu \end{pmatrix}, \quad \begin{pmatrix} x(0) \\ \mu(T) \end{pmatrix} = \begin{pmatrix} x_0 \\ S_T x(T) \end{pmatrix}.$$

This is the  $n$ -dimensional analogue of the linear Hamiltonian system considered in Sect. 1.4. Its solution is easily obtained from the solution of the associated matrix differential equation

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} A(t) & -B(t)R(t)^{-1}B(t)^T \\ -Q(t) & -A(t)^T \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \begin{pmatrix} X(0) \\ Y(T) \end{pmatrix} = \begin{pmatrix} \text{Id} \\ S_T \end{pmatrix}.$$

The following classical result generalizes Proposition 1.4.1 to the multidimensional case and establishes the connections between solutions to Riccati equations and quotients of solutions to linear differential equations in general. In the engineering literature, e.g., [64], this technique and its generalizations are known as the *sweep method*.

**Proposition 2.4.1.** *Suppose  $A(\cdot)$ ,  $B(\cdot)$ ,  $M(\cdot)$ , and  $N(\cdot)$  are continuous  $n \times n$  matrices defined on  $[0, T]$  and let  $(X, Y)^T$  be the solution to the initial value problem*

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} A & -M \\ -N & -B \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \begin{pmatrix} X(0) \\ Y(0) \end{pmatrix} = \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}. \quad (2.16)$$

*Suppose  $X_0$  is nonsingular. Then the solution  $X(t)$  is nonsingular on the full interval  $[0, T]$  if and only if the solution  $S$  to the Riccati equation*

$$\dot{S} + SA(t) + B(t)S - SM(t)S + N(t) \equiv 0, \quad S(0) = Y_0X_0^{-1}, \quad (2.17)$$

*exists on the full interval  $[0, T]$ , and in this case we have that*

$$Y(t) = S(t)X(t). \quad (2.18)$$

*The solution  $S$  to the Riccati equation (2.17) has a finite escape time at  $t = \tau$  if and only if  $\tau$  is the first time when the matrix  $X(t)$  becomes singular.*

*Proof.*  $[ \implies ]$  Suppose  $X(t)$  is nonsingular for all  $t \in [0, T]$ . Then  $S(t) = Y(t)X(t)^{-1}$  is well-defined over  $[0, T]$ , and we need only verify that  $S$  satisfies the Riccati equation (2.17). This is shown with a direct calculation: omitting the variable  $t$ , we have that

$$\dot{S} = \frac{d}{dt}(YX^{-1}) = \dot{Y}X^{-1} + Y \frac{d}{dt}(X^{-1}).$$

Since  $X(t)X(t)^{-1} = \text{Id}$ , it follows that

$$0 = \frac{d}{dt}(XX^{-1}) = \dot{X}X^{-1} + X \frac{d}{dt}(X^{-1}),$$

or

$$\frac{d}{dt}(X^{-1}) = -X^{-1}\dot{X}X^{-1},$$

and thus

$$\dot{S} = \dot{Y}X^{-1} - YX^{-1}\dot{X}X^{-1}.$$

Substituting the differential equations for  $\dot{X}$  and  $\dot{Y}$  gives

$$\dot{S} = (-NX - BY)X^{-1} - S(AX - MY)X^{-1} = -SA - BS + SMS - N.$$

[ $\Leftarrow$ ] Conversely, suppose a solution  $S$  to the Riccati equation exists on all of  $[0, T]$ . The linear equation

$$\dot{U} = (A(t) - M(t)S(t))U, \quad U(t_0) = X_0,$$

has a solution  $U = U(t)$  defined over the full interval  $[0, T]$ . Setting  $V(t) = S(t)U(t)$ , we have  $V(0) = S(0)X_0 = Y_0$  and

$$\dot{V} = \dot{S}U + S\dot{U} = (-SA - BS + SMS - N)U + S(A - MS)U = -NU - BV. \quad (2.19)$$

Thus the pair  $(U, V)^T$  is a solution to the initial value problem (2.16). But so is  $(X, Y)^T$ , and by the uniqueness of solutions we have  $(X, Y) = (U, V)$ , i.e.,  $Y(t) = S(t)X(t)$ .

Suppose that there exists a time  $\tau$  for which  $X(\tau)$  is singular. Pick  $x_0 \neq 0$  such that  $X(\tau)x_0 = 0$  and let  $x(t) = X(t)x_0$  and  $y(t) = Y(t)x_0$ . Then  $x(\tau) = 0$  and  $y(\tau) = Y(\tau)x_0 = S(\tau)x(\tau) = 0$ , and thus since  $(x, y)^T$  satisfies a homogeneous linear differential equation, both  $x$  and  $y$  vanish identically. But  $x(0) = X(0)x_0 \neq 0$ , since  $X(0)$  is nonsingular. Contradiction. Thus  $X(t)$  is nonsingular over all of  $[0, T]$ .  $\square$

For the linear-quadratic problem we already have seen in Theorem 2.1.1 that the associated Riccati equation has a solution over the full interval  $[0, T]$ , and thus we have  $\mu(t) = S(t)x(t)$ , or in the original notation,  $\lambda^T(t) = S(t)x(t)$ . Hence, as we already know, the optimal control is given by

$$u(t) = -R(t)^{-1}B(t)^T S(t)x(t).$$

This argument, however, is based only on necessary conditions and thus by itself does not prove the optimality of this control law. But of course, we already know that the control is optimal from Sect. 2.1.

### 2.4.2 Two Scalar Examples

We illustrate the solution procedure with two scalar examples in which the Riccati equation can be solved in analytic form.

*Example 2.4.1.* Let  $x$  and  $u$  be scalar and consider the problem to minimize the objective

$$J(u) = \frac{1}{2} \int_0^T (qx^2 + u^2) dt$$

subject to the dynamics

$$\dot{x} = ax + bu, \quad x(0) = x_0, \quad 0 \leq t \leq T.$$

For example, this is a simple model of regulating the pH value of some chemical component [139]. The variable  $x$  denotes the deviation of the pH value from a preset nominal value and the pH value is regulated through a controlling agent with the rate of change in pH proportional to a weighted sum of its current value and the strength of the controlling ingredient  $u$ , also measured by its deviation from the nominal pH value;  $a$  and  $b$  are known positive constants and  $x_0$  is the known initial value.

This formulation fits the model exactly, and thus the optimal control is given in feedback form as

$$u_*(t) = -bS(t)x_*(t), \quad 0 \leq t \leq T,$$

where  $S(t)$  is the solution to the Riccati equation (2.3),

$$\dot{S} + 2aS - b^2S^2 + q = 0, \quad S(T) = 0.$$

A scalar Riccati equation can always be reduced to a second-order homogeneous linear differential equation by making the substitution

$$\frac{\dot{\phi}}{\phi} = -b^2S,$$

which gives

$$-\frac{1}{b^2} \left( \frac{\ddot{\phi}\phi - (\dot{\phi})^2}{\phi^2} \right) = \dot{S} = \frac{2a}{b^2} \left( \frac{\dot{\phi}}{\phi} \right) + \frac{1}{b^2} \left( \frac{\dot{\phi}}{\phi} \right)^2 - q.$$

Equivalently,

$$\frac{1}{b^2} \left( \frac{\ddot{\phi}}{\phi} \right) = -\frac{2a}{b^2} \left( \frac{\dot{\phi}}{\phi} \right) + q,$$

and thus we obtain the following second-order homogeneous equation with constant coefficients:

$$\ddot{\phi} + 2a\dot{\phi} - qb^2\phi = 0.$$

From the terminal condition on  $S$  we get  $\dot{\phi}(T) = 0$ , and since we are just looking for a nontrivial solution, we may take  $\phi(T) = 1$ . Setting  $\kappa = \sqrt{a^2 + qb^2}$ , the explicit solution is given as

$$\phi(t) = e^{-a(t-T)} \left[ \cosh(\kappa(t-T)) + \frac{a}{\kappa} \sinh(\kappa(t-T)) \right], \quad 0 \leq t \leq T,$$

and thus

$$S(t) = -\frac{1}{b^2} \left( \frac{\dot{\phi}(t)}{\phi(t)} \right) = \frac{1}{b^2} \left( a - \kappa \frac{\kappa \sinh(\kappa(t-T)) + a \cosh(\kappa(t-T))}{\kappa \cosh(\kappa(t-T)) + a \sinh(\kappa(t-T))} \right),$$

with the optimal time-varying feedback gain given by  $-bS(t)$ .

*Example 2.4.2 (Inventory control).* [139] In most regulator problems, the variables are normalized as deviations from predetermined set points. In this example, a simple inventory control problem, the desired values are left as predetermined time-varying quantities, and we illustrate the changes that arise in the argument for such a model that involves a modified form of the Lagrangian in the objective. The reasoning given here easily extends to the general case (for example, see [64, 144]).

Consider a company that produces some good and has desired levels for the production and inventory over a planning horizon  $[0, T]$  represented by  $u_d(t)$  and  $x_d(t)$ , respectively. If the demand at time  $t$  is denoted by  $d(t)$ , then the rate of change of the inventory level  $x(t)$  is given by

$$\dot{x}(t) = u(t) - d(t), \quad x(0) = x_0.$$

If the firm's objective is to maintain the inventory and production levels, then it is reasonable to minimize a functional of the form

$$J(u) = \frac{1}{2} \int_0^T q [x(t) - x_d(t)]^2 + r [u(t) - u_d(t)]^2 dt,$$

where  $r$  and  $q$  are positive weights selected by the company. In this problem, we have the restrictions  $x(t) \geq 0$  and  $u(t) \geq 0$  that do not fit into the linear-quadratic regulator model, but making the natural assumption that  $x_d$  and  $u_d$  are positive continuous functions, for sufficiently high weights  $r$  and  $q$  we can assume that these conditions will not be violated. In other words, we solve the problem ignoring these constraints, but then need to verify that the optimal solution does not violate them. The other, less significant change to the model formulation analyzed so far is that the objective, when multiplied out, contains linear terms in  $x$  and  $u$  as well. These are easily incorporated into the sweep method described above. (This topic will still be picked up in greater generality in Sect. 5.3.)

The above change in the problem formulation does not alter the fact that extremals are normal, and the Hamiltonian for the problem is

$$H(t, \lambda, x, u) = \frac{q}{2} (x - x_d(t))^2 + \frac{r}{2} (u - u_d(t))^2 + \lambda (u - d(t)).$$

Minimization of the Hamiltonian over  $u \in \mathbb{R}$  leads to  $\lambda = -r(u_* - u_d)$  and hence

$$u_*(t) = -\frac{\lambda_*(t)}{r} + u_d(t), \quad 0 \leq t \leq T.$$

Substituting this relation into the dynamics and combining with the adjoint equation gives the inhomogeneous linear system

$$\begin{aligned}\dot{x}_*(t) &= -\frac{\lambda_*(t)}{r} + u_d(t) - d(t), \\ \dot{\lambda}_*(t) &= -qx_*(t) + qx_d(t),\end{aligned}$$

with boundary conditions  $x_*(0) = x_0$  and  $\lambda_*(T) = 0$ . In this case, the solutions are related by

$$\lambda_*(t) = a(t) + b(t)x_*(t), \quad (2.20)$$

for some  $C^1$ -functions  $a$  and  $b$  that satisfy the terminal conditions  $a(T) = 0$  and  $b(T) = 0$ . Differentiating Eq. (2.20), we get that

$$\dot{\lambda}_* = \dot{a} + \dot{b}x_* + b\dot{x}_*,$$

which, upon substituting for  $\dot{\lambda}_*$  and  $\dot{x}_*$ , yields

$$\dot{a} + b(u_d(t) - d(t)) - qx_d(t) - \frac{ab}{r} + \left(\dot{b} - \frac{b^2}{r} + q\right)x_* = 0.$$

This equation will be satisfied if we choose  $a$  and  $b$  such that

$$\dot{a} + b(u_d(t) - d(t)) - qx_d(t) - \frac{ab}{r} = 0, \quad a(T) = 0, \quad (2.21)$$

$$\dot{b} - \frac{b^2}{r} + q = 0, \quad b(T) = 0. \quad (2.22)$$

Equation (2.22) is the Riccati equation for a related standard linear-quadratic optimal control problem [LQ] and has a solution over the full interval  $[0, T]$  because of the positivity of  $q$  and  $r$ . Equation (2.21) then is a time-varying linear ODE defined over the full interval and thus also has a solution over the interval  $[0, T]$ . As for Example 2.4.1, the solution  $b$  to the Riccati equation can be calculated explicitly by making a substitution of the form

$$\frac{\phi}{\dot{\phi}} = -\frac{1}{r}b,$$

yielding the second-order equation

$$\ddot{\phi} = \frac{q}{r}\phi.$$

Setting  $\kappa = \sqrt{\frac{q}{r}}$ , the solution for terminal conditions  $\phi(T) = 1$  and  $\dot{\phi}(T) = 0$  is given by  $\phi(t) = \cosh(\kappa(t - T))$  and thus



$$b(t) = -r\kappa \tanh(\kappa(t - T)) = r\kappa \tanh(\kappa(T - t)).$$

We still need to find  $a(t)$ . This solution depends on the demand  $d(t)$  and the specified production levels  $x_d(t)$ . If these are constants, say the firm controlling the inventory desires to have the production rate equal to the demand rate,  $u_d = d = \text{const}$ , while at the same time maintaining a constant level of inventory,  $x_d = \text{const}$ , then Eq. (2.21) becomes

$$\dot{a} - \frac{b(t)}{r}a - qx_d = 0, \quad a(T) = 0.$$

Solving this equation, it follows that

$$a(t) = \frac{qx_d}{\kappa} \tanh(\kappa(t - T)) = -\sqrt{qr}x_d \tanh(\kappa(T - t)).$$

Hence the optimal feedback control  $u_*(t, x)$  is given by

$$\begin{aligned} u_*(t, x) &= -\frac{\lambda_*(t)}{r} + d = -\frac{a(t) + b(t)x_*(t)}{r} + d \\ &= \kappa \tanh(\kappa(T - t))(x_d - x_*(t)) + d, \quad 0 \leq t \leq T. \end{aligned}$$

Thus the optimal control equals the constant demand rate  $d$  plus a time-varying inventory correction factor proportional to the deviation from the set point.

## 2.5 Time-Optimal Control for Linear Time-Invariant Systems

The two classes of problems considered so far, the simplest problem in the calculus of variations and the linear-quadratic regulator, were both problems without constraints on the control set and as such, are examples that still could be fully analyzed with techniques from the calculus of variations. We now consider examples from another class of classical problems for which this no longer is the case: time-optimal control to a point for a time-invariant linear system with bounded controls.

[LTOC] Given a time-invariant linear control system

$$\Sigma : \quad \dot{x} = Ax + Bu, \quad A \in \mathbb{R}^{n \times n}, \quad B \in \mathbb{R}^{n \times m},$$

find among all piecewise continuous controls  $u$  that take values in the hypercube

$$U = \left\{ u \in \mathbb{R}^m : \|u\|_\infty = \max_{i=1, \dots, m} |u_i| \leq 1 \right\},$$

one that steers a given (but arbitrary) initial point  $x_0 \in \mathbb{R}^n$  into the origin 0 in minimum time.

In the formulation of Sect. 2.2, we have  $M = \mathbb{R}^n$ ,  $L(t, x, u) \equiv 1$ ,  $f(t, x, u) = Ax + Bu$ ,  $\phi \equiv 0$ , and  $\Psi$  is given by  $\Psi : [0, \infty) \times M \rightarrow \mathbb{R}^n$ ,  $(t, x) \mapsto \Psi(t, x) = x$ , i.e.,  $N = \{x \in \mathbb{R}^n : x = 0\}$ . Since both initial and terminal points on the state are specified, in this case the transversality conditions give no information about the multiplier  $\lambda$ . Formally, we have  $\frac{\partial \Psi}{\partial x}(t, x) = \text{Id}$  and thus  $\lambda(T) = v$ , an arbitrary covector from  $(\mathbb{R}^n)^*$ . But the transversality condition on the final time  $T$  implies that  $H(T, \lambda_0, \lambda, x, u) = 0$ . Since

$$H = \lambda_0 + \lambda(Ax + Bu)$$

is time-invariant, it follows that along any extremal, we have

$$H(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) \equiv 0.$$

In particular,  $\lambda(t)$  can never vanish, since otherwise also  $\lambda_0 = 0$ . The adjoint equation is given by

$$\dot{\lambda} = -\lambda A,$$

and the minimum condition implies that for each  $i = 1, \dots, m$ , the  $i$ th component  $u_*^{(i)}(t)$  of an optimal control must satisfy

$$u_*^{(i)}(t) = \begin{cases} +1 & \text{if } \lambda(t)b_i < 0, \\ -1 & \text{if } \lambda(t)b_i > 0, \end{cases}$$

where  $b_i$  is the  $i$ th column of  $B$ . Summarizing, we thus have the following version of the Maximum Principle for the optimal control problem [LTOC]:

**Theorem 2.5.1 (Maximum principle for problem [LTOC]).** *Let  $(x_*, u_*)$  be a controlled trajectory defined over the interval  $[t_0, T]$  that minimizes the time of transfer from  $x_0 \in \mathbb{R}^n$  to the origin. Then there exists a nontrivial solution  $\lambda : [t_0, T] \rightarrow (\mathbb{R}^n)^*$  to the adjoint equation  $\dot{\lambda} = -\lambda A$  so that the control  $u_*$  satisfies*

$$u_*^{(i)}(t) = \begin{cases} +1 & \text{if } \lambda(t)b_i < 0, \\ -1 & \text{if } \lambda(t)b_i > 0, \end{cases} \quad (2.23)$$

and the Hamiltonian is identically zero on  $[t_0, T]$ ,  $H(t, \lambda_0, \lambda(t), x_*(t), u_*(t)) \equiv 0$ .

The necessary conditions of this theorem are also sufficient for optimality under some easily verifiable controllability assumption on the system  $\Sigma$ . For the moment, consider a system  $\Sigma$  of the form

$$\Sigma : \quad \dot{x} = Ax + Bu, \quad x(0) = p, \quad u \in U, \quad (2.24)$$

with a general control set  $U \subset \mathbb{R}^m$ . Since the system is time-invariant, without loss of generality we normalize the initial time to  $t_0 = 0$ , and the solution  $x(\cdot; p)$  to the initial value problem (2.24) is given by

$$x(t; p) = e^{At}p + \int_0^t e^{A(t-s)}Bu(s)ds. \quad (2.25)$$

**Definition 2.5.1 (Reachable and controllable sets).** The time- $t$ -reachable set from  $p$  is the set of all points  $q \in \mathbb{R}^n$  that can be reached from  $p$  by means of an admissible control defined on the interval  $[0, t]$ ,

$$\text{Reach}_{\Sigma, t}(p) = \left\{ q \in \mathbb{R}^n : \exists u \in \mathcal{U} \text{ such that } q = e^{At}p + \int_0^t e^{A(t-s)}Bu(s)ds \right\}.$$

The *reachable set* from  $p$  is the union of all time- $t$ -reachable sets for  $t > 0$ ,

$$\text{Reach}_{\Sigma}(p) = \bigcup_{t>0} \text{Reach}_{\Sigma, t}(p).$$

The time- $t$ -controllable set to  $q$  is the set of all points  $p \in \mathbb{R}^n$  that can be steered into  $q$  by means of an admissible control defined on the interval  $[0, t]$ ,

$$\text{Contr}_{\Sigma, t}(q) = \left\{ p \in \mathbb{R}^n : \exists u \in \mathcal{U} \text{ such that } q = e^{At}p + \int_0^t e^{A(t-s)}Bu(s)ds \right\}.$$

The *controllable set* to  $q$  is the union of all time- $t$ -controllable sets for  $t > 0$ ,

$$\text{Contr}_{\Sigma}(q) = \bigcup_{t>0} \text{Contr}_{\Sigma, t}(q).$$

Clearly, a point  $q$  is reachable from  $p$  in time  $t$  if and only if  $p$  is controllable to  $q$  in time  $t$ . Thus, generally, we restrict our attention to reachable sets. It is clear from Eq. (2.25) that

$$\text{Reach}_{\Sigma, t}(p) = e^{At}p + \text{Reach}_{\Sigma, t}(0)$$

and henceforth we consider only the case  $p = 0$ .

A special situation arises if there are no restrictions on the control set, i.e., if  $U = \mathbb{R}^m$ . That is, we are considering the problem of whether *in principle* it is possible to steer a point  $p$  into another point  $q$ . In this case, for every  $t > 0$  the reachable set  $\text{Reach}_{\Sigma, t}(0)$  is a linear subspace (and in fact, the same one regardless of the size of the interval), known as the *controllable subspace*  $\mathcal{C}(A, B)$ . This is the subspace spanned by the columns of the so-called Kalman matrix, i.e.,

$$\mathcal{C}(A, B) = \text{Im} (B, AB, A^2B, \dots, A^{n-1}B). \quad (2.26)$$

**Theorem 2.5.2.** *If  $U = \mathbb{R}^m$ , then for every  $t > 0$*

$$\text{Reach}_{\Sigma,t}(0) = \mathcal{C}(A, B) = \text{Contr}_{\Sigma,t}(0).$$

*Proof.* We fix  $t$  and first show that the reachable set  $\text{Reach}_{\Sigma,t}(0)$  is given by the image  $\text{Im } W(t)$  of the matrix

$$W(t) = \int_0^t e^{A(t-s)} B B^T e^{A^T(t-s)} ds.$$

Choosing continuous time-varying controls of the form

$$u(s) = B^T e^{A^T(t-s)} p,$$

it follows that

$$x(t) = \int_0^t e^{A(t-s)} B u(s) ds = W(t) p,$$

and thus  $\text{Im } W(t) \subset \text{Reach}_{\Sigma,t}(0)$ . But  $W(t)$  is a symmetric matrix, and hence the full space  $\mathbb{R}^n$  is the direct sum of the image and the kernel of  $W(t)$  [113],

$$\mathbb{R}^n = \text{Im } W(t) \oplus \ker W(t).$$

Furthermore, the kernel is the orthogonal complement of the image,  $\ker W(t) = \text{Im } W(t)^\perp$ , and it therefore suffices to show that

$$\ker W(t) \subset \text{Reach}_{\Sigma,t}(0)^\perp.$$

Given any point  $y \in \ker W(t)$ , we have that

$$0 = \langle y, W(t)y \rangle = \int_0^t y^T e^{A(t-s)} B B^T e^{A^T(t-s)} y ds = \int_0^t \left\| B^T e^{A^T(t-s)} y \right\|_2^2 ds,$$

and thus  $y^T e^{A(t-s)} B \equiv 0$  on the interval  $[0, t]$ . Since any point  $q$  in the reachable set  $\text{Reach}_{\Sigma,t}(0)$  is of the form

$$q = \int_0^t e^{A(t-s)} B u(s) ds$$

for some control  $u$ , we thus have that

$$\langle y, q \rangle = \int_0^t y^T e^{A(t-s)} B u(s) ds = 0$$

for all  $q \in \text{Reach}_{\Sigma,t}(0)$ . Hence  $y \in \text{Reach}_{\Sigma,t}(0)^\perp$  as claimed. Overall, it therefore follows that

$$\text{Reach}_{\Sigma,t}(0) = \text{Im } W(t) \quad \text{for all } t > 0.$$

It remains to compute this image, or equivalently, the kernel of  $W(t)$ . If  $y \in \ker W(t)$ , then as shown above,  $y^T e^{A(t-s)} B \equiv 0$  on the interval  $[0, t]$ . Since this function is real-analytic, this is equivalent to the fact that all derivatives vanish at  $t = 0$ , i.e.,

$$y^T A^k B = 0 \quad \text{for all } k \in \mathbb{N}.$$

By the Cayley–Hamilton theorem [113],  $A^n$  can be expressed as a linear combination of the powers  $A^i$  for  $i = 0, 1, \dots, n-1$ , and thus this is equivalent to

$$y^T (B, AB, A^2 B, \dots, A^{n-1} B) = 0.$$

Hence the columns of the Kalman matrix

$$K = (B, AB, A^2 B, \dots, A^{n-1} B)$$

span the orthogonal complement to  $\ker W(t)$ ; that is, they span  $\text{Im } W(t)$ . This proves the result.  $\square$

**Definition 2.5.2 (Completely controllable).** The linear system  $\Sigma$  is said to be completely controllable if  $\mathcal{C}(A, B) = \mathbb{R}^n$ .

Thus, if the system  $\Sigma$  is completely controllable, then in principle, it is possible to go from any point  $p_0 \in \mathbb{R}^n$  to any other point  $p_1 \in \mathbb{R}^n$  in arbitrarily short time  $T$ . (Simply take the control that steers the point 0 into the point  $p_1 - e^{AT} p_0$  in time  $T$ .) Obviously, the shorter the time-interval is, the larger the control values need to become, and if the controls are bounded, then complete controllability no longer ensures that such a transfer is possible. In fact, as we shall see in the examples in the next section, with a bound on the controls, it may no longer be possible to steer  $p_0$  into  $p_1$  at all. However, for the system  $\Sigma$  with control set given by the hypercube

$$U = \left\{ u \in \mathbb{R}^m : \|u\|_\infty = \max_{i=1, \dots, m} |u_i| \leq 1 \right\}$$

(more generally, for any control set  $U$  that has 0 as interior point), this notion of complete controllability makes the conditions of the maximum principle also sufficient for optimality.

**Theorem 2.5.3.** *Consider the time-optimal control problem to the origin for the time-invariant linear system*

$$\Sigma : \quad \dot{x} = Ax + Bu, \quad A \in \mathbb{R}^{n \times n}, \quad B \in \mathbb{R}^{n \times m},$$

with control set

$$U = \left\{ u \in \mathbb{R}^m : \|u\|_\infty = \max_{i=1, \dots, m} |u_i| \leq 1 \right\}.$$

If the system  $\Sigma$  is completely controllable, then a control  $u : [0, T] \rightarrow U$  is time-optimal if and only if there exists a nontrivial solution  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$  to the adjoint equation  $\dot{\lambda} = -\lambda A$  such that

$$\lambda(t)Bu(t) = \min_{v \in U} \lambda(t)Bv. \quad (2.27)$$

This theorem will be proven in Sect. 3.5. Thus, for the time-optimal control problem for linear time-invariant systems, the conditions of the maximum principle are both necessary and sufficient for optimality under an easily verifiable algebraic condition. Note that the minimum condition (2.27) gives no information about  $u_*^{(i)}(t)$  for times  $t$  when  $\lambda(t)b_i = 0$ . The function  $\Phi_i(t) = \lambda(t)b_i$  is called the *i*th switching function, and its properties determine the structure of optimal controls. For instance, if  $\Phi_i(t)$  has a simple zero at time  $\tau$ , then the control switches between  $+1$  and  $-1$  at  $\tau$ . Controls that oscillate only between the upper and lower values  $\pm 1$  are called *bang-bang controls*. For general nonlinear systems with locally bounded Lebesgue measurable functions as controls, the switching functions may have complicated zerosets (see Sect. 2.8). But for linear systems, as we shall show in Chap. 3, these phenomena play a minor role, and we therefore do not discuss these features here. Rather, we close this section with a useful criterion on the eigenvalues of the matrix  $A$  that ensures that optimal controls are bang-bang and gives a bound on the number of switching times.

**Proposition 2.5.1.** *If all eigenvalues of the matrix  $A$  are real, then optimal controls for the single-input linear control system*

$$\dot{x} = Ax + bu, \quad A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n, \quad |u| \leq 1,$$

*are bang-bang with at most  $n - 1$  switching times.*

*Proof.* It follows from the adjoint equation,  $\dot{\lambda} = -\lambda A$ , that the derivatives of the switching function  $\Phi(t) = \lambda(t)b$  are given by

$$\dot{\Phi}(t) = -\lambda(t)Ab, \quad \ddot{\Phi}(t) = \lambda(t)A^2b, \quad \dots, \quad \Phi^{(r)}(t) = (-1)^r A^r b, \quad \dots$$

By the Cayley–Hamilton theorem, the matrix  $-A$  is a root of its characteristic polynomial,  $\chi_{-A}(-A) = 0$ , say

$$\chi_{-A}(t) = \det(t \cdot \text{Id} + A) = t^n + a_{n-1}t^{n-1} + \dots + a_1t + a_0.$$

Thus  $(-A)^n$  can be written as a linear combination of lower powers of  $A$ ,

$$(-A)^n = -a_{n-1}(-A)^{n-1} - \dots - a_1(-A) - a_0 \text{Id}.$$

The switching function therefore satisfies the  $n$ th-order linear differential equation

$$\Phi^{(n)}(t) + a_{n-1}\Phi^{(n-1)}(t) + \dots + a_1\Phi(t) + a_0 \equiv 0$$

with constant coefficients, where the polynomial is the characteristic polynomial of the matrix  $-A$ . Since all eigenvalues of  $A$ , and thus also those of  $-A$ , are real, the general solution  $\Phi$  to this differential equation is of the form

$$\Phi(t) = \sum_{i=1}^k p_i(t) e^{-\alpha_i t}, \quad (2.28)$$

where  $\alpha_1, \dots, \alpha_k$  are the *distinct* eigenvalues of the matrix  $A$  and  $p_i$  are polynomials of degree at most  $d_i$ , where  $d_i$  is the algebraic multiplicity of the eigenvalue  $\alpha_i$ , that is, the multiplicity of  $\alpha_i$  as a zero of the characteristic polynomial of  $A$ . Expressions of this type are called exponential polynomials, and the result of the proposition follows from a general property of these functions. Define the degree  $\text{Deg } \Phi$  of an exponential polynomial of the form (2.28) as

$$\text{Deg } \Phi = \sum_{i=1}^k (1 + \deg p_i),$$

where  $\deg p_i$  denotes the usual degree of the polynomial  $p_i$ . Then the proposition follows from the following lemma:

**Lemma 2.5.1.** *A nontrivial exponential polynomial of the form*

$$\Phi(t) = \sum_{i=1}^k p_i(t) e^{-\alpha_i t}$$

*of degree  $\text{Deg } \Phi = r$  has at most  $r - 1$  zeros.*

*Proof.* The proof is by induction on the degree  $r$ . If  $\text{Deg } \Phi = 1$ , then  $\Phi$  is of the form  $\Phi(t) = ce^{-\alpha t}$  with  $c \neq 0$  and hence  $\Phi$  has no zeros. Thus, inductively, assume that the statement is correct for all exponential polynomials of degree at most  $r$  and assume that  $\Phi$  is of degree  $r + 1$ . Then

$$\Psi(t) = \Phi(t) e^{\alpha_1 t} = p_1(t) + \sum_{i=2}^k p_i(t) e^{(\alpha_1 - \alpha_i)t}$$

also is an exponential polynomial of degree  $r + 1$ , and we have

$$\dot{\Psi}(t) = \dot{p}_1(t) + \sum_{i=2}^k (\dot{p}_i(t) + (\alpha_i - \alpha_1) p_i(t)) e^{(\alpha_1 - \alpha_i)t}.$$

Since differentiation of the polynomial lowers the degree,  $\deg \dot{p}_1 = \deg p_1 - 1$ , the derivative  $\dot{\Psi}$  is an exponential polynomial of strictly smaller degree,  $\text{Deg } \dot{\Psi} \leq r$ . Hence, by the inductive assumption,  $\dot{\Psi}$  has at most  $r - 1$  zeros. By the mean value theorem,  $\Psi$  therefore has at most  $r$  zeros. Hence so does  $\Phi$ .  $\square$

## 2.6 Time-Optimal Control for Planar Linear Time-Invariant Systems: Examples

We give several examples that illustrate how the conditions of the maximum principle can be used to construct optimal solutions for linear time-optimal control problems. The examples are two-dimensional, but the procedures are generally applicable. We start with the classical model of time-optimal control to the origin for the double integrator.

### 2.6.1 The Double Integrator

The double integrator is a mathematical model of an object moving along a horizontal line without friction, and the goal is to bring it to rest at the origin in minimum time. Here  $x(t)$  denotes the position of the object at time  $t$ ,  $\dot{x}(t)$  its velocity, and  $u(t)$  the external force applied to the object. Mathematically, we take as variable  $x = (x_1, x_2)^T = (x, \dot{x})^T$ , and the dynamics can be written in the form

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u.$$

The Hamiltonian  $H$  is given by

$$H = \lambda_0 + \lambda \left[ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \right] = \lambda_0 + \lambda_1 x_2 + \lambda_2 u,$$

and thus the minimum condition implies that

$$u(t) = \begin{cases} +1 & \text{if } \lambda_2(t) < 0, \\ -1 & \text{if } \lambda_2(t) > 0. \end{cases}$$

Obviously, the matrix  $A$  has the double eigenvalue 0, and thus by Proposition 2.5.1, optimal controls are bang-bang with at most one switching. Naturally, for this simple model this also is easily seen directly: The adjoint equation is given by

$$\dot{\lambda} = -\lambda \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

or

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = -\lambda_1,$$

and thus

$$\ddot{\lambda}_2 \equiv \frac{d}{dt}(-\lambda_1) = 0.$$



Hence any solution of the adjoint equation is an affine function  $\lambda_2(t) = \alpha t + \beta$  and has at most one zero. Therefore *optimal controls are bang-bang with at most one switching*.

Once this structure is known, it is straightforward to synthesize all possible extremals. We simply need to analyze the phase portraits of the two systems corresponding to the constant controls  $u \equiv +1$  and  $u \equiv -1$  and then consider all possible combinations that steer the system into the origin and have no more than one switching. Let  $X$  denote the vector field corresponding to control  $u \equiv -1$ , i.e.,  $\dot{x}_1 = x_2$  and  $\dot{x}_2 = -1$ . Forming  $\frac{dx_1}{dx_2} = -x_2$ , we see that the integral curves have the form  $x_1 = -\frac{1}{2}x_2^2 + a$  with  $a \in \mathbb{R}$  some constant. Analogously, if  $Y$  denotes the vector field corresponding to control  $u \equiv +1$ , then we have  $\dot{x}_1 = x_2$  and  $\dot{x}_2 = 1$ , and now the integral curves are given by  $x_1 = \frac{1}{2}x_2^2 + b$  with  $b \in \mathbb{R}$  another constant. Thus, all integral curves are parabolas opening left for  $u = -1$  and right for  $u = +1$ . Among all these curves, however, there are only two that steer the system into the origin directly, namely

$$\Gamma_+ : x_1 = \frac{1}{2}x_2^2 \quad \text{for } x_2 < 0$$

and

$$\Gamma_- : x_1 = -\frac{1}{2}x_2^2 \quad \text{for } x_2 > 0.$$

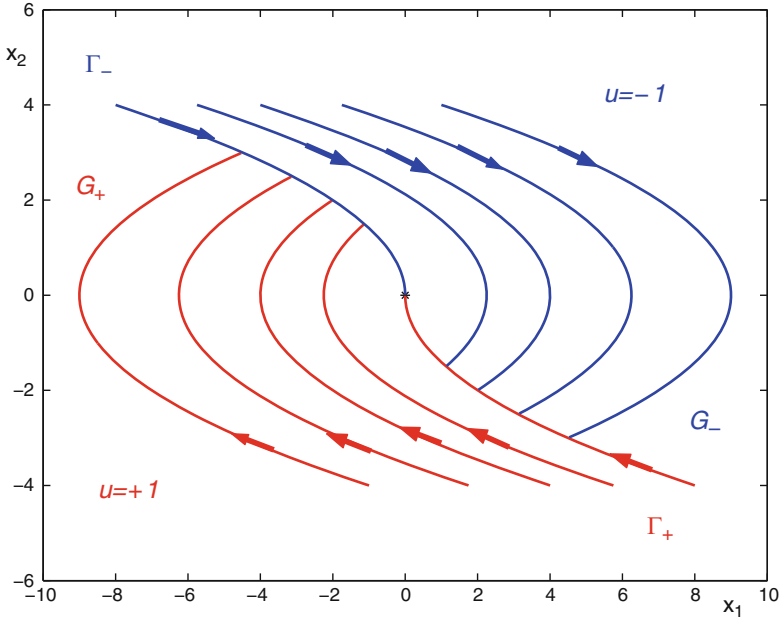
Only these two half-parabolas are integral curves that steer the system into the origin; the other two halves that were dropped steer the system away from the origin. Thus any optimal trajectory needs to arrive at the origin along either  $\Gamma_+$  or  $\Gamma_-$ . Bang-bang trajectories that have exactly one switching are now constructed by integrating the vector field  $X$  backward from any point in  $\Gamma_+$  and integrating  $Y$  backward from any point in  $\Gamma_-$ .

Denote the resulting family of extremal controlled trajectories by  $\mathcal{F}$ . It is clear that away from  $\Gamma_+$  and  $\Gamma_-$ , this family  $\mathcal{F}$  covers the entire state space injectively and for every initial condition  $(x_1^0, x_2^0) \neq (0, 0)$  there exists a unique extremal in  $\mathcal{F}$  that is bang-bang with at most one switching and steers the system into the origin *forward* in time. This family is shown in Fig. 2.1. In general, such a family of controlled trajectories is called an *extremal synthesis* (and this will be the main topic of Chap. 6). Note that for each trajectory, the control at  $(x_1, x_2)$  depends only on the actual point  $(x_1, x_2)$ , but not on the path along which this point was reached and thus we can describe the controls associated with this family as a discontinuous feedback control. If we define regions

$$G_+ = \left\{ (x_1, x_2) : x_1 < -\text{sgn}(x_2)\frac{1}{2}x_2^2 \right\}$$

and

$$G_- = \left\{ (x_1, x_2) : x_1 > -\text{sgn}(x_2)\frac{1}{2}x_2^2 \right\},$$



**Fig. 2.1** Synthesis of optimal controlled trajectories for the time-optimal control problem to the origin for the double integrator

then the corresponding controls are given by

$$u_*(x) = \begin{cases} +1 & \text{for } x \in \Gamma_+ \cup G_+, \\ -1 & \text{for } x \in \Gamma_- \cup G_-. \end{cases}$$

It follows from Theorem 2.5.3 that the controlled trajectories in this family are optimal. For this simple example, this can also easily be verified directly [41]: Let  $(\bar{x}_1, \bar{x}_2) \neq (0, 0)$  be an arbitrary initial condition and let  $T$  denote the time it takes for the system to reach the origin along the controlled trajectory in the family  $\mathcal{F}$ . Suppose there exists another control  $\bar{u}$  that steers  $(\bar{x}_1, \bar{x}_2)$  into the origin in time  $\bar{T} < T$ . Without loss of generality, consider the case that the control in the family  $\mathcal{F}$  is given by

$$u(t) = \begin{cases} -1 & \text{for } 0 < t \leq \alpha, \\ +1 & \text{for } \alpha < t \leq T. \end{cases}$$

Define functions

$$\Phi(t) = -x_1(t) + x_2(t)(t - \alpha)$$

and

$$\Psi(t) = -\bar{x}_1(t) + \bar{x}_2(t)(t - \alpha),$$

where  $(x_1(\cdot), x_2(\cdot))$  is the solution from the family  $\mathcal{F}$  and  $(\bar{x}_1(\cdot), \bar{x}_2(\cdot))$  is the solution corresponding to the control  $\bar{u}$ . Then we have that

$$\dot{\Phi}(t) = -\dot{x}_1(t) + \dot{x}_2(t)(t - \alpha) + x_2(t) = u(t)(t - \alpha) = |t - \alpha|$$

and

$$\dot{\Psi}(t) = -\dot{\bar{x}}_1(t) + \dot{\bar{x}}_2(t)(t - \alpha) + \bar{x}_2(t) = \bar{u}(t)(t - \alpha).$$

Since  $U = [-1, 1]$ , it follows that  $\dot{\Psi}(t) \leq |\dot{\Psi}(t)| \leq \dot{\Phi}(t)$  and thus

$$\int_0^{\bar{T}} \dot{\Phi}(t) dt \geq \int_0^{\bar{T}} \dot{\Psi}(t) dt.$$

Hence

$$\Phi(\bar{T}) - \Phi(0) \geq \Psi(\bar{T}) - \Psi(0).$$

But by construction,

$$\Phi(0) = -\bar{x}_1 - \alpha\bar{x}_2 = \Psi(0),$$

and since  $\Psi(\bar{T}) = 0$  (the system is at the origin at time  $\bar{T}$ ), we have  $\Phi(\bar{T}) \geq 0$ . But then

$$0 < \int_{\bar{T}}^T |t - \alpha| dt = \int_{\bar{T}}^T \dot{\Phi}(t) dt = \Phi(T) - \Phi(\bar{T}) = -\Phi(\bar{T}) \leq 0.$$

Contradiction. This proves that the family  $\mathcal{F}$  is an *optimal synthesis of controlled trajectories*.

The general question of optimality of an extremal synthesis will be considered in the context of sufficient conditions for optimality in Chap. 6. For the linear systems considered in this section, the optimality of all the syntheses constructed here follows from Theorem 2.5.3.

### 2.6.2 A Hyperbolic Saddle

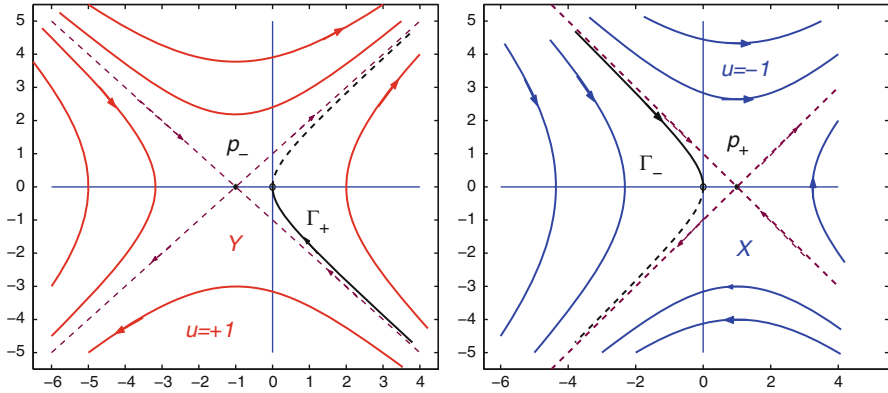
We now consider a system that has both a positive and negative eigenvalue:

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \quad |u| \leq 1.$$

Again, the system is completely controllable,

$$K = (b, Ab) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and the eigenvalues of  $A$  are  $\mu_1 = -1$  and  $\mu_2 = +1$ . Hence optimal controls are bang-bang with at most one switching, and an extremal synthesis is sufficient for optimality.



**Fig. 2.2** Phase portrait for  $u = +1$  (left) and for  $u = -1$  (right)

As for the double integrator, geometric properties of the phase portrait of the uncontrolled system determine the structure of the overall synthesis of optimal controlled trajectories. The origin is a hyperbolic saddle for the system  $\dot{z} = Az$ , and the stable and unstable subspaces at the equilibria are spanned by the eigenvectors  $v_1$  and  $v_2$  of the eigenvalues  $\mu_1$  and  $\mu_2$ , respectively,

$$v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

That is, if  $p$  is a multiple of  $v_1$ , then the solution  $z(t)$  to the initial value problem  $\dot{z} = Az$ ,  $z(0) = p$ , is given by  $z(t) = e^{At}p = e^{-t}p$  and thus satisfies  $\lim_{t \rightarrow \infty} z(t) = 0$ , while for multiples of  $v_2$  the solution is given by  $z(t) = e^{At}p = e^t p$  and thus satisfies  $\lim_{t \rightarrow -\infty} z(t) = 0$ . The phase portraits for the controlled vector fields with  $u = +1$  or  $u = -1$  are simply shifted versions of the phase portrait of the homogeneous system  $\dot{z} = Az$  along the  $x_1$ -axis and are shown in Fig. 2.2.

If, as above, we denote by  $X$  the vector field corresponding to the control  $u \equiv -1$  and by  $Y$  the vector field corresponding to the control  $u \equiv +1$ , i.e.,

$$X(x) = Ax - b = \begin{pmatrix} x_2 \\ x_1 - 1 \end{pmatrix}, \quad Y(x) = Ax + b = \begin{pmatrix} x_2 \\ x_1 + 1 \end{pmatrix},$$

then these vector fields now have a hyperbolic saddle at the points  $p_+ = (+1, 0)$  and  $p_- = (-1, 0)$  and the stable and unstable subspaces of the matrix  $A$  are translated to become lines through  $p_+$  and  $p_-$ . Note that there again exist unique trajectories  $\Gamma_-$  of  $X$  and  $\Gamma_+$  of  $Y$  that steer the system into the origin forward in time, shown as solid black curves in Fig. 2.2. Their continuations, which will not be part of the synthesis, are shown dashed. As with the double integrator, an extremal synthesis is then constructed by integrating  $X$  backward from points in  $\Gamma_+$  and  $Y$  backward

from points in  $\Gamma_-$ . However, it is now no longer possible to steer every point into the origin as it was the case with the double integrator, and the controllable set is bounded by the stable manifolds of the equilibria  $p_+$  and  $p_-$ , that is, by the lines

$$E_+ = p_+ + \text{linspan}\{v_1\} = \{x \in \mathbb{R}^2 : x_1 + x_2 = +1\}$$

and

$$E_- = p_- + \text{linspan}\{v_1\} = \{x \in \mathbb{R}^2 : x_1 + x_2 = -1\}.$$

Clearly, for any admissible control  $u$ , we have that

$$\frac{d}{dt}(x_1 + x_2) = (x_1 + x_2) + u$$

and since  $|u| \leq 1$ , we always have  $\frac{d}{dt}(x_1 + x_2) \leq 0$  at points  $(x_1, x_2)$  satisfying  $x_1 + x_2 \leq -1$  and  $\frac{d}{dt}(x_1 + x_2) \geq 0$  at points satisfying  $x_1 + x_2 \geq 1$ . Thus no point outside of

$$\mathcal{C} = \{(x_1, x_2) : -1 < x_1 + x_2 < 1\}$$

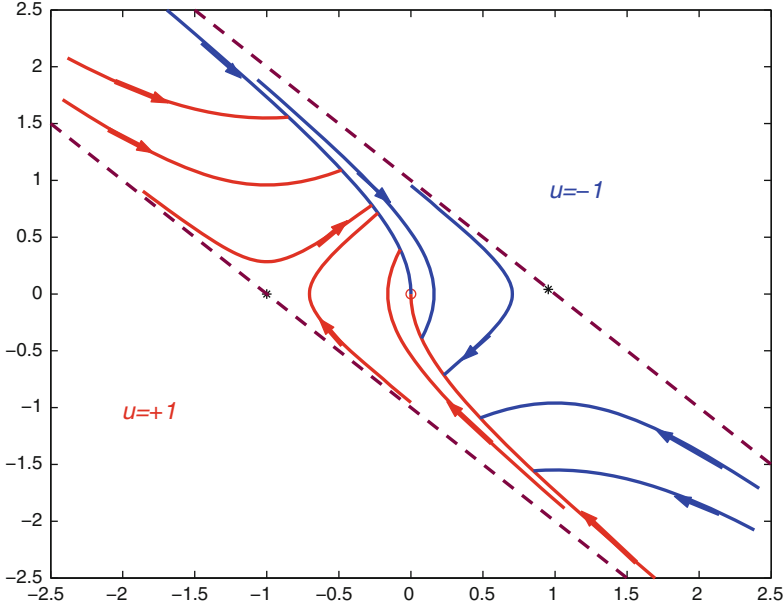
can be steered into the origin. On the other hand, if a point lies in  $\mathcal{C}$ , then it is clear from the phase portraits that there exists a unique bang-bang control with at most one switching that steers this initial condition into the origin. This family of controlled trajectories is illustrated in Fig. 2.3. By construction this family of controlled trajectories is an extremal synthesis, and hence it is optimal by Theorem 2.5.3.

This example illustrates the obvious fact that complete controllability does not allow one to freely steer the system into arbitrary points if constraints are imposed on the control. As seen in Example 2.4.1, if the eigenvalues are critical, i.e., lie on the imaginary axis, then the instability can be fully overcome by any kind of control action (of course, the control set needs to contain the origin in its interior). Generally, for unstable systems with eigenvalues with positive real parts a certain degree of instability can be overcome depending on the size of the control that is allowed. In Example 2.4.2, when there still existed a one-dimensional stable subspace for the system, it was this subspace (and the size on the control) that determined the controllable set. The next example shows what happens if the system is an unstable node without any stable trajectories at all. Even in this case, the control is still able to overcome some of the instabilities, and the controllable set still is open.

### 2.6.3 An Unstable Node

Consider the system

$$\dot{x} = \begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \quad u \in [-1, 1].$$



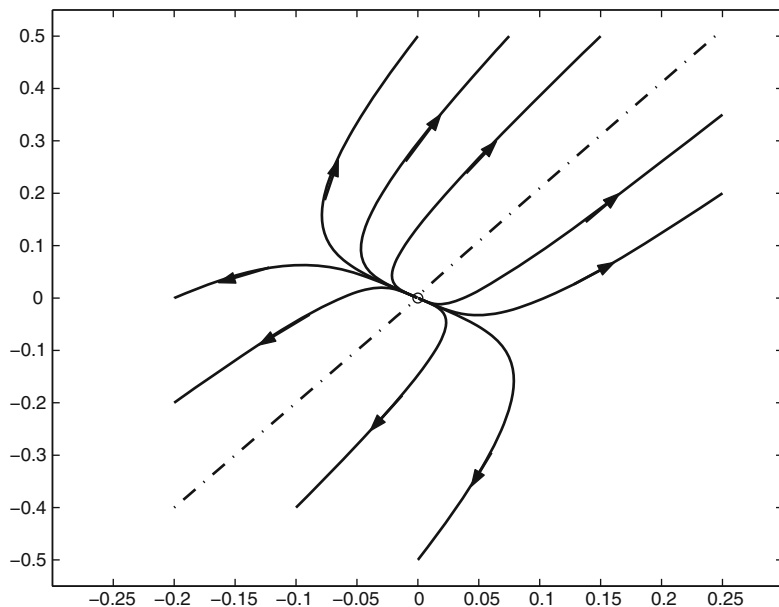
**Fig. 2.3** Synthesis of optimal controlled trajectories for the time-optimal control problem to the origin for a hyperbolic saddle

As above, the system is completely controllable,

$$K = (b, Ab) = \begin{pmatrix} 0 & 1 \\ 1 & 3 \end{pmatrix},$$

with two real eigenvalues,  $\mu_1 = 1$  and  $\mu_2 = 4$ , and as before, optimal controls are bang-bang with at most one switching and an extremal synthesis is optimal.

The uncontrolled system is an unstable node, and the phase portraits for the controlled vector fields  $X$  and  $Y$  corresponding to the constant controls  $u = -1$  and  $u = +1$ , respectively, again are simply shifted versions along the  $x_1$ -axis of the phase portrait of  $\dot{z} = Az$  shown in Fig. 2.4. In this case, the solutions along the eigenvectors do not play an important role, but instead the boundary of the controllable set is given by two specific trajectories  $\Lambda_+$  and  $\Lambda_-$  of the vector fields  $X$  and  $Y$ :  $\Lambda_+$  is the backward orbit of the trajectory of the vector field  $Y$  that passes through the equilibrium point  $p_- = (-\frac{1}{4}, \frac{1}{2})$  of the vector field  $X$  at time 0 and converges to the equilibrium  $p_+ = (\frac{1}{4}, -\frac{1}{2})$  of the vector field  $Y$  as  $t \rightarrow -\infty$ , and symmetrically,  $\Lambda_-$  is the backward orbit of the trajectory of the vector field  $X$  that passes through the equilibrium point  $p_+$  of the vector field  $Y$  at time 0 and converges to the equilibrium  $p_-$  of the vector field  $X$  as  $t \rightarrow -\infty$ . The concatenation of these two curves with the equilibria  $p_+$  and  $p_-$  forms a simple closed curve, and the controllable set  $\mathcal{C}$  is the interior of this closed curve with  $\mathcal{C}$  as its boundary. The optimal synthesis is constructed analogously as for the double integrator and the hyperbolic saddle by



**Fig. 2.4** Phase portrait of  $\dot{z} = Az$  for an unstable node

integrating the vector fields  $X$  and  $Y$  backward from the unique trajectories  $\Gamma_-$  of  $X$  and  $\Gamma_+$  of  $Y$  that steer the system into the origin forward in time, and it is illustrated in Fig. 2.5.

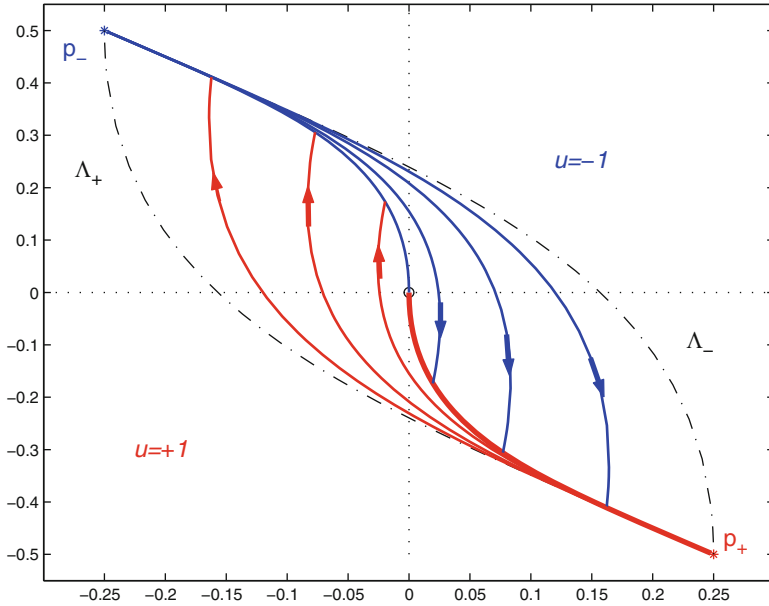
### 2.6.4 The Harmonic Oscillator

We close this section with an example of a matrix  $A$  that has complex eigenvalues. Because of the inherent oscillatory character of these systems, the number of switchings no longer can be bounded. We consider the harmonic oscillator. As before,  $x(t)$  denotes the position of the object at time  $t$ ,  $\dot{x}(t)$  its velocity, and  $u(t)$  the external force applied to the object, and we write the state as  $x = (x_1, x_2)^T = (x, \dot{x})^T$ . Now the dynamics takes the form

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u,$$

and the Hamiltonian  $H$  is given by

$$H = \lambda_0 + \lambda \left[ \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \right] = \lambda_0 + \lambda_1 x_2 + \lambda_2 (-x_1 + u).$$



**Fig. 2.5** Synthesis of optimal controlled trajectories for the time-optimal control problem to the origin for an unstable node

Thus again, the minimum condition implies that

$$u(t) = \begin{cases} +1 & \text{if } \lambda_2(t) < 0, \\ -1 & \text{if } \lambda_2(t) > 0. \end{cases}$$

As for all linear systems, the adjoint equation is given by the system itself run backward, but written as a row vector

$$\dot{\lambda} = -\lambda \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

or

$$\dot{\lambda}_1 = \lambda_2, \quad \dot{\lambda}_2 = -\lambda_1,$$

and thus

$$\ddot{\lambda}_2 \equiv \frac{d}{dt}(-\lambda_1) = -\lambda_2.$$

Hence, all solutions of the adjoint equation are integral curves of the harmonic oscillator. Thus again *optimal controls are bang-bang*, but now we cannot give an a priori bound on the number of switchings. In fact, depending on the initial condition, this number can be arbitrarily large. However, since switchings are the zeros of a solution to the harmonic oscillator, it follows that *all switchings  $\tau_n$  are*



spaced exactly  $\pi$  units apart, and the first switching  $\tau_1$  can take any value in the interval  $(0, \pi]$ . Analytically, any solution  $\lambda_2$  of the adjoint equation is of the form  $\lambda_2(t) = a \cos t + b \sin t$  for some constants  $a$  and  $b$  and therefore can be written in phase-angle form as

$$\lambda_2(t) = A \cos(t - \varphi)$$

with amplitude  $A = \sqrt{a^2 + b^2}$  and phase  $\varphi = \arctan\left(\frac{b}{a}\right)$ .

With this information, as with the examples above, it is again straightforward to synthesize all possible extremals by analyzing the phase portraits of the systems corresponding to the constant controls  $u \equiv +1$  and  $u \equiv -1$  and then consider all possible concatenations that switch exactly  $\pi$  units of time apart. As before, let  $X$  and  $Y$  denote the vector fields corresponding to the controls  $u \equiv -1$  and  $u \equiv +1$ , respectively. Integral curves of  $X$  are circles with center at the point  $p_- = (-1, 0)$ , and integral curves of  $Y$  are circles with center at  $p_+ = (1, 0)$ , both traversed clockwise. Exactly as in the case of the double integrator, among all these trajectories there are only two that steer the system into the origin directly, namely

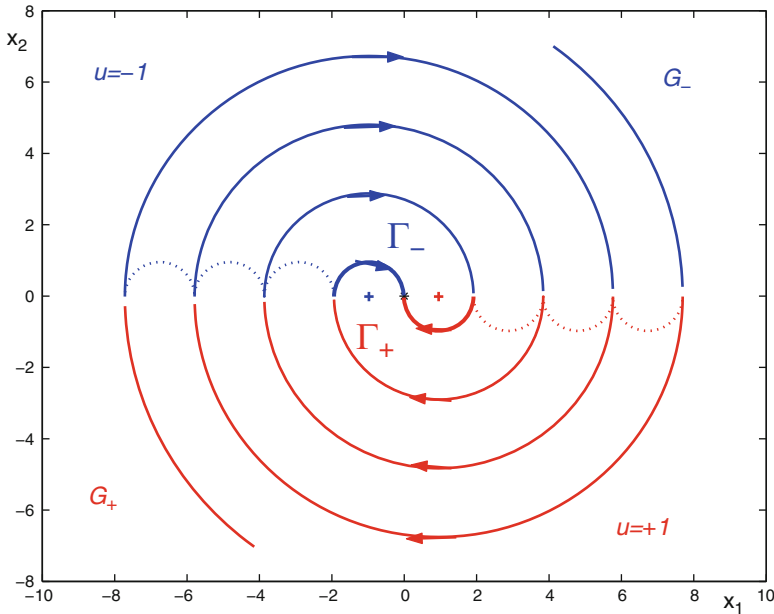
$$\Gamma_+ : [-\pi, 0] \rightarrow \mathbb{R}^2, \quad t \mapsto (x_1(t), x_2(t)) = (1 - \cos(t), \sin(t)),$$

and

$$\Gamma_- : [-\pi, 0] \rightarrow \mathbb{R}^2, \quad t \mapsto (x_1(t), x_2(t)) = (-1 + \cos(t), -\sin(t)).$$

Note that  $\Gamma_-$  is the curve obtained by reflecting  $\Gamma_+$  at the origin, and only these two semicircles are admissible extremal trajectories, since switchings must be spaced  $\pi$  units apart. Thus there cannot be any segment of an optimal  $X$  or  $Y$  trajectory longer than  $\pi$ . Any extremal control that steers the system into the origin needs to do so along either  $\Gamma_+$  or  $\Gamma_-$  as final segment. The full family  $\mathcal{F}$  is now constructed by picking a point  $q_+(t) \in \Gamma_+$  (respectively  $q_-(t) \in \Gamma_-$ ) for a time  $t \in [-\pi, 0]$  and integrating the vector fields  $X$  and  $Y$  (respectively,  $Y$  and  $X$ ) backward from  $q_+(t)$  (respectively  $q_-(t)$ ), switching between these vector fields at all times precisely  $\pi$  units apart. Thus, with the final time normalized to 0, the switchings occur at times  $t, t - \pi, t - 2\pi, \dots$  and the curves where the switchings occur, the so-called *switching curves*, are obtained inductively by following the flow of  $X$ , respectively  $Y$ , for exactly  $\pi$  units of time starting with  $\Gamma_+$  and  $\Gamma_-$ . Since integral curves are concentric circles centered at  $p_{\pm}$ , this generates a family of shifted semicircles of type  $\Gamma_+$  below the positive  $x_1$ -axis and of type  $\Gamma_-$  above the negative  $x_1$ -axis as depicted in Fig. 2.6. In this figure, the curves  $\Gamma_+$  and  $\Gamma_-$  are shown as solid curves since these are actually integral curves of  $X$  and  $Y$ , while all their translates are strictly switching curves that do not correspond to integral curves and are shown dashed. On all points on these translates, the controls switch between  $+1$  and  $-1$  and the corresponding trajectories cross the switching curves.

As with the double integrator, the family  $\mathcal{F}$  covers the entire state space injectively, and for every initial condition  $(x_1^0, x_2^0)$  there exists a unique bang-bang extremal that steers the system into the origin. So again we have an extremal synthesis, and the control can be given as a feedback control. If we denote the switching locus by  $\mathcal{Y}$  and let  $G_+$  denote the region below  $\mathcal{Y}$  in the  $(x_1, x_2)$ -plane



**Fig. 2.6** Synthesis of optimal controlled trajectories for time-optimal control to the origin for the harmonic oscillator

and  $G_-$  the region above  $\mathcal{Y}$ , then the control is again a discontinuous feedback of the form

$$u_*(x) = \begin{cases} +1 & \text{for } x \in \Gamma_+ \cup G_+, \\ -1 & \text{for } x \in \Gamma_- \cup G_-. \end{cases}$$

As with the other examples considered in this section, by Theorem 2.5.3 the family  $\mathcal{F}$  of controlled trajectories is an *optimal synthesis*.

We close this section with pointing out the special nature of the trajectories that end with the full semicircles  $\Gamma_+$  and  $\Gamma_-$ . Let  $\gamma_+$  and  $\gamma_-$  be the controlled trajectories in the field  $\mathcal{F}$  that end at the origin at time 0 by following the full arcs  $\Gamma_+$  and  $\Gamma_-$ , respectively, and have switchings at all negative integer multiples of  $\pi$ . These two extremal trajectories are *strictly abnormal*, i.e., the only way to satisfy the conditions of the maximum principle is with  $\lambda_0 = 0$ . Thus, the trajectories  $\gamma_+$  and  $\gamma_-$  are examples of optimal trajectories whose extremals are abnormal.

**Proposition 2.6.1.** *The extremals corresponding to  $\gamma_+$  and  $\gamma_-$  are unique (up to a positive multiple) and are strictly abnormal.*

*Proof.* Without loss of generality, we consider  $\gamma_-$ . Since the system is time-invariant, we can normalize the final time to be  $T = 0$  and integrate backward. Then, as used already above, the parametrization of  $\gamma_-$  (respectively  $\Gamma_-$ ) over the interval  $[-\pi, 0]$  is given by

$$x_1(t) = -1 + \cos(t) \quad \text{and} \quad x_2(t) = -\sin(t).$$

The times  $t = -\pi$  and  $t = 0$  are switching times. Since  $\lambda$  itself is a solution of the harmonic oscillator, only multiples of  $\sin(t)$  will satisfy this condition, and since the control is  $u = -1$  on  $[-\pi, 0]$ , we must have  $\lambda_2(t) = \alpha \sin(t)$  for some  $\alpha < 0$ . Hence  $\lambda_1(t) = -\dot{\lambda}_2(t) = -\alpha \cos(t)$ , and on  $[-\pi, 0]$  the Hamiltonian  $H$  takes the form

$$\begin{aligned} H &= \lambda_0 + \lambda_1(t)x_2(t) + \lambda_2(t)(-x_1(t) + u(t)) \\ &= \lambda_0 + \alpha \cos(t) \sin(t) + \alpha \sin(t)(1 - \cos(t) - 1) \\ &= \lambda_0. \end{aligned}$$

But it follows from Theorem 2.5.1 that  $H \equiv 0$ , and thus we must have  $\lambda_0 = 0$ .  $\square$

## 2.7 Extensions of the Model: Two Examples

The examples considered so far fall into well-established classes, linear-quadratic optimal control and time-optimal control for time-invariant linear systems. But the techniques that were used apply more generally, and as further illustration of how to use the conditions of the maximum principle, we shall solve a basic trading problem in economics and a classical example of a nonlinear system, the so-called moon landing problem, that will lead us to a discussion of general nonlinear control-affine systems in the next section.

### 2.7.1 An Economic Trading Model

We consider a simple model of a firm that buys and sells a product and has cash and the quantity of this product as its two assets; denote the values of these assets at time  $t$  by  $x_1(t)$  and  $x_2(t)$ , respectively [139]. The initial values of the assets,  $x_1(0)$  and  $x_2(0)$ , are given. If the company's reservation utility for the price of the product at the end of some planning period  $[0, T]$  is denoted by  $\pi$ , then the firm's goal is to *maximize*

$$C(u) = x_1(T) + \pi x_2(T).$$

Ideally, the reservation utility would agree with the price at time  $T$ , but a priori this price is unknown. The control in the problem, represented by  $u(t)$ , is the rate of buying and selling the product at time  $t$  with  $u(t) > 0$  considered a purchase and  $u(t) < 0$  a sale. We assume that at any time, there are (self-imposed) limits on the amount of the product the company wants to buy or sell, say  $m \leq u(t) \leq M$  with  $m < 0$  and  $M > 0$  given constants. If  $p(t)$  denotes the price of the product at time  $t$ , then the effect of a trading operation on the company's assets is given by

$$\dot{x}_1(t) = -\alpha x_2(t) - p(t)u(t), \quad \dot{x}_2(t) = u(t),$$

where  $\alpha > 0$  is a constant associated with the cost of storing a unit of the product and the term  $p(t)u(t)$  gives the cost of purchase or the revenue from sales at time  $t$ .

The dynamics now has the form  $\dot{x} = Ax + B(t)u$ , where  $A$  is time-invariant, but  $B$  is time-varying,

$$A = \begin{pmatrix} 0 & -\alpha \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B(t) = \begin{pmatrix} -p(t) \\ 1 \end{pmatrix}.$$

The Hamiltonian  $H$  for the problem is

$$H = H(t, \lambda, x, u) = \lambda_1(-\alpha x_2 - p(t)u) + \lambda_2 u = -\alpha \lambda_1 x_2 + (\lambda_2 - \lambda_1 p(t))u,$$

and the adjoint equations are given by

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = \alpha \lambda_1,$$

with transversality conditions

$$\lambda_1(T) = -\lambda_0, \quad \lambda_2(T) = -\lambda_0 \pi.$$

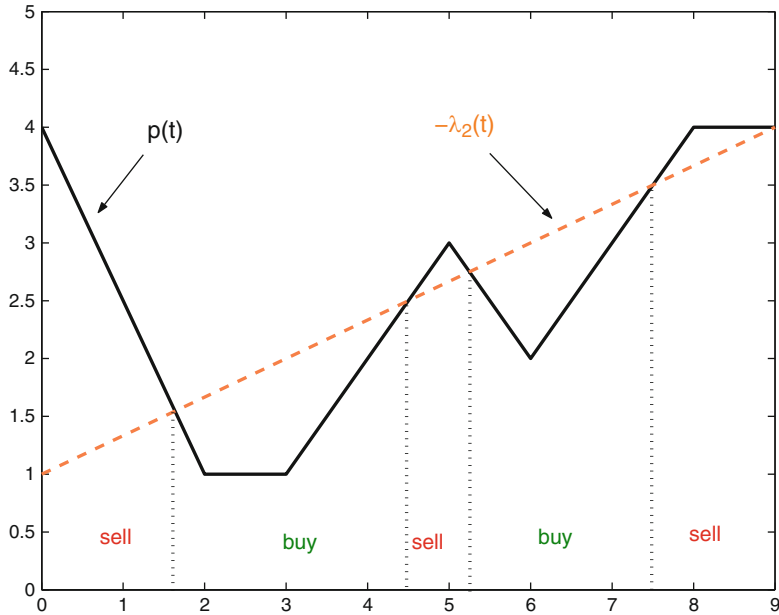
Notice the minus signs in the transversality conditions that arise, since in our formulation, we minimize the objective  $J(u) = -C(u)$ . If  $\lambda_0 = 0$ , then also  $\lambda(t) \equiv 0$ , contradicting the nontriviality of the multiplier, and thus we normalize  $\lambda_0 = 1$ . Hence

$$\lambda_1(t) \equiv -1 \quad \text{and} \quad \lambda_2(t) = \alpha(T - t) - \pi, \quad 0 \leq t \leq T.$$

The minimization condition on the Hamiltonian therefore implies that the optimal control  $u_*(t)$  satisfies

$$u_*(t) = \begin{cases} m & \text{if } p(t) > \alpha(t - T) + \pi, \\ M & \text{if } p(t) < \alpha(t - T) + \pi, \end{cases}$$

while it is not specified through the minimization condition if  $p(t) = \alpha(t - T) + \pi$ . In fact, if the price were to follow this linear relationship, then the minimization condition would be inconclusive, and this leads to the concept of singular controls that we shall describe in Sect. 2.8. Here we consider only the simpler case in which  $p$  is continuous and piecewise continuously differentiable with  $\dot{p}(t) \neq \alpha$  everywhere. In this case, whenever  $p(t) = \alpha(t - T) + \pi$ , then  $p$  crosses the line  $\ell(t) = \alpha(t - T) + \pi$ , and at every such crossing a switch from  $m$  to  $M$  or vice versa occurs. Let us illustrate this solution with a particular case of price function as given below for the numerical values  $T = 9$ ,  $m = -1$ ,  $M = 1$ , and  $\alpha = \frac{1}{3}$ :



**Fig. 2.7** An optimal trading strategy

$$p(t) = \begin{cases} -\frac{3}{2}t + 4 & \text{for } 0 \leq t \leq 2, \\ 1 & \text{for } 2 \leq t \leq 3, \\ t - 2 & \text{for } 3 \leq t \leq 5, \\ -t + 8 & \text{for } 5 \leq t \leq 6, \\ t - 4 & \text{for } 6 \leq t \leq 8, \\ 4 & \text{for } 8 \leq t \leq 9. \end{cases}$$

Choosing  $\pi = 4$ , we get that  $\lambda_2(t) = -\frac{1}{3}t - 1$ , and Fig. 2.7 illustrates the optimal buy–sell decisions for this price function. The optimal control for the problem thus is

$$u_*(t) = \begin{cases} m & \text{for } 0 \leq t \leq 1.64, \\ M & \text{for } 1.64 \leq t \leq 4.5, \\ m & \text{for } 4.5 \leq t \leq 5.25, \\ M & \text{for } 5.25 \leq t \leq 7.5, \\ m & \text{for } 7.5 \leq t \leq 9. \end{cases}$$

### 2.7.2 The Moon-Landing Problem

We now consider a problem with a nonlinear dynamics, but for which the synthesis of optimal controlled trajectories can still easily be obtained with the procedure used for time-invariant linear systems. This is a highly simplified version of the dynamics underlying the real version of a spacecraft making a vertical soft landing on the surface of the moon while minimizing fuel consumption [95]. The state variables are  $h$ , the height of the space craft above the lunar surface;  $v$ , its vertical velocity oriented upward; and  $m$ , its mass. Fuel consumption lowers the mass and because of the orientation, increases the velocity, since the jets are used to slow down the free fall of the craft. The simplified dynamical equations therefore take the form

$$\dot{h} = v, \quad h(0) = h_0, \quad (2.29)$$

$$\dot{v} = -g + \frac{u}{m}, \quad v(0) = v_0, \quad (2.30)$$

$$\dot{m} = -ku, \quad m(0) = M + F, \quad (2.31)$$

where  $g$  is the moon's gravitational constant and  $u$  denotes the control of the system. By means of the constant  $k$ , we normalize the control set to be  $U = [0, 1]$ . The coefficients  $M$  and  $F$  in the initial condition for  $m$  denote the mass of the spacecraft and the total mass of the fuel at the beginning of descent. The optimal control problem then becomes the following:

[ML] For a free terminal time  $T$ , minimize the total amount of fuel used,

$$J(u) = \int_0^T u(t) dt,$$

over all piecewise continuous functions  $u: [0, T] \rightarrow [0, 1]$ , subject to the dynamics (2.29)–(2.31) and terminal conditions

$$h(T) = 0 \quad \text{and} \quad v(T) = 0.$$

Clearly, an implicit assumption in the model is the state constraint  $h \geq 0$ , and obviously we also cannot allow that  $h = 0$  at some intermediate time with negative velocity  $v$ . However, for the moment we ignore these constraints, and it will be seen that the optimal solution fulfills these obvious physical side conditions.

The Hamiltonian function for the moon-landing problem is given as

$$H = \lambda_0 u + \lambda_1 v + \lambda_2 \left( -g + \frac{u}{m} \right) - \lambda_3 k u = \lambda_1 v - \lambda_2 g + u \left( \lambda_0 - \frac{\lambda_2}{m} - \lambda_3 k \right).$$

If  $(x_*, u_*)$  is an optimal controlled trajectory defined over the interval  $[0, T]$ , then there exist a constant  $\lambda_0 \geq 0$  and an adjoint variable  $\lambda = (\lambda_1, \lambda_2, \lambda_3) : [0, T] \rightarrow$

$(\mathbb{R}^3)^*$  such that the following conditions are satisfied: (a)  $\lambda_0$  and  $\lambda(t)$  do not vanish simultaneously over  $[0, T]$ , (b)  $\lambda(t)$  satisfies the adjoint equations

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = -\lambda_1, \quad \dot{\lambda}_3 = \lambda_2 \frac{u}{m^2},$$

with transversality condition  $\lambda_3(T) = 0$ , and (c) the control  $u_*(t)$  minimizes the Hamiltonian  $H$  as a function of  $u$  over the control set  $[0, 1]$  with minimum value 0.

Since the Hamiltonian  $H$  is linear in  $u$ , this minimum is determined by the sign of the function

$$\Phi(t) = \lambda_0 + \frac{\lambda_2(t)}{m(t)} - \lambda_3(t)k,$$

and we have that

$$u_*(t) = \begin{cases} 0 & \text{if } \Phi(t) > 0, \\ \text{undefined} & \text{if } \Phi(t) = 0, \\ 1 & \text{if } \Phi(t) < 0. \end{cases}$$

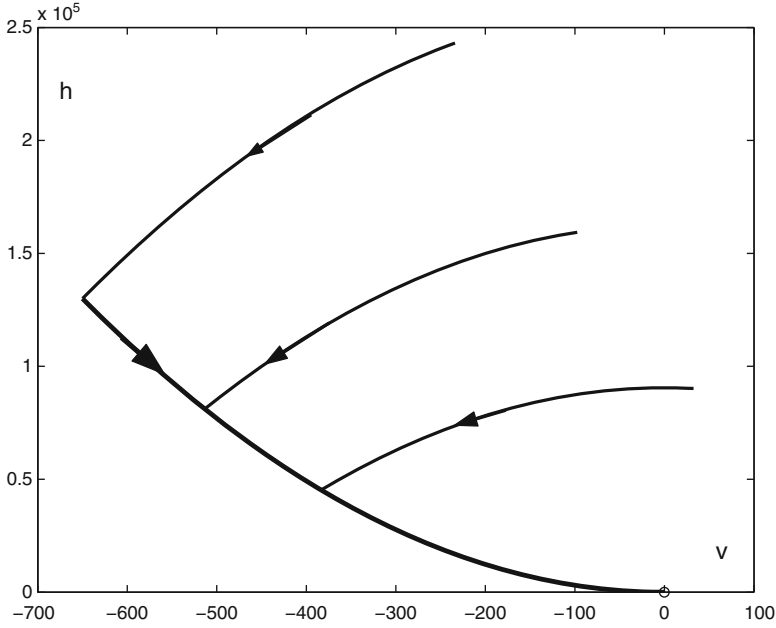
Again  $\Phi$  is the *switching function* of the problem.

For the time-optimal control problem, intuition would say that the optimal solution should be free fall ( $u_* = 0$ ) followed, at the right moment, by a maximum thrust ( $u_* = 1$ ) to slow down the craft to make a soft landing. This corresponds to a bang-bang control that has exactly one switching from  $u = 0$  to  $u = 1$ . For this problem, this also is the minimum-fuel-consumption solution. To see this, we analyze the derivative of the switching function. It follows from the dynamics and adjoint equation that

$$\dot{\Phi}(t) = \frac{\dot{\lambda}_2(t)}{m(t)} - \lambda_2(t) \frac{\dot{m}(t)}{m(t)^2} - \dot{\lambda}_3(t)k = -\frac{\lambda_1}{m(t)}.$$

If  $\lambda_1 = 0$ , then the switching function  $\Phi$  is constant. But  $\Phi$  cannot vanish identically, since the condition that  $H = \lambda_1 v - \lambda_2 g + u\Phi(t) \equiv 0$  then also gives that  $\lambda_2 = 0$ , which implies  $\lambda_3 = 0$  as well and thus also  $\lambda_0 = 0$  from  $\Phi = 0$ , contradicting the nontriviality of the multipliers. Clearly,  $\Phi$  also cannot be positive, since  $v$  decreases along the control  $u \equiv 0$ , and thus we cannot meet the terminal condition  $v = 0$ . Hence, in this case,  $\Phi$  must be negative, giving the constant control  $u_*(t) \equiv 1$ . This corresponds to braking with full thrust throughout, and clearly this is the optimal control for specific initial conditions. If  $\lambda_1 \neq 0$ , then the switching function is strictly monotone and thus has at most one zero. Again, the only choice that can satisfy the terminal condition is  $\lambda_1 > 0$ , and hence optimal controls must be bang-bang with at most one switching from  $u = 0$  to  $u = 1$ .

Once this is known, a field of extremals can be constructed as before in the examples for linear systems. Suppose the control is given by  $u_* \equiv 1$  on the interval  $[\zeta, T]$ . It then follows from the terminal conditions  $h(T) = 0$  and  $v(T) = 0$  that



**Fig. 2.8** Switching curve and optimal controlled trajectories near the final time  $T$

$$h(\zeta) = -\frac{1}{2}g(T-\zeta)^2 - \frac{M+F}{k^2} \ln \left( 1 - \frac{k(T-\zeta)}{M+F} \right) - \frac{T-\zeta}{k},$$

$$v(\zeta) = g(T-\zeta) + \frac{1}{k} \ln \left( 1 - \frac{k(T-\zeta)}{M+F} \right).$$

Plotting  $h(\zeta)$  against  $v(\zeta)$ , we get a curve  $\mathcal{Z}$ ,  $\zeta \mapsto (h(\zeta), v(\zeta))$ , that represents the set of all initial conditions (height and velocity pairs) that would result in a soft landing with full thrust  $u_* \equiv 1$ . Since there exists a restriction that the total amount of fuel will burn in time  $\frac{F}{k}$  seconds, one further needs to restrict the curve to the  $\zeta$ -values in the interval  $[T - \frac{F}{k}, T]$ . The first part of the trajectory simply is free fall ( $u_* = 0$ ), and the initial portions of the equations are given by

$$h(t) = -\frac{1}{2}gt^2 + v_0t + h_0 \quad \text{and} \quad v(t) = -gt + v_0$$

so that

$$h(t) = h_0 - \frac{1}{2g} (v^2(t) - v_0^2), \quad t \geq 0. \quad (2.32)$$

Once this parabola meets the curve  $\mathcal{Z}$ , the thrusters need to be engaged at full force to make a soft landing. If the parabola does not meet the switching curve  $\mathcal{Z}$ , a soft landing is impossible. Figure 2.8 illustrates the synthesis.



## 2.8 Singular Controls and Lie Derivatives

Both the linear time-optimal control problems in the plane and also the examples considered in the previous section lead to minimizing a Hamiltonian function that is linear in a scalar control  $u$  over a compact interval  $[a, b]$ . Clearly, this minimum is attained at  $u = a$  if the function  $\Phi$  multiplying  $u$  is positive and at  $u = b$  if this function is negative. In the examples we have considered so far, it always turned out that optimal controls were *bang-bang*, i.e., consisted of a finite number of switchings between  $u = a$  and  $u = b$ . We shall show in Sect. 3.6, that this is “always” the case for a time-invariant linear system whose control set is a compact polyhedron. More precisely, for these systems, it is always possible to find an optimal control that switches finitely many times between controls that take their values in one of the vertices of the control set; in addition, the number of switchings over a finite interval  $[0, T]$  can be bounded. This no longer holds once the dynamics of the control system becomes nonlinear: optimal controls need not be bang-bang, and even when optimal controls switch only between  $u = a$  and  $u = b$ , the number of switchings on a compact interval  $[0, T]$  can be countably infinite. We shall see in the remaining sections of this chapter that these phenomena are linked with controls that arise when the function  $\Phi$  multiplying  $u$  vanishes over some interval, so-called *singular controls*. We now develop geometric tools and techniques required for their analysis.

### 2.8.1 Time-Optimal Control for a Single-Input Control-Affine Nonlinear System

Again we use the time-optimal control problem as the vehicle to develop these tools, but now allow for nonlinearities in the state. We consider a time-invariant, single-input, control-affine system  $\Sigma$  of the form

$$\Sigma : \quad \dot{x} = f(x) + g(x)u, \quad f, g \in V^\infty(\Omega), \quad x \in \Omega. \quad (2.33)$$

Here,  $\Omega$  is a domain (i.e., an open and connected subset) in  $\mathbb{R}^n$ , and  $f : \Omega \rightarrow \mathbb{R}^n$  and  $g : \Omega \rightarrow \mathbb{R}^n$  are two infinitely often continuously differentiable vector fields defined on  $\Omega$ . We use  $V^r(\Omega)$  to denote the set of all vector fields defined on  $\Omega$  for which all components are  $C^r(\Omega)$ -functions, i.e., are defined and  $r$  times continuously differentiable on  $\Omega$ . Clearly, the  $C^\infty$  assumption is without loss of generality and can be replaced by requiring that the vector fields be sufficiently smooth, say  $f, g \in V^r(\Omega)$ , with  $r$  large enough for all the derivatives that arise to exist.

[NTOC] Given a time-invariant, single-input, control-affine control system  $\Sigma$  of the form (2.33), among all piecewise continuous (more generally, Lebesgue measurable) controls  $u$  that take values in the compact interval  $[-1, 1]$ ,  $u : [0, T] \rightarrow [-1, 1]$ , find one that steers a given initial point  $p \in \Omega$  into a target point  $q \in \Omega$  in minimum time.

In the formulation of Sect. 2.2, we have that  $M = \Omega$ ,  $L(t, x, u) \equiv 1$ ,  $f(t, x, u) = f(x) + g(x)u$ ,  $\varphi \equiv 0$ , and  $\Psi$  is given by  $\Psi : [0, \infty) \times \Omega \rightarrow \Omega$ ,  $(t, x) \mapsto \Psi(t, x) = x - q$ , i.e.,  $N = \{q\}$ . As with the linear system, since both initial and terminal points on the state are specified, the transversality conditions give no information about the multiplier  $\lambda$ , and the terminal value  $\lambda(T)$  is free. But the transversality condition on the final time  $T$  implies that  $H(T, \lambda_0, \lambda, x, u) = 0$ , and since

$$H = \lambda_0 + \lambda (f(x) + g(x)u) = \lambda_0 + \langle \lambda, f(x) + g(x)u \rangle$$

is time-invariant, it follows that the Hamiltonian vanishes identically along any extremal. Equivalently, we have that

$$\langle \lambda(t), f(x_*(t)) + u_*(t)g(x_*(t)) \rangle \equiv \text{const} \leq 0.$$

We freely use the notation  $\langle \cdot, \cdot \rangle$  for the inner product. In particular, note that this implies that  $\lambda(t) \neq 0$ , since otherwise also  $\lambda_0 = 0$  contradicting the nontriviality condition of the maximum principle. The adjoint equation is given by

$$\dot{\lambda}(t) = -\lambda(t) (Df(x_*(t)) + u_*(t)Dg(x_*(t))),$$

where  $Df$  and  $Dg$  denote the matrices of the partial derivatives of the vector fields  $f$  and  $g$ , respectively, and the minimum condition implies that

$$u_*(t) = \begin{cases} +1 & \text{if } \langle \lambda(t), g(x_*(t)) \rangle < 0, \\ -1 & \text{if } \langle \lambda(t), g(x_*(t)) \rangle > 0. \end{cases}$$

Summarizing, we thus have the following result:

**Theorem 2.8.1 (Maximum principle for problem (NTOC)).** *Let  $(x_*, u_*)$  be a controlled trajectory defined over the interval  $[0, T]$ . If  $(x_*, u_*)$  minimizes the time of transfer from  $p \in \Omega$  to  $q \in \Omega$ , then there exists a nontrivial solution  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$  to the adjoint equation*

$$\dot{\lambda}(t) = -\lambda(t) (Df(x_*(t)) + u_*(t)Dg(x_*(t))) \quad (2.34)$$

such that

$$\langle \lambda(t), f(x_*(t)) + u_*(t)g(x_*(t)) \rangle \equiv \text{const} \leq 0$$

and the control  $u_*$  satisfies

$$u_*(t) = -\text{sgn} \langle \lambda(t), g(x_*(t)) \rangle.$$

## 2.8.2 The Switching Function and Singular Controls

**Definition 2.8.1 (Switching function).** Let  $\Gamma$  be an extremal lift for the problem [NTOC] consisting of a controlled trajectory  $(x_*, u_*)$  defined over the interval  $[0, T]$  with corresponding adjoint vector  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$ . The function

$$\Phi_\Gamma(t) = \lambda(t)g(x_*(t)) = \langle \lambda(t), g(x_*(t)) \rangle \quad (2.35)$$

is called the (corresponding) switching function.

We usually drop the subscript  $\Gamma$  in the notation if the extremal under consideration is understood. Clearly, properties of the switching function  $\Phi$  determine the structure of the optimal controls. As long as  $\Phi$  is not zero, the optimal control is simply given by

$$u_*(t) = -\operatorname{sgn} \Phi(t)$$

and thus takes its value in one of the vertices of the control set. A priori, the control is not determined by the minimum condition at times when  $\Phi(t) = 0$ . Naturally, if  $\Phi(\tau) = 0$  and the derivative  $\dot{\Phi}(\tau)$  exists and does not vanish, then the control switches between  $u = -1$  and  $u = +1$  with the order depending on the sign of  $\dot{\Phi}(\tau)$ . Such a time  $\tau$  simply is a bang-bang junction, exactly as with linear systems. On the other hand, if  $\Phi(t)$  were to vanish identically on an open interval  $I$ , then although the minimization property by itself gives no information about the control, in this case, also all the derivatives of  $\Phi(t)$  must vanish, and this, except for some degenerate situations, generally does determine the control as well. Controls of this kind are the *singular* controls referred to above, while we refer to the constant controls  $u = -1$  and  $u = +1$  as *bang* controls. Strictly speaking, to be singular is not a property of the control, but of the extremal lift, since it clearly also depends on the multiplier  $\lambda$  defining the switching function.

**Definition 2.8.2 (Singular controls and extremals).** Let  $\Gamma$  be an extremal lift for the problem [NTOC] consisting of a controlled trajectory  $(x_*, u_*)$  defined over the interval  $[0, T]$  with corresponding adjoint vector  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$ . We say that the control  $u$  is singular on an open interval  $I \subset [0, T]$  if the switching function vanishes identically on  $I$ . The corresponding portion of the trajectory  $x$  defined over  $I$  is called a singular arc, and  $\Gamma$  a singular extremal (respectively, singular extremal lift).

Historically, this terminology has its origin in the following simple observation: in terms of the Hamiltonian  $H$  for problem [NTOC], the switching function can be expressed as

$$\Phi(t) = \frac{\partial H}{\partial u}(\lambda_0, \lambda(t), x_*(t), u_*(t)),$$

and thus the condition  $\Phi(t) = 0$  formally is the first-order necessary condition for the Hamiltonian to have a minimum in the interior of the control set. For a general optimal control problem, extremal lifts are called singular, respectively nonsingular, over an open interval  $I$  if the first-order necessary condition

$$\frac{\partial H}{\partial u}(\lambda_0, \lambda(t), x_*(t), u_*(t)) = 0$$

is satisfied for  $t \in I$  and if the matrix of the second-order partial derivatives,

$$\frac{\partial^2 H}{\partial u^2}(\lambda_0, \lambda(t), x_*(t), u_*(t)),$$

is singular, respectively nonsingular, on  $I$ . For problem [NTOC], this quantity is always zero, and thus any optimal control that takes values in the interior of the control set is necessarily singular. On the other hand, for example, for the linear-quadratic optimal control problem [LQ] considered earlier, this matrix is always positive definite and all extremals are nonsingular.

In order to solve the problem [NTOC], optimal controls need to be synthesized from bang and singular controls, the potential candidates for optimality, through an analysis of the zero set  $\mathcal{Z}_\Phi$  of the switching function,

$$\mathcal{Z}_\Phi = \{t \in [0, T] : \Phi(t) = 0\}.$$

This, however, can become a very difficult problem, since a priori, all we know about  $\mathcal{Z}_\Phi$  is that it is a closed set.

**Proposition 2.8.1.** *Given any closed subset  $Z \subset \mathbb{R}^n$ , there exists a nonnegative  $C^\infty$ -function  $\varphi$  such that  $Z = \{y \in \mathbb{R}^n : \varphi(y) = 0\}$ .*

*Proof.* [108] Let  $B = Z^c$ , the complement of  $Z$ . Since  $\mathbb{R}^n$  is second countable (see Appendix C), there exists a sequence of open balls  $B_i$ ,  $i \in \mathbb{N}$ , such that  $B = \bigcup_{i \in \mathbb{N}} B_i$ . It is a standard calculus exercise to verify that the function  $\Gamma$  defined by

$$\Gamma(y) = \begin{cases} \exp\left(-\frac{1}{(y-1)^2}\right) & \text{for } y < 1, \\ 0 & \text{for } y \geq 1, \end{cases}$$

is  $C^\infty$ : derivatives of arbitrary order exist and all derivatives at  $y = 1$  from the left vanish. Let  $D = B_r(p)$  be an open ball with radius  $r$  centered at  $p$ . Defining a radially symmetric function  $\psi : D \rightarrow \mathbb{R}^n$  by

$$\psi(y) = \Gamma\left(\frac{\|y - p\|^2}{r^2}\right),$$

it then follows that  $\psi$  is nonnegative,  $\psi \in C^\infty(D)$ , and  $\psi$  vanishes identically outside of  $D$ . For each open ball  $B_i$ ,  $i \in \mathbb{N}$ , let  $\psi_i$  be the correspondingly defined function. For a multi-index  $\alpha = (\alpha_1, \dots, \alpha_n)$ ,  $\alpha_i \in \mathbb{N}$ , let  $|\alpha| = \alpha_1 + \dots + \alpha_n$  and denote the corresponding partial derivatives of  $\psi_i$  by  $D^\alpha \psi_i$ ,

$$D^\alpha \psi_i = \frac{\partial^{\alpha_1} \dots \partial^{\alpha_n} \psi_i}{\partial y^{\alpha_1} \dots \partial y^{\alpha_n}}.$$

Let

$$M_i = \sup_{|\alpha| \leq i} \|D^\alpha \psi_i\|;$$

since all functions  $\psi_i$  have compact support and the summation is finite, all the numbers  $M_i$  are finite. Thus the series

$$\varphi(y) = \sum_{i=1}^{\infty} \frac{\psi_i}{2^i M_i}$$

converges uniformly (we have  $\|\psi_i\| \leq M_i$  for all  $i \in \mathbb{N}$ ), and so do all its partial derivatives. For since also  $\|D^\alpha \psi_i\| \leq M_i$  for all  $i \geq |\alpha|$ , the termwise differentiated series

$$\sum_{i=1}^{\infty} \frac{D^\alpha \psi_i}{2^i M_i}$$

converges uniformly and its limit is the  $\alpha$ th derivative of  $\varphi$ ,  $D^\alpha \varphi$ . Thus  $\varphi$  is a  $C^\infty$ -function that is positive on each ball  $B_i$  and vanishes identically outside  $\cup_{i \in \mathbb{N}} B_i$ , i.e., on  $Z$ .  $\square$

Thus, in principle, the zero set  $\mathcal{Z}_\Phi$  of the switching function can be an arbitrary closed subset of the interval  $[0, T]$ , and a better understanding of this set is needed to solve the optimal control problem [NTOC]. In order to achieve this, we now analyze the derivatives of the switching function. Since both  $\lambda$  and the state  $x$  satisfy differential equations, the switching function  $\Phi$  is differentiable, and we obtain

$$\begin{aligned} \dot{\Phi}(t) &= \dot{\lambda}(t)g(x_*(t)) + \lambda(t)Dg(x_*(t))\dot{x}_*(t) \\ &= -\lambda(t)[Df(x_*(t)) + u_*(t)Dg(x_*(t))]g(x_*(t)) \\ &\quad + \lambda(t)Dg(x_*(t))f(x_*(t)) + u_*(t)g(x_*(t)) \\ &= \lambda(t)Dg(x_*(t))f(x_*(t)) - Df(x_*(t))g(x_*(t)). \end{aligned} \quad (2.36)$$

The coefficients at  $u_*(t)$  cancel, and thus the derivative of the switching function does not depend on the control  $u_*$ . Hence  $\dot{\Phi}(t)$  is once more differentiable, and we can iterate this calculation to find higher-order derivatives. This very much is the approach pursued in older textbooks on the subject. However, brute force is not necessarily always a good strategy, and now it is of benefit to develop the proper formalism. The key is to observe that the tangent vector that multiplies  $\lambda$  in (2.36) is the coordinate expression of the *Lie bracket* of the vector fields  $f$  and  $g$ , and this quantity is of fundamental importance in the control of nonlinear systems. We therefore digress to give some of the background that not only is fundamental for nonlinear optimal control theory in general, but also provides us with an elegant and transparent scheme to carry out the required calculations.

### 2.8.3 Lie Derivatives and the Lie Bracket

As before, let  $\Omega$  be a domain in  $\mathbb{R}^n$  and denote the space of all infinitely often continuously differentiable functions on  $\Omega$  by  $C^\infty(\Omega)$ . Also let  $X : \Omega \rightarrow \mathbb{R}^n$  be a  $C^\infty$  vector field defined on  $\Omega$ ,  $X \in V^\infty(\Omega)$ . As before, the assumption  $r = \infty$  is taken for simplicity of notation, and it suffices to have all functions and vector fields to be  $r$ -times continuously differentiable with the blanket assumption that  $r$  is large enough for all the required differentiations to be permissible. The vector field  $X$  can be viewed as defining a first-order differential operator from the space  $C^\infty(\Omega)$  into  $C^\infty(\Omega)$  by taking at every point  $q \in \Omega$  the directional derivative of a function  $\varphi \in C^\infty(\Omega)$  in the direction of the vector field  $X(q)$ , i.e.,

$$X : C^\infty(\Omega) \rightarrow C^\infty(\Omega), \quad \varphi \mapsto X\varphi,$$

defined by

$$(X\varphi)(q) = \nabla\varphi(q) \cdot X(q),$$

where  $\nabla\varphi$  denotes the gradient of the function  $\varphi$ , as always written as a row vector. While this is a convenient notation, which we freely use, in order to distinguish the values of the vector field from its action when considered as an operator, it is more customary to denote this operator by  $L_X$ , i.e.,  $L_X(\varphi)(q) = (X\varphi)(q)$ , and this function is called the *Lie derivative* of the function  $\varphi$  along the vector field  $X$ .

**Definition 2.8.3 (Lie bracket).** The Lie bracket of two vector fields  $X$  and  $Y$  defined on  $\Omega$  is the operator defined by the commutator

$$[X, Y] = X \circ Y - Y \circ X = XY - YX.$$

Formally, this is a second-order differential operator. But in fact, all second-order terms cancel, and the Lie bracket defines another first-order differential operator. For if we denote the Hessian matrix of a function  $\varphi$  by  $H(\varphi)$  and the action of this symmetric matrix on the vector fields  $X$  and  $Y$  by  $H(\varphi)(X, Y)$ , then we simply have that

$$\begin{aligned} [X, Y](\varphi) &= X(Y\varphi) - Y(X\varphi) = X(\nabla\varphi \cdot Y) - Y(\nabla\varphi \cdot X) \\ &= \nabla(\nabla\varphi \cdot Y) \cdot X - \nabla(\nabla\varphi \cdot X) \cdot Y \\ &= H(\varphi)(Y, X) + \nabla\varphi \cdot DY \cdot X - H(\varphi)(X, Y) - \nabla\varphi \cdot DX \cdot Y \\ &= \nabla\varphi \cdot (DY \cdot X - DX \cdot Y). \end{aligned}$$

This calculation verifies that if  $X : \Omega \rightarrow \mathbb{R}^n$ ,  $z \mapsto X(z)$ , and  $Y : \Omega \rightarrow \mathbb{R}^n$ ,  $z \mapsto Y(z)$ , are coordinates for these vector fields, then the coordinate expression for the Lie bracket is given by

$$[X, Y](z) = DY(z) \cdot X(z) - DX(z) \cdot Y(z). \quad (2.37)$$

This computation directly extends to calculating Lie brackets if we consider  $C^\infty$  vector fields as a module over  $C^\infty(\Omega)$ , i.e., multiply the vector fields by smooth functions.

**Lemma 2.8.1.** *Suppose  $\alpha$  and  $\beta$  are smooth functions on  $\Omega$ ,  $\alpha, \beta \in C^\infty(\Omega)$ , and  $X$  and  $Y$  are  $C^\infty$  vector fields on  $\Omega$ . Then*

$$[\alpha X, \beta Y] = \alpha\beta[X, Y] + \alpha(L_X\beta)Y - \beta(L_Y\alpha)X.$$

*Proof.* This simply follows from the product rule:

$$\begin{aligned} [aX, \beta Y] &= (\alpha X(\beta Y)) - (\beta Y(\alpha X)) \\ &= \alpha\{(X\beta)Y + \beta XY\} - \beta\{(Y\alpha)X + \alpha YX\} \\ &= \alpha\beta[X, Y] + \alpha(X\beta)Y - \beta(Y\alpha)X. \end{aligned}$$

□

A more important, and less obvious identity is the Jacobi identity.

**Proposition 2.8.2.** *For any  $C^\infty$  vector fields  $X, Y$ , and  $Z$  defined on  $\Omega$  we have that*

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] \equiv 0.$$

*Proof.* Again, computing as operators, we have that

$$\begin{aligned} [X, [Y, Z]] &= X[Y, Z] - [Y, Z]X \\ &= X(YZ - ZY) - (YZ - ZY)X \\ &= XYZ - XZY - YZX + ZYX. \end{aligned}$$

Adding the corresponding terms for the other brackets thus gives

$$\begin{aligned} &[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] \\ &\equiv (XYZ - XZY - YZX + ZYX) + (YZX - YXZ - ZXY + XZY) \\ &\quad + (ZXY - ZYX - XYZ + YXZ), \end{aligned}$$

and all terms cancel.

□

Note that the Jacobi identity can be written in the form

$$[X, [Y, Z]] = [[X, Y], Z] + [Y, [X, Z]],$$

and this simply states that taking the Lie bracket with  $X$  (i.e., the Lie derivative of a vector field along  $X$ ) satisfies the product rule. These rules show that the vector fields, understood as differential operators, form a Lie algebra. A *Lie algebra* over

$\mathbb{R}$  is a real vector space  $\mathfrak{G}$  together with a bilinear operator  $[\cdot, \cdot] : \mathfrak{G} \times \mathfrak{G} \rightarrow \mathfrak{G}$  such that for all  $X, Y$ , and  $Z \in \mathfrak{G}$  we have  $[X, Y] = -[Y, X]$  and  $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$ . Many of the essential concepts and computational tools that will be developed in Sect. 4.5 depend only on these general identities abstracted from the above properties of vector fields.

These notions allow us to restate the formula for the derivative of the switching function in a more general format, equally simple, but of great importance.

**Theorem 2.8.2.** *Let  $Z : \Omega \rightarrow \mathbb{R}^n$  be a differentiable vector field defined on  $\Omega$  and let  $(x, u)$  be a controlled trajectory defined over an interval  $I$  with trajectory in  $\Omega$ . Let  $\lambda$  be a solution to the corresponding adjoint equation and define the function*

$$\Psi(t) = \langle \lambda(t), Z(x(t)) \rangle.$$

*Then  $\Psi$  is differentiable with derivative given by*

$$\dot{\Psi}(t) = \langle \lambda(t), [f + ug, Z](x(t)) \rangle.$$

*Proof.* This is the same calculation as above. Note that for any row vector  $\lambda \in (\mathbb{R}^n)^*$ , matrix  $A \in \mathbb{R}^{n \times n}$ , and column vector  $x \in \mathbb{R}^n$  we have that  $\langle \lambda, Ax \rangle = \lambda Ax = \langle \lambda A, x \rangle$ . Thus, and dropping the argument  $t$  in the calculation, we have that

$$\begin{aligned} \dot{\Psi}(t) &= \left\langle \dot{\lambda}, Z(x) \right\rangle + \langle \lambda, DZ(x)\dot{x} \rangle \\ &= -\langle \lambda (Df(x) + uDg(x)), Z(x) \rangle + \langle \lambda, DZ(x)(f(x) + ug(x)) \rangle \\ &= \langle \lambda, DZ(x)f(x) - Df(x)Z(x) \rangle + u \langle \lambda, DZ(x)g(x) - Dg(x)Z(x) \rangle \\ &= \langle \lambda, [f, Z](x) \rangle + u \langle \lambda, [g, Z](x) \rangle, \end{aligned}$$

which, for simplicity of notation, we also write as  $\dot{\Psi}(t) = \langle \lambda(t), [f + ug, Z](x(t)) \rangle$ , noting that  $u(t)$  simply is a real number under differentiation with respect to the state variables involved in the calculation of the Lie brackets.  $\square$

### 2.8.4 The Order of a Singular Control and the Legendre–Clebsch Conditions

It follows from Theorem 2.8.2 that the first and second derivatives of the switching function  $\Phi(t) = \langle \lambda(t), g(x(t)) \rangle$  are given by

$$\dot{\Phi}(t) = \langle \lambda(t), [f, g](x(t)) \rangle \quad (2.38)$$

and

$$\ddot{\Phi}(t) = \langle \lambda(t), [f, [f, g]](x(t)) \rangle + u(t) \langle \lambda(t), [g, [f, g]](x(t)) \rangle. \quad (2.39)$$



If now  $\Gamma = ((x, u), \lambda)$  is an extremal lift for which the control is singular on an open interval  $I$ , then all derivatives of  $\Phi$  vanish identically on  $I$ , so that we have

$$\langle \lambda(t), [f, g](x(t)) \rangle \equiv 0$$

and

$$\langle \lambda(t), [f, [f, g]](x(t)) \rangle + u(t) \langle \lambda(t), [g, [f, g]](x(t)) \rangle \equiv 0.$$

Clearly, at times  $t$  when  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$  does not vanish, this equation determines the singular control, and this leads to the following definition:

**Definition 2.8.4 (Order 1 singular control).** Let  $\Gamma = ((x, u), \lambda)$  be an extremal lift for the problem [NTOC] consisting of a controlled trajectory  $(x, u)$  defined over the interval  $[0, T]$  with corresponding adjoint vector  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$ . If  $\Gamma$  is a singular extremal lift over an open interval  $I$ , then  $\Gamma$ , and also the control  $u$ , are said to be singular of order 1 over  $I$  if  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$  does not vanish on the interval  $I$ .

We thus immediately have the following formula for the singular control in terms of the state and multiplier.

**Proposition 2.8.3.** *If  $\Gamma = ((x, u), \lambda)$  is a singular extremal lift of order 1 over an open interval  $I$ , then the singular control is given by*

$$u_{\text{sing}}(t) = - \frac{\langle \lambda(t), [f, [f, g]](x(t)) \rangle}{\langle \lambda(t), [g, [f, g]](x(t)) \rangle}. \quad (2.40)$$

Note that this formula defines the singular control as a function of the state and the multiplier, and thus it depends on the extremal lift. In differential-geometric terms, it defines the singular control in the cotangent bundle (see Appendix C). Generally, it is not a feedback function in the state space. However, more can be said in low dimensions  $n$  of the state space and this will be pursued later on. Naturally, this formula in no way guarantees that the control bounds imposed in the problem are satisfied, and thus  $u_{\text{sing}}(t)$  is admissible only if the values of this expression lie in the control set, the interval  $[-1, 1]$ .

Similar to the Legendre condition in the calculus of variations, for singular controls a generalized version of the Legendre condition also is necessary for optimality. This result will be proven in Sect. 4.6.1.

**Theorem 2.8.3 (Legendre–Clebsch condition).** *Suppose the controlled trajectory  $(x_*, u_*)$  defined over the interval  $[0, T]$  minimizes the time of transfer from  $p \in \Omega$  to  $q \in \Omega$  for problem [NTOC], and the control  $u_*$  is singular over an open interval  $I \subset [0, T]$ . Then there exists an extremal lift  $\Gamma = ((x_*, u_*), \lambda)$  with the property that*

$$\langle \lambda(t), [g, [f, g]](x(t)) \rangle \leq 0 \quad \text{for all } t \in I.$$

**Definition 2.8.5 (Strengthened Legendre–Clebsch condition).** Let  $\Gamma = ((x, u), \lambda)$  be an extremal lift for the problem [NTOC] consisting of a controlled trajectory  $(x, u)$

defined over the interval  $[0, T]$  and corresponding adjoint vector  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$  that is singular of order 1 over an open interval  $I$ . We say that the strengthened Legendre–Clebsch condition is satisfied along  $\Gamma$  over  $I$  if  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$  is negative on  $I$ , and that it is violated if this expression is positive.

An important property of singular extremals that satisfy the strengthened Legendre–Clebsch condition is that if the singular control takes values in the interior of the control set, then at any time  $t \in I$ , it can be concatenated with either of the two bang controls  $u = -1$  and  $u = +1$  in the sense that this generates junctions that satisfy the conditions of the maximum principle. As before, let  $X = f - g$  and  $Y = f + g$  denote the corresponding vector fields. We write  $XS$  for a concatenation of a trajectory corresponding to the control  $u = -1$  with a singular arc; i.e., for some  $\varepsilon > 0$  the control is given by

$$u(t) = \begin{cases} -1 & \text{for } t \in (\tau - \varepsilon, \tau), \\ u_{\text{sing}}(t) & \text{for } t \in [\tau, \tau + \varepsilon). \end{cases}$$

The time  $\tau$  is called a junction time, and the corresponding point  $x(\tau)$  a junction point. Similarly, concatenations of the type  $YS$ ,  $SX$ , and  $SY$  are defined, and we use the symbol  $B$  to denote any one of  $X$  or  $Y$ .

**Proposition 2.8.4.** *Let  $\Gamma = ((x, u), \lambda)$  be an extremal lift for the problem [NTOC] defined over the interval  $[0, T]$  that is singular over an open interval  $I$  and suppose the strengthened Legendre–Clebsch condition is satisfied on  $I$ . If the singular control at the time  $\tau \in I$  has a value in the open interval  $(-1, 1)$ , then there exists an  $\varepsilon > 0$  such that any concatenation of the singular control with a bang control  $u = -1$  or  $u = +1$  at time  $\tau$  satisfies the necessary conditions of the maximum principle; i.e., concatenations of the types  $BS$  and  $SB$  are allowed.*

*Proof.* It follows from Eq. (2.40) that the singular control is continuous if the strengthened Legendre–Clebsch condition is satisfied and so trivially are the constant controls  $u = \pm 1$ . For any control  $u$  that is continuous from the left (–) or right (+), the second derivative of the switching function at time  $\tau$  is given by

$$\ddot{\Phi}(\tau_{\pm}) = \langle \lambda(\tau), [f, [f, g]](x(\tau)) \rangle + u(\tau_{\pm}) \langle \lambda(\tau), [g, [f, g]](x(\tau)) \rangle,$$

and it vanishes identically along the singular control. If the strengthened Legendre–Clebsch condition is satisfied, then we have  $\lambda(\tau)[g, [f, g]](x(\tau)) < 0$ . By assumption, the singular control takes values in the interior of the control set,  $u(\tau) \in (-1, 1)$ , and thus we get for  $u = -1$  that

$$\begin{aligned} \ddot{\Phi}(\tau_{\pm}) &= \langle \lambda(\tau), [X, [f, g]](x(\tau)) \rangle \\ &= \langle \lambda(\tau), [f, [f, g]](x(\tau)) \rangle - \langle \lambda(\tau), [g, [f, g]](x(\tau)) \rangle \\ &> \langle \lambda(\tau), [f, [f, g]](x(\tau)) \rangle + u(\tau_{\pm}) \langle \lambda(\tau), [g, [f, g]](x(\tau)) \rangle = 0, \end{aligned}$$

and for  $u = +1$  we have

$$\begin{aligned}\ddot{\Phi}(\tau_{\pm}) &= \langle \lambda(\tau), [Y, [f, g]](x(\tau)) \rangle \\ &= \langle \lambda(\tau), [f, [f, g]](x(\tau)) \rangle + \langle \lambda(\tau), [g, [f, g]](x(\tau)) \rangle \\ &< \langle \lambda(\tau), [f, [f, g]](x(\tau)) \rangle + u(\tau_{\pm}) \langle \lambda(\tau), [g, [f, g]](x(\tau)) \rangle = 0.\end{aligned}$$

For each control, these signs are consistent with both entry and exit from the singular arc. For example, if  $u = -1$  on an interval  $(\tau - \varepsilon, \tau)$ , then the switching function has a local minimum at time  $t = \tau$  with minimum value 0, and thus  $\Phi$  is positive over this interval, consistent with the minimum condition of the maximum principle.  $\square$

The order of a singular control over an interval  $I$  need not be constant, since the function  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$  may vanish on some portions of  $I$ . If these are isolated times, then typically at those times the local optimality status of the singular control changes from minimizing to maximizing, and the resulting subintervals simply need to be analyzed separately. A more degenerate situation would arise if  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$  were to vanish identically on some subinterval  $J \subset I$ . In this case, many more relations need to be satisfied for the conditions to be consistent. Since we have both

$$\langle \lambda(t), [f, [f, g]](x(t)) \rangle \equiv 0 \quad \text{and} \quad \langle \lambda(t), [g, [f, g]](x(t)) \rangle \equiv 0 \quad \text{for all } t \in J,$$

differentiating both identities, we get the following two equations on  $J$ :

$$\begin{aligned}0 &\equiv \langle \lambda(t), [f, [f, [f, g]]](x(t)) \rangle + u(t) \langle \lambda(t), [g, [f, [f, g]]](x(t)) \rangle, \\ 0 &\equiv \langle \lambda(t), [f, [g, [f, g]]](x(t)) \rangle + u(t) \langle \lambda(t), [g, [g, [f, g]]](x(t)) \rangle.\end{aligned}$$

Each condition by itself determines the control if the functions multiplying the control  $u(t)$  are not zero. Since the pair  $(1, u(t))$  is a nontrivial solution to this homogeneous system, however, we also need to have the compatibility condition

$$\langle \lambda(t), [f, [f, [f, g]]](x(t)) \rangle \langle \lambda(t), [g, [g, [f, g]]](x(t)) \rangle = \langle \lambda(t), [g, [f, [f, g]]](x(t)) \rangle^2,$$

where we use that by the Jacobi identity,

$$[g, [f, [f, g]]] = [f, [g, [f, g]]].$$

It is clear that these are increasingly more and more demanding requirements for the singular extremal to satisfy, and it seems plausible that “typically” these conditions should be difficult to satisfy, even in higher dimensions. This indeed is correct and can be made precise in the sense that “generically” singular extremals are of order 1, as shown by Bonnard and Chyba in [44, Sects. 8.3 and 8.5].

While this result, and also the results by Chitour, Jean and Trélat [71, 72] imply that we should not expect higher-order singular extremals for too many systems, this

does not mean that these do not exist nor that these may not be of particular interest for some specific problem. One common way in which these higher-order singular extremals arise is that the control vector field  $g$  and the Lie bracket  $[f, g]$  commute, i.e., that

$$[g, [f, g]] = 0.$$

In this case, the brackets  $[f, [g, [f, g]]]$  and  $[g, [g, [f, g]]]$  also are zero, and thus the calculation of the derivatives of the switching function simply continues as

$$\Phi^{(3)}(t) = \langle \lambda(t), [f, [f, [f, g]]](x(t)) \rangle = 0$$

and

$$\Phi^{(4)}(t) = \langle \lambda(t), [f, [f, [f, [f, g]]]](x(t)) \rangle + u(t) \langle \lambda(t), [g, [f, [f, [f, g]]]](x(t)) \rangle = 0. \quad (2.41)$$

This seems an adequate place to introduce a shorter notation for the iterated Lie brackets. It is common (for reasons that are connected with what is called the adjoint representation in Lie theory [256]) to think of taking the Lie bracket of a fixed vector field  $X$  with another vector field as a linear operator on the set of all smooth vector fields defined on  $\Omega$ ,  $V^\infty(\Omega)$ , and to denote it by  $\text{ad}_X$ ,

$$\text{ad}_X : V^\infty(\Omega) \rightarrow V^\infty(\Omega), \quad Y \mapsto \text{ad}_X(Y) = [X, Y].$$

The composition of these operators is then defined as

$$\text{ad}_X^i = \text{ad}_X^{i-1} \circ \text{ad}_X,$$

so that, for example, we have

$$[f, [f, [f, [f, g]]]] = \text{ad}_f^4(g).$$

In this notation, Eq. (2.41) can be written more compactly as

$$\Phi^{(4)}(t) = \langle \lambda(t), \text{ad}_f^4(g)(x(t)) \rangle + u(t) \langle \lambda(t), [g, \text{ad}_f^3(g)](x(t)) \rangle = 0.$$

**Definition 2.8.6 (Higher-order singular control).** Let  $\Gamma$  be an extremal lift for the problem [NTOC] consisting of a controlled trajectory  $(x, u)$  defined over the interval  $[0, T]$  and corresponding adjoint vector  $\lambda : [0, T] \rightarrow (\mathbb{R}^n)^*$  that is singular over an open interval  $I$ . The singular control is said to be of *intrinsic order*  $k$  over  $I$  if the following conditions are satisfied: (1) the first  $2k - 1$  derivatives of the switching function do not depend on the control  $u$  and vanish identically, i.e., for  $i = 1, \dots, 2k - 1$  we have that

$$\Phi^{(i)}(t) = \langle \lambda(t), \text{ad}_f^i(g)(x(t)) \rangle \equiv 0,$$

and (2)  $\langle \lambda(t), \text{ad}_f^{2k}(g)(x(t)) \rangle$  does not vanish on  $I$ .

**Theorem 2.8.4 (Generalized Legendre–Clebsch condition).** *Suppose the controlled trajectory  $(x_*, u_*)$  defined over the interval  $[0, T]$  minimizes the time of transfer from  $p \in \Omega$  to  $q \in \Omega$  for problem [NTOC] and the control  $u_*$  is singular of intrinsic order  $k$  over an open interval  $I \subset [0, T]$ . Then there exists an extremal lift  $\Gamma = ((x_*, u_*), \lambda)$  with the property that*

$$\begin{aligned} & (-1)^k \frac{\partial}{\partial u} \frac{d^{2k}}{dt^{2k}} \frac{\partial H}{\partial u} (\lambda_0, \lambda(t), x_*(t), u_*(t)) \\ &= (-1)^k \left\langle \lambda(t), [g, \text{ad}_f^{2k-1}(g)](x(t)) \right\rangle \geq 0 \quad \text{for all } t \in I. \end{aligned}$$

This result is also known as the Kelley condition [131, 132, 262]. For a singular extremal of intrinsic order 2, it states that

$$\left\langle \lambda(t), [g, \text{ad}_f^3(g)](x(t)) \right\rangle \geq 0 \quad \text{for all } t \in I, \quad (2.42)$$

and a proof of this condition will be given in Sect. 4.6.2. The strengthened version of this condition has very interesting consequences.

**Proposition 2.8.5.** *Let  $\Gamma = ((x, u), \lambda)$  be an extremal lift for the problem [NTOC] defined over the interval  $[0, T]$  that is singular of intrinsic order 2 over an open interval  $I$  for which*

$$\left\langle \lambda(t), [g, \text{ad}_f^3(g)](x(t)) \right\rangle > 0 \quad \text{for all } t \in I.$$

*Suppose that the singular control  $u$  takes values in the interior of the control set over the interval  $I$ . Then at no time  $\tau \in I$  can the control  $u$  be concatenated with a bang control  $u = -1$  or  $u = +1$ : concatenations of the types BS and SB violate the conditions of the maximum principle and are not optimal.*

*Proof.* Without loss of generality, we consider a concatenation of the type SX at time  $\tau \in I$ . That is, we assume that for some  $\varepsilon > 0$  the control is singular over the interval  $(\tau - \varepsilon, \tau)$  and is given by  $u = -1$  over the interval  $(\tau, \tau + \varepsilon)$ . Since the singular control is of order 2, the first three derivatives of the switching function do not depend on the control and thus are all continuous and given by  $\Phi^{(i)}(t) = \left\langle \lambda(t), \text{ad}_f^i(g)(x(t)) \right\rangle$ ,  $i = 1, 2, 3$ . The fourth derivative of  $\Phi$  at  $\tau$  from the right is thus given by

$$\begin{aligned} \Phi^{(4)}(\tau) &= \left\langle \lambda(\tau), \text{ad}_f^4(g)(x(\tau)) \right\rangle - \left\langle \lambda(\tau), [g, \text{ad}_f^3(g)](x(\tau)) \right\rangle \\ &< \left\langle \lambda(\tau), [f + u(\tau)g, \text{ad}_f^3(g)](x(\tau)) \right\rangle = 0, \end{aligned}$$

since the singular control  $u(\tau)$  takes a value in  $(-1, 1)$ . Thus the switching function has a local maximum for  $t = \tau$  and is negative over the interval  $(\tau, \tau + \varepsilon)$ . But then the minimization property of the Hamiltonian implies that the control must be

$u = +1$ . The analogous contradiction arises for concatenations of the type  $SY$  or for the order  $BS$ .  $\square$

This result implies that an optimal singular arc of order 2 cannot be concatenated with a bang control. In fact, an optimal control needs to switch infinitely many times between the controls  $u = -1$  and  $u = +1$  on any interval  $(\tau, \tau + \varepsilon)$  if a singular junction occurs at time  $\tau$ . Corresponding trajectories are called *chattering arcs*. In Sect. 2.11 we shall give an example that shows that these can be optimal for the seemingly most innocent-looking system.

### 2.8.5 Multi-input Systems and the Goh Condition

We close this section with some comments about the multi-input case in which the dynamics takes the form

$$\dot{x} = f(x) + \sum_{i=1}^m g_i(x)u_i, \quad x \in \Omega, \quad u \in U. \quad (2.43)$$

As before, for simplicity, we assume that all vector fields are  $C^\infty$  on  $\Omega$ . Clearly, now geometric properties of the control set  $U \subset \mathbb{R}^m$  matter. If  $U$  is a compact polyhedron, then the Hamiltonian will be minimized at one of the vertices, and singular controls arise as the minimum is attained along one of the faces of the polyhedron. The situation that most closely resembles the structures for the single-input case above, and probably is the practically most important one, occurs when the control set is a multi-dimensional rectangle,

$$U = [\alpha_1, \beta_1] \times \cdots \times [\alpha_m, \beta_m].$$

In this case, the minimization of the Hamiltonian function

$$H = \lambda_0 + \left\langle \lambda, f(x) + \sum_{i=1}^m g_i(x)u_i \right\rangle = \lambda_0 + \langle \lambda, f(x) \rangle + \sum_{i=1}^m \Phi_i u_i$$

still splits into  $m$  scalar minimization problems as in the single-input case, and optimal controls satisfy

$$u_i(t) = \begin{cases} \alpha_i & \text{if } \Phi_i(t) = \langle \lambda(t), g_i(x(t)) \rangle < 0, \\ \beta_i & \text{if } \Phi_i(t) = \langle \lambda(t), g_i(x(t)) \rangle > 0. \end{cases}$$

As before, now the switching functions  $\Phi_i$ ,  $i = 1, \dots, m$ , need to be analyzed to determine the optimal controls, and in principle, this follows the pattern dis-

cussed above. For example, Theorem 2.8.2 applies to give the derivatives of the switching functions as

$$\begin{aligned}\dot{\Phi}_i(t) &= \left\langle \lambda(t), \left[ f + \sum_{j=1}^m g_j u_j, g_i \right] (x(t)) \right\rangle \\ &= \langle \lambda(t), [f, g_i](x(t)) \rangle + \sum_{j \neq i} u_j(t) \langle \lambda(t), [g_j, g_i](x(t)) \rangle.\end{aligned}$$

In contrast to the single-input case, now the derivative  $\dot{\Phi}_i$  depends on the controls; on the controls other than  $u_i$ , that is. Hence, whether higher derivatives can be computed depends on the type of the controls, since these now need to be differentiated in time. Clearly, this is no issue for those components that are bang controls, but it needs to be checked if some of the controls are singular. All this leads to a much more elaborate analysis, which is best left for the particular problem under consideration. For example, if only one of the components is singular, with all other controls held constant, all the necessary conditions for optimality for the single-input control system are applicable. If more than one component is singular at the same time, the following result, the so-called *Goh condition*, [107] provides an extra necessary condition for optimality. This condition will be derived in Sect. 4.6.3.

**Theorem 2.8.5 (Goh condition).** [107] *Suppose the controlled trajectory  $(x_*, u_*)$  defined over the interval  $[0, T]$  minimizes the time of transfer from  $p \in \Omega$  to  $q \in \Omega$  for the multi-input control system with dynamics given by Eq. (2.43) and a rectangular control set  $U$ . Suppose the  $i$ th and  $j$ th controls are simultaneously singular over an open interval  $I \subset [0, T]$ . Then there exists an extremal lift  $\Gamma = ((x_*, u_*), \lambda)$  with the property that*

$$\langle \lambda(t), [g_i, g_j](x(t)) \rangle \equiv 0 \quad \text{for all } t \in I.$$

## 2.9 Time-Optimal Control for Nonlinear Systems in the Plane

We use the time-optimal control problem [NTOC] in the plane as an instrument to provide a first illustration of the use of geometric methods and the Lie-derivative-based techniques introduced above in the analysis of optimal control problems. These results are due to H. Sussmann, who in a series of papers [230, 236–238], gave a complete solution for this optimal control problem in dimension 2. The two-dimensional problem allows for easy visualization of the results, yet the general problem quickly gets very difficult, both in dimension 2 and even more so in higher dimensions. While Sussmann's results and the monograph by Boschain and Piccoli [51] provide a comprehensive analysis of the time-optimal control problem for two-dimensional systems, only partial results about the structure of time-optimal controls in higher dimensions (mostly in  $\mathbb{R}^3$  [54, 141, 210, 211] and some in  $\mathbb{R}^4$  [221]) are currently known.

[TOC in  $\mathbb{R}^2$ ] Let  $\Omega$  be an open and simply connected subset of  $\mathbb{R}^2$  and let  $f : \Omega \rightarrow \mathbb{R}^2$  and  $g : \Omega \rightarrow \mathbb{R}^2$  be two  $C^\infty$  vector fields defined on  $\Omega$ . For the control-affine system  $\Sigma$  with dynamics given by

$$\dot{x} = f(x) + g(x)u,$$

among all piecewise continuous (more generally, Lebesgue measurable) controls  $u, u : [0, T] \rightarrow [-1, 1]$ , find one that steers a given initial point  $q_1 \in \Omega$  into a target point  $q_2 \in \Omega$  (while remaining in  $\Omega$ ) in minimum time.

Our aim is to *determine the concatenation structure of optimal controls whose trajectories entirely lie in  $\Omega$* . More precisely, we are asking the question what can be said about time-optimal controlled trajectories that lie in  $\Omega$  if certain assumptions are made on the vector fields  $f$  and  $g$  at some reference point  $p \in \Omega$ . We consider  $\Omega$  to be a sufficiently small neighborhood of the point  $p$ , and thus by continuity, any inequality-type condition imposed on the values of  $f$  and  $g$  and/or their Lie brackets at  $p$  can also be assumed to hold in  $\Omega$ . But we shall develop the arguments as much as possible semiglobally, i.e., state them in a way that they are valid for sets  $\Omega$  that satisfy the required conditions throughout. It is natural to tackle this question by proceeding from the most general to increasingly more and more degenerate situations. That is, we first assume that the vectors  $f(p)$  and  $g(p)$  and other relevant Lie brackets are in general position, i.e., are linearly independent, and then proceed to consider more degenerate cases in which dependencies are allowed. In this spirit, throughout this section we make the following assumption:

(A0) The vector fields  $f$  and  $g$  are linearly independent everywhere on  $\Omega \subset \mathbb{R}^2$ .

Under this assumption, in this and the next section, we fully determine the structure of time-optimal controlled trajectories that lie in  $\Omega$  under generic conditions. These results are due to H. Sussmann, and in our presentation we follow his arguments that beautifully illustrate the use of geometric techniques in optimal control theory. In particular, as we proceed, it will become clear how these methods are needed to complement the first-order conditions of the Pontryagin maximum principle in order to arrive at deep and sharp results such as Proposition 2.9.5. Some of these results that we shall develop go well beyond the conditions of the maximum principle.

### 2.9.1 Optimal Bang-Bang Controls in the Simple Subcases

**Lemma 2.9.1.** *Any control corresponding to an abnormal extremal whose trajectory lies in  $\Omega$  is constant equal to  $u \equiv +1$  or  $u \equiv -1$ .*

*Proof.* Let  $\Gamma = ((x, u), \lambda)$  be an extremal and suppose  $\lambda_0 = 0$ . If the switching function  $\Phi(t) = \langle \lambda(t), g(x(t)) \rangle$  vanishes at some time  $\tau$ , then it follows from



$$H(t) = \langle \lambda(t), f(x(t)) \rangle + u(t) \langle \lambda(t), g(x(t)) \rangle \equiv -\lambda_0$$

that we also must have  $\langle \lambda(\tau), f(x(\tau)) \rangle = 0$ . Hence  $\lambda(\tau)$  vanishes against both  $f(x(\tau))$  and  $g(x(\tau))$ . Since these two vectors are linearly independent, it follows that  $\lambda(\tau) = 0$ . But this contradicts the nontriviality of the multipliers. Hence there cannot be any zeros for the switching function, and thus the corresponding controls must be constant.  $\square$

Having taken care of this special case, we henceforth assume that all extremals are normal and set  $\lambda_0 = 1$ . In particular, whenever  $\tau$  is a switching time, it follows that

$$\langle \lambda(\tau), f(x(\tau)) \rangle = -1. \quad (2.44)$$

Using  $f$  and  $g$  as a basis, we can express any higher-order bracket of  $f$  and  $g$  as a linear combination of this basis. In particular, there exist smooth functions  $\alpha$  and  $\beta$ ,  $\alpha, \beta \in C^\infty(\Omega)$ , such that for all  $x \in \Omega$  we have that

$$[f, g](x) = \alpha(x)f(x) + \beta(x)g(x). \quad (2.45)$$

We say an optimal controlled trajectory is of type  $XY$  if the corresponding control is bang-bang with at most one switching from  $u = -1$  to  $u = +1$  and use analogous labels for controlled trajectories that are concatenations of more segments. For example, a controlled trajectory of type  $XYSY$  is a concatenation of an  $X$ -trajectory followed by a  $Y$ -trajectory, a singular arc, and one more  $Y$ -trajectory. However, we always allow for the possibility that some of the segments are absent and thus a specific trajectory of type  $XYSY$  may simply be a concatenation of an  $X$ -trajectory with a single  $Y$ -trajectory.

**Proposition 2.9.1.** *If  $\alpha$  does not vanish on  $\Omega$ , then optimal controlled trajectories that lie in  $\Omega$  are of type  $XY$  if  $\alpha$  is positive and of type  $YX$  if  $\alpha$  is negative. Corresponding optimal controls are bang-bang with at most one switching.*

*Proof.* Recall that as always,  $X = f - g$  and  $Y = f + g$ . Let  $(x, u)$  be an optimal controlled trajectory that transfers a point  $q_1 \in \Omega$  into the point  $q_2 \in \Omega$  in minimum time with the trajectory  $x$  lying in  $\Omega$  and let  $\lambda$  be an adjoint vector such that the conditions of the maximum principle are satisfied. If  $\tau$  is a zero of the corresponding switching function, then we have that

$$\begin{aligned} \dot{\Phi}(\tau) &= \langle \lambda(\tau), [f, g](x(\tau)) \rangle \\ &= \alpha(x(\tau)) \cdot \langle \lambda(\tau), f(x(\tau)) \rangle + \beta(x(\tau)) \cdot \langle \lambda(\tau), g(x(\tau)) \rangle \\ &= -\alpha(x(\tau)). \end{aligned}$$

Since  $\alpha$  has constant sign on  $\Omega$ , it follows that at every zero of  $\Phi$ , the derivative of the switching function  $\Phi$  is nonzero and has the same sign. But then  $\Phi$  can have at most one zero, changing from positive to negative if  $\alpha > 0$  and from negative to positive if  $\alpha < 0$ . Thus optimal controls must switch from  $u = -1$  to  $u = +1$  if  $\alpha > 0$  and from  $u = +1$  to  $u = -1$  if  $\alpha < 0$ . This proves the result.  $\square$

For **example**, for the harmonic oscillator of Sect. 2.6, we have that

$$[f, g](x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} = -\frac{1}{x_2} \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} - \frac{x_1}{x_2} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \alpha(x)Ax + \beta(x)g.$$

We can take as  $\Omega$  either the upper or lower half-plane,  $\Omega_+ = \{(x_1, x_2) : x_2 > 0\}$  or  $\Omega_- = \{(x_1, x_2) : x_2 < 0\}$ , and it follows that optimal trajectories that entirely lie in  $\Omega_+$  or  $\Omega_-$  are bang-bang with at most one switching and that the switchings are from  $u = +1$  to  $u = -1$  in  $\Omega_+$  and from  $u = -1$  to  $u = +1$  in  $\Omega_-$ . Also, note that the controls corresponding to the abnormal trajectories  $\gamma_+$  and  $\gamma_-$  that lie in  $\Omega_+$  and  $\Omega_-$  are constant. As this example shows, there clearly can be more switchings, but the trajectories need to leave and reenter the region  $\Omega$  for this to be possible.

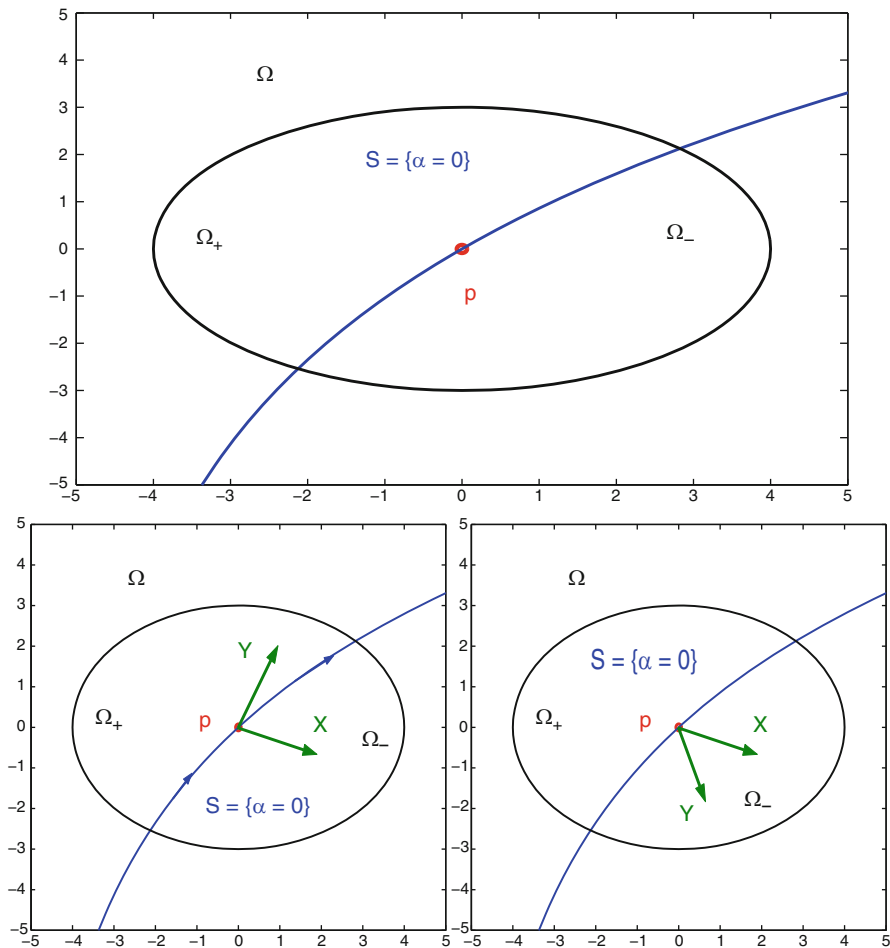
This proposition settles the local structure of time-optimal controlled trajectories near all points  $p$  where  $\alpha$  does not vanish, i.e., where  $g$  and the Lie-bracket  $[f, g]$  are in general position as well. Clearly, this does not suffice to settle the structure of optimal controls since there may and generally will exist some points where the vector fields  $g$  and  $[f, g]$  are linearly dependent and the local structure near these points will need to be determined too. Proceeding from the general case to the more special ones, but still maintaining condition (A0), we now assume that  $\alpha(p) = 0$ . At the same time, however, we want for this to occur in as nondegenerate a scenario as possible. That is, no other equality relations that would matter should hold at  $p$ . In terms of *singularity theory*, after determining the structure of optimal controlled trajectories near points of *codimension* 0 (only two inequality relations are imposed, one in the form of assumption (A0), the other as  $\alpha(p) \neq 0$ , but no equality relations hold at the reference point), we now proceed to the analysis of the *codimension* 1 scenario when we allow for exactly one equality constraint, but otherwise again only impose inequality relations. More specifically we assume that

- (A1) The vector fields  $f$  and  $g$  are linearly independent everywhere on  $\Omega \subset \mathbb{R}^2$  and there exists a point  $p \in \Omega$  with  $\alpha(p) = 0$ , but the Lie derivatives of  $\alpha$  along  $X = f - g$  and  $Y = f + g$  do not vanish on  $\Omega$ ,

$$(X\alpha)(x) = L_X(\alpha)(x) \neq 0, \quad (Y\alpha)(x) = L_Y(\alpha)(x) \neq 0 \quad \text{for all } x \in \Omega.$$

Furthermore, we assume that the zero set  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$  is a curve (embedded one-dimensional submanifold) in  $\Omega$  that divides  $\Omega$  into two connected subregions  $\Omega_+ = \{x \in \Omega : \alpha(x) > 0\}$  and  $\Omega_- = \{x \in \Omega : \alpha(x) < 0\}$  so that  $\Omega = \Omega_- \cup \mathcal{S} \cup \Omega_+$ .

With the understanding that  $\Omega$  is a sufficiently small neighborhood of  $p$ , this geometric assumption on the structure of the zero set of  $\alpha$  simply follows from the implicit function theorem, since the assumption on the Lie derivatives implies that the gradient of  $\alpha$  is non zero at  $p$ . On the other hand, several of the results below are valid as long as  $\Omega$  has this geometric separation property, not just in small neighborhoods, and therefore we prefer to state the results as such.



**Fig. 2.9** Assumption (A1): subcases with (left) and without (right) a singular arc

Assumption (A1) by itself does not lead to a unique structure of time-optimal trajectories, but several subcases exist, since it matters to which side of  $\mathcal{S}$  the vector fields  $X$  and  $Y$  point (see Fig. 2.9).

**Proposition 2.9.2.** *Assuming condition (A1), if  $L_X(\alpha) = X\alpha$  and  $L_Y(\alpha) = Y\alpha$  have the same sign on  $\Omega$ , then optimal controlled trajectories that lie in  $\Omega$  are of type  $YXY$  if the Lie derivatives are positive and of type  $XYX$  if they are negative. Optimal controls are bang-bang with at most two switchings.*

*Proof.* As above, let  $(x, u)$  be an optimal controlled trajectory that transfers a point  $q_1 \in \Omega$  into the point  $q_2 \in \Omega$  in minimum time with the trajectory  $x$  lying in  $\Omega$  and let  $\lambda$  be an adjoint vector such that the conditions of the maximum principle are

satisfied. Note that under these assumptions, the directional derivative of  $\alpha$  along the trajectory  $x$  is strictly increasing or decreasing. For at any point  $x(t)$ , the dynamics  $f(x(t)) + u(t)g(x(t))$  is a convex combination of the vectors  $X(x(t))$  and  $Y(x(t))$ ,

$$\begin{aligned} f(x(t)) + u(t)g(x(t)) &= \frac{1}{2}(X + Y)(x(t)) + u(t)\frac{1}{2}(Y - X)(x(t)) \\ &= \frac{1}{2}(1 - u(t))X(x(t)) + \frac{1}{2}(1 + u(t))Y(x(t)), \end{aligned}$$

and thus

$$(L_{f+ug}\alpha)(x(t)) = \frac{1}{2}(1 - u(t))L_X\alpha(x(t)) + \frac{1}{2}(1 + u(t))L_Y\alpha(x(t)). \quad (2.46)$$

Regardless of the control value  $u(t) \in [-1, 1]$ , this quantity is positive if  $L_X\alpha$  and  $L_Y\alpha$  are positive and negative if these quantities are negative. But then the trajectory  $x$  can cross the curve  $\mathcal{S}$  at most once. By Proposition 2.9.1, optimal controlled trajectories are of type  $YX$  in  $\Omega_-$  and of type  $XY$  in  $\Omega_+$ . Thus, if  $L_X\alpha$  and  $L_Y\alpha$  are positive, then trajectories move from the region  $\Omega_-$  into  $\Omega_+$ , and overall trajectories that lie in  $\Omega$  can at most be of type  $YXY$ . Similarly, if  $L_X\alpha$  and  $L_Y\alpha$  are negative, then trajectories move from  $\Omega_+$  into  $\Omega_-$ , and now trajectories that lie in  $\Omega$  can at most be of type  $XYX$ . In either case, optimal controls are bang-bang with at most two switchings.  $\square$

## 2.9.2 Fast and Slow Singular Arcs

If the vector fields  $X$  and  $Y$  point to opposite sides of the curve  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$ , then this curve is a singular arc.

**Proposition 2.9.3.** *Assuming condition (A1), if  $L_X(\alpha) = X\alpha$  and  $L_Y(\alpha) = Y\alpha$  have opposite signs on  $\Omega$ , then  $\mathcal{S}$  is a singular arc. If  $\Gamma = ((x, u), \lambda)$  is a corresponding singular extremal lift, then the strengthened Legendre–Clebsch condition is satisfied if  $L_X\alpha$  is negative, and it is violated if  $L_X\alpha$  is positive.*

*Proof.* In this case,  $X$  and  $Y$  always point to opposite sides of  $\mathcal{S}$ . Hence, at every point  $x \in \mathcal{S}$  there exists a convex combination  $u = u(x)$  such that the vector  $f(x) + u(x)g(x)$  is tangent to  $\mathcal{S}$  at  $x$ . This control is the unique solution to the equation  $L_{f+ug}\alpha = 0$ , i.e., solving from Eq. (2.46), we have that

$$u(x) = \frac{L_X\alpha(x) + L_Y\alpha(x)}{L_X\alpha(x) - L_Y\alpha(x)}.$$

Since  $L_X\alpha$  and  $L_Y\alpha$  have opposite signs, it follows that this value  $u(x)$  lies strictly between  $-1$  and  $+1$ , i.e., lies in the interior of the control set. In particular, it is admissible. Thus, if this control is optimal, then it must be singular.

We verify that the associated controlled trajectory through an initial condition  $q \in \mathcal{S}$  is extremal by constructing a singular extremal lift  $\Gamma = ((x, u), \lambda)$ . Let  $x = x(t)$  be the solution to the initial value problem

$$\dot{x} = f(x) + u(x)g(x), \quad x(0) = q.$$

This solution exists over a maximal interval  $(t_-, t_+)$  with  $t_- < 0 < t_+$ . Let  $\psi \in (\mathbb{R}^2)^*$  be a covector such that

$$\langle \psi, g(q) \rangle = 0 \quad \text{and} \quad \langle \psi, f(q) \rangle = -1$$

and let  $\lambda = \lambda(t)$  be the solution of the corresponding adjoint equation

$$\dot{\lambda} = -\lambda(Df(x(t)) + u(x(t))Dg(x(t)))$$

with initial condition  $\lambda(0) = \psi$ . This triple defines a singular extremal lift if the switching function  $\Phi(t) = \langle \lambda(t), g(x(t)) \rangle$  vanishes identically on  $(t_-, t_+)$ . But this is clear by construction: we have  $(L_{f+ug}\alpha)x(t) \equiv 0$ , and since  $\alpha(q) = 0$ , it follows that  $\alpha(x(t)) \equiv 0$  on  $(t_-, t_+)$ . Hence we get

$$\begin{aligned} \dot{\Phi}(t) &= \langle \lambda(t), [f, g](x(t)) \rangle \\ &= \alpha(x(t)) \langle \lambda(t), f(x(t)) \rangle + \beta(x(t)) \cdot \langle \lambda(t), g(x(t)) \rangle \\ &= \beta(x(t))\Phi(t). \end{aligned}$$

But  $\Phi(0) = \langle \psi, g(q) \rangle = 0$ , and so the switching function vanishes identically. Hence  $\Gamma$  is a normal singular extremal lift.

It remains to check the Legendre–Clebsch condition. Using Lemma 2.8.1, it follows that

$$[g, [f, g]] = [g, \alpha f + \beta g] = (L_g \alpha) f - \alpha[f, g] + (L_g \beta)g.$$

Along the singular extremal,  $\langle \lambda, g(x) \rangle \equiv 0$  and  $\langle \lambda, [f, g](x) \rangle \equiv 0$ , and thus, using Eq. (2.44), it also follows that  $\langle \lambda(t), f(x(t)) \rangle \equiv -1$ . Hence

$$\begin{aligned} \langle \lambda(t), [g, [f, g]](x(t)) \rangle &= -L_g \alpha(x(t)) \\ &= -L_{\frac{1}{2}(Y-X)} \alpha(x(t)) = \frac{1}{2} (L_X \alpha - L_Y \alpha)(x(t)). \end{aligned} \quad (2.47)$$

This quantity has the same sign as  $L_X \alpha$ , and the strengthened Legendre–Clebsch condition is satisfied if  $\langle \lambda(t), [g, [f, g]](x(t)) \rangle < 0$ . Hence the result follows.  $\square$

The Legendre–Clebsch condition distinguishes *fast* from *slow singular arcs*. On a set  $\Omega$  in the plane where  $f$  and  $g$  are linearly independent, this can be seen with an instructive geometric argument by introducing a 1-form that measures the time

along the trajectories. Differential forms provide a superior formalism for these computations, and for the sake of completeness, we provide the needed definitions and results. These are standard concepts from differential geometry and can be found in any text on the subject, such as, for example, [50, 256]. One-forms are simply linear functionals on the space of all tangent vectors; hence the space of 1-forms on  $\Omega$  is a two-dimensional vector space as well. If we write  $x = x_1 e_1 + x_2 e_2$ , where  $\{e_1, e_2\}$  is the canonical ordered basis for  $\mathbb{R}^2$ , we denote the corresponding dual basis by  $dx_1$  and  $dx_2$ ; that is,  $dx_i$  is the linear functional that satisfies

$$\langle dx_i, e_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Since  $f$  and  $g$  are linearly independent on  $\Omega$ , there exists a unique 1-form  $\omega$  on  $\Omega$  that satisfies

$$\langle \omega(x), f(x) \rangle \equiv 1 \quad \text{and} \quad \langle \omega(x), g(x) \rangle \equiv 0 \quad \text{for all } x \in \Omega. \quad (2.48)$$

This form  $\omega$  is easily computed: if  $f$  and  $g$  have the representations

$$f(x) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} \quad \text{and} \quad g(x) = \begin{pmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{pmatrix},$$

then

$$\omega(x) = \frac{g_2(x)dx_1 - g_1(x)dx_2}{f_1(x)g_2(x) - f_2(x)g_1(x)} = \frac{g_2(x)dx_1 - g_1(x)dx_2}{\det(f(x), g(x))}, \quad (2.49)$$

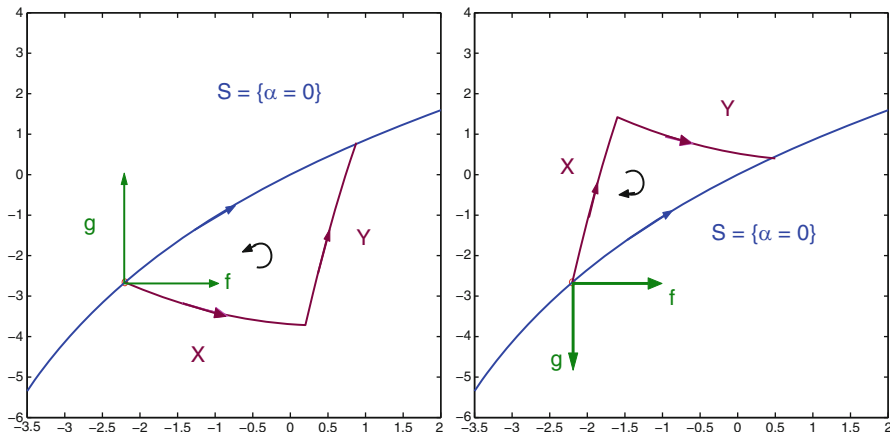
with  $\det(f(x), g(x))$  denoting the determinant of the matrix

$$\begin{pmatrix} f_1 & g_1 \\ f_2 & g_2 \end{pmatrix}.$$

This determinant does not vanish on  $\Omega$ , since the vector fields  $f$  and  $g$  are linearly independent. Depending on the sign of this determinant, the ordered basis  $\mathcal{B} = \{f, g\}$  is said to be positively, respectively negatively, oriented.

Let  $(x, u)$  be a controlled trajectory defined over an interval  $[t_0, t_1]$  with trajectory  $x$  lying in  $\Omega$ . Then the line integral of  $\omega$  along the curve  $x(\cdot)$  is given by

$$\begin{aligned} \int_{x(\cdot)} \omega &= \int_{t_0}^{t_1} \langle \omega(x(t), \dot{x}(t)) \rangle dt \\ &= \int_{t_0}^{t_1} \langle \omega(x(t), f(x(t))) \rangle dt + \int_{t_0}^{t_1} u(t) \langle \omega(x(t), g(x(t))) \rangle dt \\ &= \int_{t_0}^{t_1} dt = t_1 - t_0, \end{aligned}$$



**Fig. 2.10** Positively (*left*) and negatively (*right*) oriented vector fields  $f$  and  $g$

so that  $\omega$  measures the time along trajectories. For this reason,  $\omega$  sometimes is called the clock form [44].

We now show how this differential form  $\omega$  can be used to determine which type of trajectory is faster. Consider a point  $q_1 \in \mathcal{S}$  and let  $(x, u)$  be the controlled trajectory that steers  $q_1$  to another point  $q_2 \in \mathcal{S}$  along the singular arc  $\mathcal{S} \subset \Omega$  in time  $\tau$ . If  $\tau$  is small, then there exists a unique  $XY$ -trajectory that also steers  $q_1$  into  $q_2$  and lies in  $\Omega$ . Simply consider the forward orbit of the  $X$ -trajectory that starts at  $q_1$  and the backward orbit of the  $Y$ -trajectory that ends at  $q_2$ . Since  $X$  and  $Y$  point to opposite sides of  $\mathcal{S}$ , it follows that these two orbits intersect in some point  $r \in \Omega$ . Suppose it takes time  $s$  to go from  $q_1$  to  $r$  along  $X$  and time  $t$  to go from  $r$  to  $q_2$  along  $Y$ . If we denote the mapping that follows the flow of the vector field  $X$  for time  $s$  by  $\Psi_s^X$ , then we can write  $r = \Psi_s^X(q_1)$ , and analogously, for  $Y$ , we have that  $q_2 = \Psi_t^Y(r)$ . Overall, therefore,

$$q_2 = \Psi_t^Y(r) = \Psi_t^Y(\Psi_s^X(q_1)) = (\Psi_t^Y \circ \Psi_s^X)(q_1).$$

Stokes's theorem allows us to compare the time  $s + t$  along the  $XY$ -trajectory with the time  $\tau$  along the singular arc. Denote the closed curve consisting of the  $XY$ -trajectory concatenated with the singular arc run backward from  $q_2$  to  $q_1$  by  $\Delta$ . The orientation of this closed curve matters in Stokes's theorem, and  $\Delta$  has the same orientation as the ordered basis  $\mathcal{B} = \{f, g\}$ : since

$$\det(X(x), Y(x)) = \det(f(x) - g(x), f(x) + g(x)) = 2\det(f(x), g(x)),$$

the ordered basis  $\{X, Y\}$  has the same orientation as  $\{f, g\}$ . But if  $\{X, Y\}$  is positively oriented, the curve  $\Delta$  is traversed counterclockwise, while it is traversed clockwise if  $\{X, Y\}$  is oriented negatively (see Fig. 2.10).

Without loss of generality, we assume that the orientation of  $\Delta$  is positive. Then Stokes's theorem [50, 256] gives that

$$s + t - \tau = \int_{\Delta} \omega = \int_R d\omega$$

where  $R$  denotes the region enclosed by  $\Delta$ . For a 1-form  $\omega$  given as

$$\omega(x) = \sum_{i=1}^n \xi_i(x) dx_i$$

with  $\xi_i$  smooth functions, the 2-form  $d\omega$  is defined as

$$d\omega(x) = \sum_{i=1}^n d\xi_i(x) \wedge dx_i$$

with

$$d\xi_i(x) = \sum_{j=1}^n \frac{\partial \xi_i}{\partial x_j}(x) dx_j,$$

the differential of  $\xi_i$ . For the wedge product,  $\wedge$ , the rules of an alternating product apply, i.e.,  $dx_1 \wedge dx_2 = -dx_2 \wedge dx_1$  and  $dx_i \wedge dx_i = 0$ . In order to evaluate the area integral on the right, we need some facts about the actions of differential forms on vector fields. If  $\phi$  is a smooth function defined on some domain  $D \subset \mathbb{R}^n$ ,  $\phi \in C^\infty(D)$ , then the action of the 1-form  $d\phi$  on a smooth vector field is simply taking the Lie derivative of  $\phi$  along  $Z$ ,

$$\langle d\phi(x), Z(x) \rangle = L_Z \phi(x).$$

For writing out the inner product in terms of the basis vectors, we have that

$$\begin{aligned} \langle d\phi(x), Z(x) \rangle &= \left\langle \sum_{i=1}^n \frac{\partial \phi}{\partial x_i}(x) dx_i, \sum_{j=1}^n Z_j(x) e_j \right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \phi}{\partial x_i}(x) Z_j(x) \langle dx_i, e_j \rangle \\ &= \sum_{i=1}^n \frac{\partial \phi}{\partial x_i}(x) Z_i(x) = L_Z \phi(x). \end{aligned}$$

If  $\psi$  is another smooth function on  $D$ ,  $\psi \in C^\infty(D)$ , then the action of the 2-form  $d\phi \wedge d\psi$  on a pair of smooth vector fields  $f$  and  $g$  is defined as the alternating product

$$\begin{aligned} &\langle d\phi(x) \wedge d\psi(x), (f(x), g(x)) \rangle \\ &= \langle d\phi(x), f(x) \rangle \cdot \langle d\psi(x), g(x) \rangle - \langle d\phi(x), g(x) \rangle \cdot \langle d\psi(x), f(x) \rangle \\ &= L_f \phi(x) \cdot L_g \psi(x) - L_g \phi(x) \cdot L_f \psi(x). \end{aligned} \tag{2.50}$$



Note, in particular, that this gives 0 if  $f = g$ . These actions are then related to the Lie bracket through the following relation:

**Lemma 2.9.2.** [50] *Given any 1-form  $\omega$  and smooth vector fields  $f$  and  $g$  defined on  $D \subset \mathbb{R}^n$ , it follows that*

$$\langle d\omega, (f, g) \rangle = L_f \langle \omega, g \rangle - L_g \langle \omega, f \rangle - \langle \omega, [f, g] \rangle.$$

*Proof.* It suffices to prove the Lemma if  $\omega$  is of the form  $\omega = \phi d\psi$ , where  $\phi$  and  $\psi$  are smooth functions on  $D$ . In this case, and dropping the argument  $x$ , we have that

$$\begin{aligned} & L_f \langle \omega, g \rangle - L_g \langle \omega, f \rangle - \langle \omega, [f, g] \rangle \\ &= L_f \langle \phi d\psi, g \rangle - L_g \langle \phi d\psi, f \rangle - \langle \phi d\psi, [f, g] \rangle \\ &= L_f(\phi) \langle d\psi, g \rangle + \phi L_f \langle d\psi, g \rangle - L_g(\phi) \langle d\psi, f \rangle - \phi L_g \langle d\psi, f \rangle - \phi \langle d\psi, [f, g] \rangle \\ &= L_f(\phi) L_g(\psi) + \phi L_f(L_g(\psi)) - L_g(\phi) L_f(\psi) - \phi L_g(L_f(\psi)) - \phi L_{[f, g]} \psi \\ &= L_f(\phi) L_g(\psi) - L_g(\phi) L_f(\psi) + \phi \{ L_f(L_g(\psi)) - L_g(L_f(\psi)) - L_{[f, g]} \psi \}. \end{aligned}$$

But

$$L_{[f, g]} \psi = L_f(L_g(\psi)) - L_g(L_f(\psi))$$

and thus since  $d\omega = d\phi \wedge d\psi$ , the result follows from Eq. (2.50).  $\square$

For the 1-form  $\omega$  defined by Eq. (2.48), we have  $\langle \omega, f \rangle \equiv 1$  and  $\langle \omega, g \rangle \equiv 0$ , and thus the Lie derivatives of these functions vanish giving

$$\begin{aligned} \langle d\omega, (f, g) \rangle &= -\langle \omega, [f, g] \rangle = -\langle \omega, \alpha f + \beta g \rangle \\ &= -\alpha \langle \omega, f \rangle - \beta \langle \omega, g \rangle = -\alpha. \end{aligned}$$

Furthermore,

$$\begin{aligned} \langle d\omega, (f, g) \rangle &= \langle d\omega, (f_1 e_1 + f_2 e_2, g_1 e_1 + g_2 e_2) \rangle \\ &= f_1 \langle d\omega, (e_1, g_1 e_1 + g_2 e_2) \rangle + f_2 \langle d\omega, (e_2, g_1 e_1 + g_2 e_2) \rangle \\ &= f_1 g_1 \langle d\omega, (e_1, e_1) \rangle + f_1 g_2 \langle d\omega, (e_1, e_2) \rangle \\ &\quad + f_2 g_1 \langle d\omega, (e_2, e_1) \rangle + f_2 g_2 \langle d\omega, (e_2, e_2) \rangle \\ &= (f_1 g_2 - f_2 g_1) \langle d\omega, (e_1, e_2) \rangle \\ &= \det(f(x), g(x)) \langle d\omega, (e_1, e_2) \rangle, \end{aligned}$$

so that

$$\langle d\omega, (e_1, e_2) \rangle = -\frac{\alpha(x)}{\det(f(x), g(x))}.$$

Hence

$$\tau - (s+t) = - \int_R d\omega = \int_R \frac{\alpha(x)}{\det(f(x), g(x))} dx. \quad (2.51)$$

By construction, the region  $R$  lies entirely in  $\Omega_+$  or  $\Omega_-$ , namely in  $\Omega_+$  if the Lie derivative  $L_X \alpha$  is positive and in  $\Omega_-$  if it is negative. In the first case, the integral is positive (recall that we assume that the basis  $\mathcal{B} = \{f, g\}$  is positively oriented), and thus the singular arc takes longer than the  $XY$ -trajectory, while it does better in the second case when the region  $R$  lies in  $\Omega_-$ . These conclusions are consistent with the strengthened Legendre–Clebsch condition. In carrying out this argument for  $YX$ -trajectories, the same consistency shows. This explicitly verifies that the Legendre–Clebsch condition distinguishes fast from slow singular arcs.

This calculation can also be used to show that *increasing the number of switchings along a bang-bang trajectory speeds up the time of transfer if the strengthened Legendre–Clebsch condition is satisfied*. Again, consider the  $XY$ -trajectory that steers  $q_1$  into  $q_2$  and is part of the curve  $\Delta$  constructed above. Construct an  $XYXY$ -trajectory that connects  $q_1$  with  $q_2$  in  $\Omega$  as follows: (i) Starting from  $q_1$ , follow the  $X$ -trajectory for time  $s_1 < s$  and let  $r_1$  denote the point reached,  $r_1 = \Psi_{s_1}^X(q_1)$ . (ii) At  $r_1$ , change to the  $Y$ -trajectory and follow it for time  $t_1$  until the  $Y$ -trajectory again reaches the singular curve  $\mathcal{S}$  in some point  $r_2$ ,  $r_2 = \Psi_{t_1}^Y(r_1) \in \mathcal{S}$ . (iii) Here once more switch to the  $X$ -trajectory and follow it for time  $s_2$  until it intersects the original  $Y$ -trajectory in the point  $r_3$ ,  $r_3 = \Psi_{s_2}^X(r_2)$ . (iv) Then follow this  $Y$ -trajectory from  $r_3$  into  $q_2$ , say  $q_2 = \Psi_{t_2}^Y(r_3)$ . Thus, overall,

$$q_2 = (\Psi_{t_2}^Y \circ \Psi_{s_2}^X \circ \Psi_{t_1}^Y \circ \Psi_{s_1}^X)(q_1).$$

Denote by  $\diamond$  the diamond-shaped curve that is obtained by concatenating the  $X$ -trajectory from  $r_1$  to  $r$  first with the  $Y$ -trajectory from  $r$  to  $r_3$ , then with the  $X$ -trajectory run backward from  $r_3$  to  $r_2$ , and finally the  $Y$ -trajectory run backward from  $r_2$  to  $r_1$  (see Fig. 2.11). Since we assume that the basis  $\{f, g\}$  is positively oriented, the curve  $\diamond$  also is mathematically positively (counterclockwise) oriented. Let  $D$  denote the region enclosed by  $\diamond$ . Using the 1-form  $\omega$ , the difference in time between the original  $XY$ -trajectory and the newly constructed  $XYXY$ -trajectory can then be calculated as

$$\begin{aligned} (s+t) - (s_1+t_1+s_2+t_2) &= (s-s_1) + (t-t_2) - s_2 - t_1 \\ &= - \int_{\diamond} \omega = - \int_D d\omega = - \int_D \frac{\alpha(x)}{\det(f(x), g(x))} dx. \end{aligned}$$

By construction of  $\diamond$ , the region  $D$  lies entirely in  $\Omega_+$  if  $L_X \alpha > 0$  and in  $\Omega_-$  if  $L_X \alpha < 0$ . Hence, the  $XYXY$ -trajectory steers  $q_1$  into  $q_2$  faster than the  $XY$ -trajectory does if  $L_X \alpha < 0$ , and it is slower if  $L_X \alpha > 0$ . Thus, *if the strengthened Legendre–Clebsch condition is satisfied, i.e., for  $L_X \alpha < 0$ , bang-bang trajectories with more switchings near the singular arc are faster, while they are slower if the strengthened*

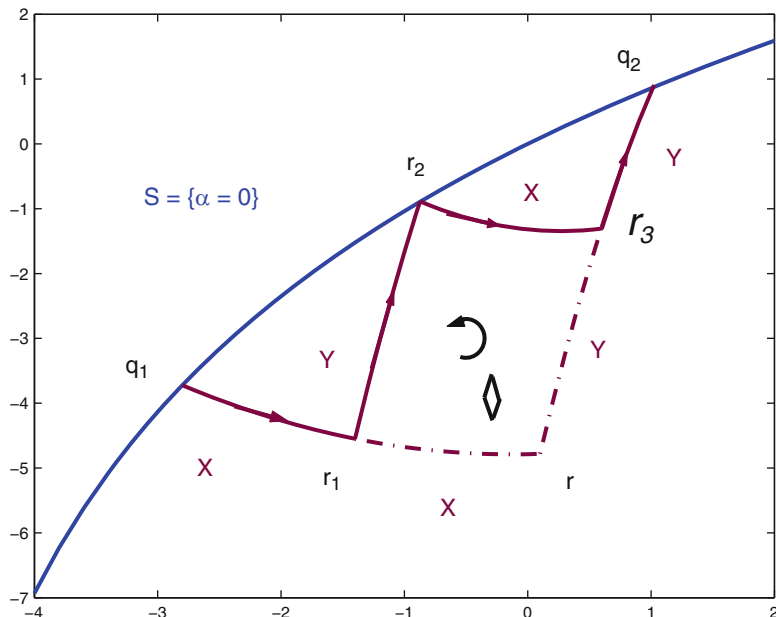


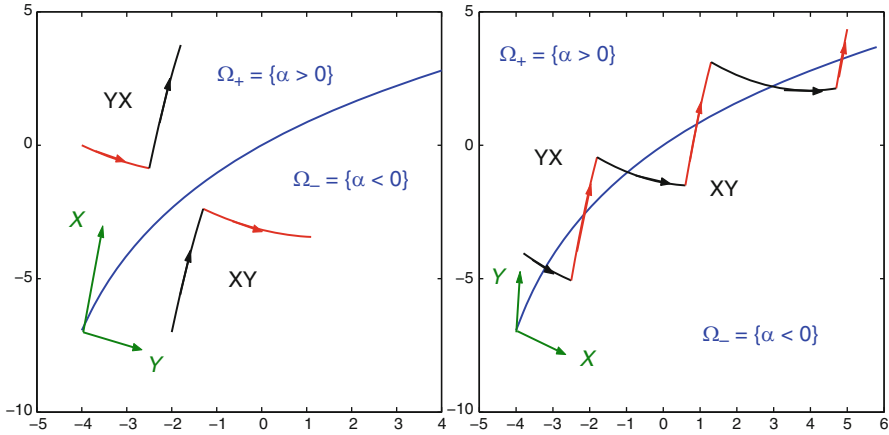
Fig. 2.11 Comparison of an XYXY-trajectory with an XY-trajectory

*Legendre–Clebsch condition is violated.* In either case, the singular arc can closely be approximated by bang-bang trajectories with an increasing number of switchings, and it is therefore to be expected that in the limit, optimal controls will follow the singular arc if the strengthened Legendre–Clebsch condition is satisfied, while they will avoid it, i.e., have as few switchings as possible, if it is violated. This indeed is the case.

**Proposition 2.9.4.** *Assuming condition (A1), if  $L_X(\alpha) = X\alpha$  is negative and  $L_Y(\alpha) = Y\alpha$  is positive on  $\Omega$ , then optimal controlled trajectories that lie in  $\Omega$  are of the type BSB, that is, are at most concatenations of a bang arc (X or Y) followed by a singular arc and possibly one more bang arc.*

*Proof.* Let  $(x, u)$  be an optimal controlled trajectory that transfers a point  $q_1 \in \Omega$  into the point  $q_2 \in \Omega$  in minimum time with the trajectory  $x$  lying in  $\Omega$  and let  $\lambda$  be an adjoint vector such that the conditions of the maximum principle are satisfied. Once more, recall that by Proposition 2.9.1, optimal controlled trajectories are at most of type YX in  $\Omega_-$  and of type XY in  $\Omega_+$ .

Suppose  $q_1 \in \Omega_-$ . Initially, since  $q_1 \notin \mathcal{S}$ , the optimal control can be only  $u = -1$  or  $u = +1$ . If the control starts with  $u = -1$ , then since  $L_X(\alpha) < 0$ , the trajectory moves away from  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$ , and no junctions from X to Y are possible in  $\Omega_-$ . Hence this trajectory simply is an X-arc, and the corresponding control is constant, given by  $u \equiv -1$ . On the other hand, if the control starts with  $u = +1$ , then the trajectory moves toward  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$ . In this case, it



**Fig. 2.12** Bang-bang switchings near fast (*left*) and slow (*right*) singular arcs

is possible that (a) the control switches to  $u = -1$  before or as the singular arc  $\mathcal{S}$  is reached, (b) the control switches to become singular as it reaches  $\mathcal{S}$ , or (c) this  $Y$ -trajectory simply crosses the singular arc. In case (a), after the switching time, the  $X$ -trajectory again moves the state away from  $\mathcal{S}$  and no further switchings to  $Y$  are allowed in  $\Omega_-$ . Hence in this case the trajectory is of type  $YX$ . In case (c), once the  $Y$ -trajectory enters the region  $\Omega_+$ , switchings to  $X$  are no longer allowed and thus this trajectory simply is a  $Y$ -arc with constant control  $u = +1$ . The interesting case is (b). It follows from Proposition 2.8.4 that switchings onto and off the singular arc for some time, the trajectory can leave  $\mathcal{S}$  with the bang control  $u = -1$  or  $u = +1$ . Using an  $X$ -trajectory, the system enters the region  $\Omega_-$ , while it enters  $\Omega_+$  along a  $Y$ -trajectory. In any case, no more switchings are possible in  $\Omega$  by Proposition 2.9.1. Thus overall, the structure is at most of type *BSB*. The analogous reasoning for an initial condition  $q_1 \in \Omega_+$  shows the same concatenation structure to be valid.  $\square$

### 2.9.3 Optimal Bang-Bang Trajectories near a Slow Singular Arc

What makes Proposition 2.9.4 work is that optimal bang-bang switchings in the regions  $\Omega_-$  and  $\Omega_+$  move the system *away* from the singular arc  $\mathcal{S}$  if  $L_X(\alpha) = X\alpha$  is negative and  $L_Y(\alpha) = Y\alpha$  is positive on  $\Omega$ . The resulting synthesis of the type *BSB* is quite common around optimal singular arcs in small dimensions and will still be encountered several times throughout this text (e.g., Sects. 6.2 and 7.3). If, however,  $L_X(\alpha) = X\alpha$  is positive and  $L_Y(\alpha) = Y\alpha$  is negative on  $\Omega$ , and in this case the singular arc is not optimal by the Legendre–Clebsch condition, or “slow,” the opposite is true. Now optimal bang-bang junctions steer trajectories *toward* the singular arc  $\mathcal{S}$  (see Fig. 2.12). In fact, in this case, there exist bang-bang extremals

(i.e., bang-bang trajectories that satisfy the necessary conditions for optimality of the maximum principle) whose trajectories lie in  $\Omega$  and have an arbitrarily large number of switchings. But as the geometric argument carried out above indicates, in this case, making more switchings slows down the trajectories, and thus none of these are optimal. This reasoning, however, is quite more intricate and goes well beyond a direct application of the conditions of the maximum principle, but involves the generalization of the concept of an *envelope* from the calculus of variations to the optimal control problem. We shall more generally develop this theory in Sect. 5.4, but here we include a self-contained proof of the result below due to Sussmann.

**Proposition 2.9.5.** [230, 236] *Let  $\Omega$  be a domain on which condition (A1) is satisfied and where  $L_X(\alpha) = X\alpha$  is positive and  $L_Y(\alpha) = Y\alpha$  is negative. If  $\Omega$  is taken sufficiently small, then optimal controls for trajectories that lie in  $\Omega$  are bang-bang with at most one switching.*

Note that in contrast to the previous results, here we need to include the requirement that  $\Omega$  be a small enough neighborhood of the reference point. This result does not hold in the more semiglobal setting without additional assumptions. The essential new concept involved in the proof of this result involves what are called conjugate points in [230]. However, for reasons that will be explained below, we prefer to use the terminology of *g-dependent points* instead.

**Definition 2.9.1 (Variational vector field).** Let  $(x, u) : [0, T] \rightarrow \Omega \times U$  be an extremal controlled trajectory with multiplier  $\lambda$ . A variational vector field  $w$  along  $\Gamma = ((x, u), \lambda)$  is a solution  $w : [0, T] \rightarrow \mathbb{R}^2$  of the corresponding variational equation

$$\dot{w}(t) = \{Df(x(t)) + u(t)Dg(x(t))\} \cdot w(t). \quad (2.52)$$

The adjoint equation for the multiplier  $\lambda$  actually is the “adjoint” in the sense of linear differential equations to this variational equation (2.52). Thus, for any variational vector field  $w$  along  $\Gamma$ , the function  $h : [0, T] \rightarrow \mathbb{R}, t \mapsto h(t) = \langle \lambda(t), w(t) \rangle$  is constant:

$$\dot{h}(t) = \langle \dot{\lambda}(t), w(t) \rangle + \langle \lambda(t), \dot{w}(t) \rangle = 0.$$

Suppose now that the switching function  $\Phi(t) = \langle \lambda(t), g(x(t)) \rangle$  vanishes at times  $t_1 < t_2$  and let  $w$  be the variational vector field that satisfies  $w(t_1) = g(x(t_1))$ . Since  $\Phi(t_1) = 0$ , it then follows that  $\langle \lambda(t_2), w(t_2) \rangle = 0$ . But  $\Phi(t_2) = \langle \lambda(t_2), g(x(t_2)) \rangle = 0$  as well, and since  $\lambda(t_2) \neq 0$ , the vectors  $g(x(t_2))$  and  $w(t_2)$  must be linearly dependent. This leads to the following definition of *g-dependent points* in the plane.

**Definition 2.9.2 (g-dependent).** Let  $(x, u) : [0, T] \rightarrow \Omega \times U$  be an extremal controlled trajectory with multiplier  $\lambda$ . Given times  $t_1$  and  $t_2$ ,  $0 \leq t_1 < t_2 \leq T$ , let  $w(\cdot)$  be the variational vector field that satisfies  $w(t_1) = g(x(t_1))$ . We call the points  $x(t_1)$  and  $x(t_2)$  *g-dependent* (along  $\Gamma = ((x, u), \lambda)$ ) if the vectors  $g(x(t_2))$  and  $w(t_2)$  are linearly dependent.

Thus, if  $\Gamma = ((x, u), \lambda)$  is an extremal lift for which the control  $u$  switches at times  $t_1 < t_2$ , then the switching points  $x(t_1)$  and  $x(t_2)$  are *g-dependent*. As the example

of time-optimal control for the harmonic oscillator shows, optimality of trajectories need not cease at  $g$ -dependent points. It does in the case that will be considered here, and thus the terminology of conjugate points is used in [230]. However, we generally prefer to restrict the terminology “conjugate point” to the case when optimality of trajectories ceases. We shall elaborate more on this in Sect. 6.1.

The key to the proof of Proposition 2.9.5 is to establish an inversion of  $g$ -dependent points around  $\mathcal{S}$ . For this calculation, a good choice of coordinates around  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$  is beneficial. The type of coordinates used here will also be needed in Sect. 2.10 and we therefore consider a slightly weaker version of assumption (A1). Let  $p \in \Omega$  be a point at which (i) the vector fields  $f$  and  $g$  (and thus also  $X$  and  $Y$ ) are linearly independent; (ii)  $\alpha(p) = 0$ , but the Lie derivative of  $\alpha$  along  $X$  does not vanish,  $L_X \alpha(p) \neq 0$ ; and (iii) the Lie derivative of  $\alpha$  along  $g$  does not vanish,  $L_g \alpha(p) \neq 0$ . Conditions (i) and (ii) imply that the geometric properties of  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$  required in assumption (A1) are satisfied on a sufficiently small neighborhood of  $p$ . The third condition ensures that the vector field

$$S(x) = f(x) + \frac{L_f \alpha(x)}{L_g \alpha(x)} g(x) = f(x) + \frac{L_X \alpha(x) + L_Y \alpha(x)}{L_X \alpha(x) - L_Y \alpha(x)} g(x)$$

is well-defined near  $p$ . If the quotient  $\frac{L_f \alpha(x)}{L_g \alpha(x)}$  lies between  $-1$  and  $+1$ , then this is the singular vector field. But for the current reasoning it is not necessary that  $S$  correspond to a trajectory of the system, only that the integral curve of  $S$  through  $p$  be the curve  $\mathcal{S}$ . (This was shown in the proof of Proposition 2.9.3.) Let  $[a, b]$  be an interval that contains 0 in its interior on which the solution to the initial value problem  $\dot{y} = S(y)$ ,  $y(0) = p$ , exists. It then follows from a standard compactness argument that there exists an  $\varepsilon > 0$  such that the solution  $z = z(\cdot; s)$  to the initial value problem  $\dot{z} = X(z)$ ,  $z(0) = y(s)$ , exists on the interval  $[-\varepsilon, \varepsilon]$ . Using the notation  $\Psi$  for the flow, we denote this solution by

$$\psi(s, t) = \Psi_t^X \circ \Psi_s^S(p).$$

If, in addition, the Lie derivative  $L_X(\alpha)$  does not vanish at  $y(t)$  for all  $t \in [a, b]$ , then the  $X$ -flow is everywhere transversal to the curve  $\mathcal{S}$  and the map  $\psi$  is a diffeomorphism from some square  $Q(\varepsilon) = (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon)$  onto some neighborhood  $\psi(Q)$  of  $p$ . If we now choose this set  $\psi(Q)$  as  $\Omega$ ,

$$\Omega = \{\psi(s, t) : -\varepsilon < s < \varepsilon, -\varepsilon < t < \varepsilon\},$$

then the times  $(s, t) \in Q$  provide us with a good set of coordinates on  $\Omega$  called *canonical coordinates of the second kind* in Lie theory (also, see Sects. 4.5 and 7.1). In these coordinates, the curve  $\mathcal{S}$  corresponds to the  $s$ -axis,  $\mathcal{S} \cong \{(s, t) \in Q : t = 0\}$ , and integral curves of the vector field  $X$  are the vertical lines  $s = \text{const}$ . We call such a mapping  $\psi : Q \rightarrow \Omega$  an  $X$ -aligned chart of coordinates centered at the point  $p$  (see Fig. 2.13).

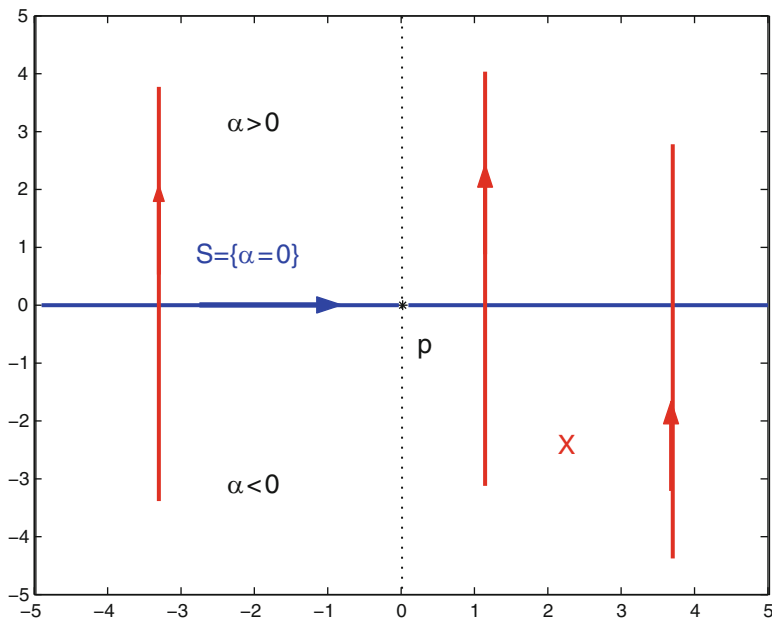


Fig. 2.13 An  $X$ -aligned coordinate chart

**Definition 2.9.3 ( $X$ -aligned chart of coordinates).** An  $X$ -aligned chart of coordinates (centered at  $p$ ) is a diffeomorphism  $\psi$ ,

$$\psi : Q(\varepsilon) \subset \mathbb{R}^2 \rightarrow \Omega, \quad (s, t) \mapsto \psi(s, t) = \Psi_t^X \circ \Psi_s^S(p),$$

such that  $X$  and  $Y$  are linearly independent everywhere on  $\Omega$ , the set  $\mathcal{S} = \{x \in \Omega : \alpha(x) = 0\}$  is the integral curve of the vector field  $S$  through  $p$ , and the Lie derivatives  $L_X \alpha$  and  $L_g \alpha$  are everywhere nonzero on  $\Omega$ .

**Lemma 2.9.3.** [230] *Given an  $X$ -aligned chart of coordinates,  $\Omega = \psi(Q(\varepsilon))$ , for  $\varepsilon$  small enough, there exists a differentiable function*

$$\zeta : Q \rightarrow \mathbb{R}, \quad (s, t) \mapsto \zeta(s, t),$$

that satisfies

$$\zeta(s, 0) = 0, \quad \frac{\partial \zeta}{\partial t}(s, 0) = -1,$$

and is such that two points  $q = \psi(s, t)$  and  $q' = \psi(s', t')$  in  $\Omega$  are  $g$ -dependent along an  $X$ -extremal if and only if  $s' = s$  and  $t' = \zeta(s, t)$ . Thus  $\zeta$  defines the mapping from  $q$  to its  $g$ -dependent point in this  $X$ -aligned chart of coordinates.

*Proof.* In these coordinates, we have that  $X \cong (0, 1)^T = \frac{\partial}{\partial t}$ , and we write  $Y \cong (a, b)^T$  for some differentiable functions  $a$  and  $b$ . Since  $X$  and  $Y$  are everywhere linearly independent on  $\Omega$ , the function  $a$  does not vanish on  $Q$ . Since  $X$ -trajectories are vertical lines, the variational equation (2.52) along  $X$ -extremals is simply  $\dot{w}(t) \equiv 0$ , and thus two points  $q = \psi(s, t)$  and  $q' = \psi(s', t')$  in  $\Omega$  are  $g$ -dependent along  $X$  if and only if  $s = s'$  and the vectors  $g(q)$  and  $g(q')$  are linearly dependent. In terms of the coordinates of the vector fields  $X$  and  $Y$ , we have that

$$g = \frac{1}{2}(Y - X) \cong \frac{1}{2} \begin{pmatrix} a \\ b - 1 \end{pmatrix},$$

and since  $a$  has constant sign in  $\Omega$ , the vectors  $g(q)$  and  $g(q')$  need to point in the same direction; that is,

$$\frac{b(s, t) - 1}{a(s, t)} = \frac{b(s, t') - 1}{a(s, t')}.$$

If we define

$$\theta : Q \rightarrow \mathbb{R}, \quad (s, t) \mapsto \theta(s, t) = \frac{b(s, t) - 1}{a(s, t)},$$

then

$$\frac{\partial \theta}{\partial t}(s, t) = \frac{\xi(s, t)}{a^2(s, t)},$$

where

$$\xi(s, t) = \frac{\partial b}{\partial t}(s, t)a(s, t) - (b(s, t) - 1)\frac{\partial a}{\partial t}(s, t).$$

This expression relates to the determinant of  $[f, g]$  and  $g$ : suppressing the arguments, we have that

$$[f, g] = [X, g] \cong Dg \cdot X = \frac{1}{2} \begin{pmatrix} \frac{\partial a}{\partial s} & \frac{\partial a}{\partial t} \\ \frac{\partial b}{\partial s} & \frac{\partial b}{\partial t} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \frac{\partial a}{\partial t} \\ \frac{\partial b}{\partial t} \end{pmatrix}$$

and thus

$$\det([f, g], g) \cong \frac{1}{4} \begin{vmatrix} \frac{\partial a}{\partial t} & a \\ \frac{\partial b}{\partial t} & b - 1 \end{vmatrix} = -\frac{1}{4}\xi(s, t).$$

Expressing the Lie bracket  $[f, g]$  in terms of  $f$  and  $g$ , we therefore get that

$$\xi(s, t) = -4 \det([f, g], g) = -4 \det(\alpha f + \beta g, g) = -4\alpha \det(f, g),$$

where  $\alpha$  and the vector fields  $f$  and  $g$  are evaluated at the point  $q = \psi(s, t) \in \Omega$ . In particular, since  $\alpha$  vanishes for  $t = 0$  ( $\psi(s, 0) \in \mathcal{S}$ ), it follows that  $\xi(s, 0) \equiv 0$ , and



therefore  $t$  can be factored from  $\xi(s, t)$ , say

$$\xi(s, t) = t\tilde{\xi}(s, t).$$

Thus, with all functions and vector fields evaluated at  $\psi(s, 0) \in \mathcal{S}$ , it follows that

$$\tilde{\xi}(s, 0) = \frac{\partial \xi}{\partial t}(s, 0) = -4 \frac{\partial}{\partial t} \Big|_{t=0} \left( \alpha \det(f, g) \right) = -4L_X \alpha \cdot \det(f, g) \neq 0.$$

Hence

$$\frac{\partial \theta}{\partial t}(s, 0) = 0$$

and

$$\frac{\partial^2 \theta}{\partial t^2}(s, t) = \frac{\frac{\partial \xi}{\partial t}(s, t)a(s, t) - 2\frac{\partial a}{\partial t}(s, t)\xi(s, t)}{a^3(s, t)}$$

gives that

$$\frac{\partial^2 \theta}{\partial t^2}(s, 0) = \frac{\tilde{\xi}(s, 0)}{a^2(s, 0)} \neq 0.$$

Overall, we therefore can write

$$\theta(s, t) = \theta(s, 0) + t^2\tilde{\theta}(s, t)$$

for some smooth function  $\tilde{\theta} = \tilde{\theta}(s, t)$  that satisfies  $\tilde{\theta}(s, 0) \neq 0$  for all  $s \in [-\varepsilon, \varepsilon]$ . By shrinking  $\varepsilon$  further, if necessary, we may assume that  $\tilde{\theta}(s, t)$  does not vanish on  $Q$ .

If one now expresses the difference

$$\delta(s, t, t') = \tilde{\theta}(s, t) - \tilde{\theta}(s, t')$$

as

$$\delta(s, t, t') = (t - t')\tilde{\delta}(s, t, t'),$$

then the equation  $\theta(s, t) = \theta(s, t')$  is equivalent to

$$\begin{aligned} 0 &= t^2\tilde{\theta}(s, t) - (t')^2\tilde{\theta}(s, t') \\ &= \left(t^2 - (t')^2\right)\tilde{\theta}(s, t') + t^2(\tilde{\theta}(s, t) - \tilde{\theta}(s, t')) \\ &= (t - t') \left[ (t + t')\tilde{\theta}(s, t') + t^2\tilde{\delta}(s, t, t') \right], \end{aligned}$$

and thus we need to solve the equation

$$\Delta(s, t, t') = (t + t')\tilde{\theta}(s, t') + t^2\tilde{\delta}(s, t, t') = 0.$$

Clearly,  $\Delta(0,0,0) = 0$  and

$$\frac{\partial \Delta}{\partial t'}(s,0,0) = \tilde{\theta}(s,0) \neq 0.$$

Hence, by the implicit function theorem, the equation  $\Delta(s,t,t') = 0$  can be solved for  $t'$  near  $(0,0,0)$  in terms of a differentiable function  $t' = \zeta(s,t)$ . Furthermore, for  $t = 0$ , we have that

$$0 = \Delta(s,0,\zeta(s,0)) = \zeta(s,0) \cdot \tilde{\theta}(s,\zeta(s,0)),$$

and since  $\tilde{\theta}(s,t)$  does not vanish, it follows that  $\zeta(s,0) \equiv 0$  for all  $s \in [-\varepsilon, \varepsilon]$ . Finally, differentiating  $\Delta(s,t,\zeta(s,t))$  with respect to  $t$  and setting  $t = 0$  gives

$$\begin{aligned} 0 &= \frac{\partial \Delta}{\partial t}(s,0,\zeta(s,0)) + \frac{\partial \Delta}{\partial t'}(s,0,\zeta(s,0)) \frac{\partial \zeta}{\partial t}(s,0) \\ &= \frac{\partial \Delta}{\partial t}(s,0,0) + \frac{\partial \Delta}{\partial t'}(s,0,0) \frac{\partial \zeta}{\partial t}(s,0). \end{aligned}$$

But

$$\frac{\partial \Delta}{\partial t}(s,0,0) = \frac{\partial \Delta}{\partial t'}(s,0,0) = \tilde{\theta}(s,0) \neq 0,$$

and therefore

$$\frac{\partial \zeta}{\partial t}(s,0) = -1.$$

□

We now prove Proposition 2.9.5: Let  $\Omega = \psi(Q(\varepsilon))$  be an  $X$ -aligned chart of coordinates and suppose  $\varepsilon$  is small enough that there exists a differentiable function  $\zeta : Q \rightarrow \mathbb{R}$ ,  $(s,t) \mapsto \zeta(s,t)$ , with the properties of Lemma 2.9.3. By making  $\varepsilon$  smaller if necessary, we also may assume that  $\frac{\partial \zeta}{\partial t}$  is negative on  $Q$ . As before,  $X \cong (0,1)^T = \frac{\partial}{\partial t}$  and we write  $Y \cong (a,b)^T$  for some differentiable functions  $a$  and  $b$ . In these coordinates,

$$L_X \alpha \cong \frac{\partial \alpha}{\partial t}$$

and

$$L_Y \alpha \cong \frac{\partial \alpha}{\partial s} \cdot a + \frac{\partial \alpha}{\partial t} \cdot b.$$

Since the singular curve  $\mathcal{S}$  is given by the  $s$ -axis, we have  $\alpha(s,0) \equiv 0$  and therefore  $\frac{\partial \alpha}{\partial s}(s,0) \equiv 0$  as well. Hence, at the reference point  $p$ , we get

$$\frac{L_Y \alpha(p)}{L_X \alpha(p)} \cong b(0,0), \quad (2.53)$$

and thus  $b(0,0)$  is negative. By choosing  $\varepsilon$  small enough, we may assume that  $b$  is negative everywhere on  $Q$ . Similarly, without loss of generality we assume that  $L_X\alpha > 0$  and  $L_Y\alpha < 0$  on all of  $\Omega$ .

Let  $(\bar{x}, \bar{u})$  be a time-optimal  $YXY$ -trajectory that transfers a point  $q_1 \in \Omega$  into  $q_2 \in \Omega$  with the entire trajectory  $\bar{x}$  lying in  $\Omega$ . Denote the switching times by  $\tau$  and  $\tau'$ ,  $\tau < \tau'$ , and the corresponding junctions by  $r$  and  $r'$ , respectively. The points  $r$  and  $r'$  are  $g$ -dependent along  $X$ , and thus if  $r = \psi(s, t)$  and  $r' = \psi(s', t')$ , then  $s' = s$  and  $t' = \zeta(s, t)$ . Note that  $t < 0$  and  $t' > 0$ . (For by Proposition 2.9.1,  $YX$ -junctions need to lie in  $\alpha \leq 0$  and  $XY$ -junctions in  $\alpha \geq 0$ . Since  $L_X\alpha > 0$  on  $\Omega$ , we thus have  $t \leq 0$  and  $t' \geq 0$ . But  $\zeta(s, 0) = 0$ , and thus neither can be zero, since otherwise  $r = r'$ .) The next lemma is one of the two key arguments in the construction, and it is only for this result that we need to make the neighborhood  $\Omega$  small.

**Lemma 2.9.4.** *Let  $\gamma$  denote the restriction of the  $YXY$ -trajectory  $\bar{x}$  to some small interval  $[\tau - \varepsilon, \tau]$ , where  $\tau$  is the first switching time and let  $\gamma'$  be the image of this curve under the mapping  $Z : (s, t) \mapsto (s, \zeta(s, t))$ . For  $\varepsilon$  sufficiently small, the curve  $\gamma'$  is a trajectory of the system.*

*Proof.* It suffices to show that the tangent vector to the curve  $\gamma'$  at the point  $r'$  is a linear combination of  $X(r')$  and  $Y(r')$  with positive coefficients. For if this is the case, then by choosing the times sufficiently close to  $\tau$ , at every point  $q'$  on the curve  $\gamma'$  there exists a continuous control  $u(q') \in (-1, 1)$  such that  $f(q') + u(q')g(q')$  is tangent to  $\gamma'$ . After a suitable reparameterization, the curve thus becomes a trajectory of  $\Sigma$ .

This property, however, can be guaranteed only in a sufficiently small neighborhood of  $p$ . The tangent vector  $t'$  to the curve  $\gamma'$  at  $r'$  is the image of the vector  $Y(r)$  under the differential of the mapping  $Z$ , i.e.,

$$\begin{aligned} t' &= \begin{pmatrix} 1 & 0 \\ \frac{\partial \zeta}{\partial s}(s, t) & \frac{\partial \zeta}{\partial t}(s, t) \end{pmatrix} \begin{pmatrix} a(s, t) \\ b(s, t) \end{pmatrix} = \begin{pmatrix} a(s, t) \\ \frac{\partial \zeta}{\partial s}(s, t)a(s, t) + \frac{\partial \zeta}{\partial t}(s, t)b(s, t) \end{pmatrix} \\ &= \frac{a(s, t)}{a(s, t')} \begin{pmatrix} a(s, t') \\ b(s, t') \end{pmatrix} + \left[ \frac{\partial \zeta}{\partial s}(s, t)a(s, t) + \frac{\partial \zeta}{\partial t}(s, t)b(s, t) - \frac{a(s, t)}{a(s, t')}b(s, t') \right] \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{a(s, t)}{a(s, t')} Y(r') + b(s, t) \left[ \frac{\partial \zeta}{\partial s}(s, t) \frac{a(s, t)}{b(s, t)} + \frac{\partial \zeta}{\partial t}(s, t) - \frac{a(s, t)}{a(s, t')} \frac{b(s, t')}{b(s, t)} \right] X(r'). \end{aligned} \tag{2.54}$$

Since  $a$  has constant sign on  $Q$ , the quotient  $\frac{a(s, t)}{a(s, t')}$  is positive. The function  $b$  is negative on  $Q$ , and by Lemma 2.9.3,

$$\frac{\partial \zeta}{\partial s}(s, 0) \frac{a(s, 0)}{b(s, 0)} + \frac{\partial \zeta}{\partial t}(s, 0) \equiv -1.$$

Hence, and once more by choosing the neighborhood  $Q$  small enough, we may assume that

$$\frac{\partial \zeta}{\partial s}(s, t) \frac{a(s, t)}{b(s, t)} + \frac{\partial \zeta}{\partial t}(s, t) < -\frac{1}{2} \quad \text{for all} \quad (s, t) \in Q.$$

Thus the coefficient at  $X(r')$  is positive as well.  $\square$

*Remark 2.9.1.* The construction of an  $X$ -aligned chart of coordinates  $\Omega = \psi(Q(\varepsilon))$  does not require that  $L_Y \alpha \neq 0$ , and it is still applicable if  $L_Y \alpha(p) = 0$ , since then  $L_g \alpha(p) = \frac{1}{2} L_X \alpha(p) > 0$ . But in this case  $b(0, 0) = 0$ , and thus the dominance argument above no longer can be made. For later reference, however, we already note here that such an argument is not needed at points where the Lie derivative of  $\zeta$  along  $Y$  is positive,

$$L_Y \zeta(s, t) = \frac{\partial \zeta}{\partial s}(s, t) a(s, t) + \frac{\partial \zeta}{\partial t}(s, t) b(s, t) > 0,$$

and where  $b(s, t')$  is negative. In this case, (2.54) directly gives that  $t'$  is a linear combination of  $X(r')$  and  $Y(r')$  with positive coefficients. This will allow us to deal with codimension-2 cases in the next section.

We now show that Lemma 2.54 precludes the optimality of the  $YXY$ -trajectory  $\bar{x}$ . In fact, the curve  $\gamma'$  is an envelope for the control system  $\Sigma$ , and the generalization of the theory of envelopes to optimal control shows that it cannot be optimal. We shall develop this theory for a general control problem in Sect. 5.4 but already here anticipate this argument with a direct calculation invoking the clock form  $\omega$  introduced earlier.

Let  $\Gamma$  be the restriction of the  $YXY$ -trajectory to the interval  $[\tau - \varepsilon, \tau']$  so that  $\Gamma$  is the concatenation of the curve  $\gamma$  with the  $X$ -trajectory that steers  $r$  into  $r'$ . Define another trajectory  $\Gamma'$  of  $\Sigma$  that steers the point  $\bar{x}(\tau - \varepsilon)$  into  $r'$  by first following the  $X$ -trajectory from  $\bar{x}(\tau - \varepsilon)$  to its  $g$ -dependent point on the curve  $\gamma'$  and then concatenating with the  $\Sigma$ -trajectory that corresponds to  $\gamma'$  (see Fig. 2.14).

**Lemma 2.9.5.** *The times along the trajectories  $\Gamma$  and  $\Gamma'$  are equal,  $T(\Gamma) = T(\Gamma')$ .*

*Proof.* The concatenation  $Y$  of  $\Gamma$  with the curve  $\Gamma'$  run backward is a closed curve, and by Stokes's theorem, the difference in the times along these two trajectories is given by

$$T(\Gamma) - T(\Gamma') = \int_Y \omega = \int_R d\omega,$$

where  $R$  denotes the region enclosed by  $Y$ . The coordinate expression for  $\omega$  (see Eq. (2.49)) is given by

$$\omega = \frac{g_2 ds - g_1 dt}{\det(f, g)} = \frac{\frac{1}{2}(b-1)ds - \frac{1}{2}adt}{-\frac{1}{2}a} = dt + \frac{1-b(s, t)}{a(s, t)} ds$$

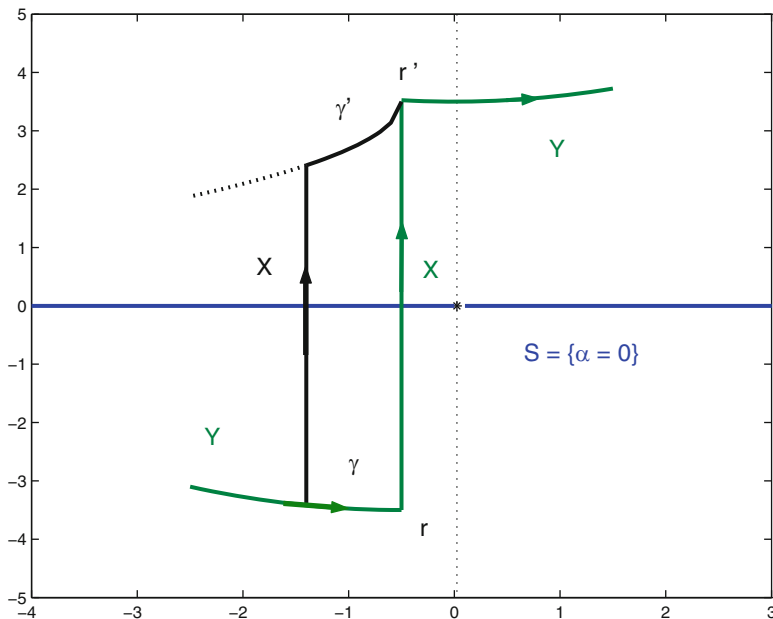


Fig. 2.14 Conjugate curve  $\gamma'$

and thus, and using the notation  $\theta$  from the proof of Lemma 2.9.3,

$$d\omega = \frac{\partial}{\partial t} \left( \frac{b(s,t) - 1}{a(s,t)} \right) ds \wedge dt = \frac{\partial \theta}{\partial t}(s,t) ds \wedge dt.$$

Since  $Y$  is transversal to  $X$ , we can parameterize the curve  $\gamma$  as the graph of a function  $\sigma$  of  $s$  over some interval  $[s_\varepsilon, s_\tau]$ , say  $\gamma: [s_\varepsilon, s_\tau] \rightarrow Q$ ,  $s \mapsto \gamma(s) = (s, \sigma(s))$ . Evaluating the double integral by integrating over the vertical segments therefore gives

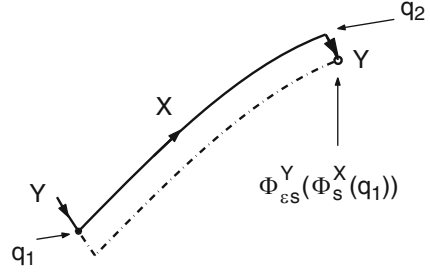
$$\begin{aligned} \int_R d\omega &= \int_{s_\varepsilon}^{s_\tau} \int_{\sigma(s)}^{\zeta(s, \sigma(s))} \frac{\partial \theta}{\partial t}(s,t) dt ds \\ &= \int_{s_\varepsilon}^{s_\tau} [\theta(s, \zeta(s, \sigma(s))) - \theta(s, \sigma(s))] ds. \end{aligned}$$

But by construction, the points  $(s, \sigma(s))$  on  $\gamma$  and  $(s, \zeta(s, \sigma(s)))$  on  $\gamma'$  are  $g$ -dependent, and therefore for all  $s \in [s_\varepsilon, s_\tau]$ ,

$$\theta(s, \zeta(s, \sigma(s))) = \theta(s, \sigma(s)).$$

Hence  $\int_R d\omega = 0$  and thus  $T(\Gamma) = T(\Gamma')$ .  $\square$

**Fig. 2.15** A variation along a  $YXY$ -trajectory



This precludes the optimality of  $\Gamma$ : for if  $\Gamma$  is time-optimal, then so is  $\Gamma'$ . But the dynamics along  $\gamma'$  is a strict convex combination of  $X$  and  $Y$ , and thus the control takes values in the interior of the control set. Hence it must be singular. But  $\alpha(r') > 0$ , and so this is not possible. Contradiction.

Since the roles of  $X$  and  $Y$  are reversible in our assumptions, it similarly can be shown that  $XYX$ -trajectories cannot be optimal either, and thus Proposition 2.9.5 is proven.  $\square$

This proof is the original one by H. Sussmann, and it beautifully illustrates the underlying *geometric* aspects (i.e., conjugate points and envelopes) of the structure of optimal bang-bang trajectories near a slow singular arc. We shall return to this topic for a general  $n$ -dimensional system in Sect. 6.1.3 about transversal folds.

There exists an alternative, and in some sense more direct, *algebraic* approach that is based on a variation analogous to the one used in [209] for the three-dimensional case. Suppose again that  $\Gamma$  is a  $YXY$ -trajectory with the switching points given by  $q_1$  and  $q_2 = \Phi_s^X(q_1)$ . It is geometrically clear (see Fig. 2.15), and not difficult to verify analytically, that there exist continuously differentiable positive functions  $r = r(\epsilon)$  and  $t = t(\epsilon)$  such that

$$\Phi_{\epsilon s}^Y(\Phi_s^X(q_1)) = \Phi_{r(\epsilon)}^X(\Phi_{t(\epsilon)}^Y(q_1)).$$

The difference in time between these trajectories is then given by

$$\Delta(\epsilon) = s(1 + \epsilon) - t(\epsilon) - s(\epsilon),$$

and the  $YXY$ -trajectory is not time-optimal if  $\Delta(\epsilon) > 0$  for small  $\epsilon > 0$ . Hence, the optimality of bang-bang trajectories with two switchings can be excluded by computing the Taylor expansion of  $\Delta$  at  $\epsilon = 0$ . It can be shown that the fact that  $q_1$  and  $q_2$  are  $g$ -dependent points is equivalent to  $\Delta'(0) = 0$ , and it thus becomes necessary to compute the second derivative  $\Delta''(0)$ . This, however, requires a good algebraic framework that is provided by a Lie-algebraic formalism that we shall establish only in Sect. 4.5. We shall return to this second approach in Sect. 7.3 when we analyze the corresponding situation—the structure of time-optimal bang-bang trajectories near a slow singular arc—in dimension three.

## 2.10 Input Symmetries and Codimension-2 Cases in the Plane

The results of the last section cover the local structure of time-optimal controlled trajectories near a point  $p$  in the plane under codimension-0 and some codimension-1 conditions. In order to classify the structure of time-optimal controlled trajectories for a generic time-invariant control-affine nonlinear system of the type [NTOC] in the plane, by Thom's transversality theorem [108] it is necessary to analyze all other possible codimension-1 and codimension-2 conditions. Other codimension-1 conditions arise if assumption (A0) is violated, i.e., if the vector fields  $f$  and  $g$  are linearly dependent at  $p$ ; codimension-2 conditions arise if two independent equality relations are imposed. In this section, we still analyze those codimension-2 situations that arise if condition (A0) is met. These correspond to situations in which in addition to  $\alpha(p)$ , also one of the Lie derivatives of  $\alpha$  along  $X$  or  $Y$  vanishes at  $p$ . The results of this and the previous section then collectively describe the structure of time-optimal controls near a reference point  $p$  where the vector fields  $f(p)$  and  $g(p)$  are linearly independent under otherwise generic conditions on the vector fields  $f$  and  $g$ .

### 2.10.1 Input Symmetries

In cases of higher codimensions, the number of possibilities increases significantly, and it now helps to use *input symmetries* and other invariances to reduce this number. Since the control set  $U = [-1, 1]$  is invariant under a reflection at the origin, the problem [NTOC] remains unchanged if we use as control  $v = -u$  instead. This transformation, however, changes the vector fields:  $g$  becomes  $-g$  while  $f$  remains the same. Thus their Lie brackets and hence also the functions  $\alpha$  and  $\beta$  and their Lie derivatives are affected. This allows us to normalize the signs of some of these functions.

**Definition 2.10.1 (Input symmetry).** An input symmetry is a linear transformation on the vector fields  $f$  and  $g$  that leaves the control system  $\Sigma : \dot{x} = f(x) + ug(x)$ ,  $u \in U$  (including the class of admissible controls), invariant.

**Definition 2.10.2 (Reflection).** For the system  $\Sigma : \dot{x} = f(x) + ug(x)$ ,  $|u| \leq 1$ , define the reflection  $\rho$  by  $\rho(f) = f$  and  $\rho(g) = -g$ , or equivalently, as the transformation that interchanges the vector fields  $X$  and  $Y$ ,

$$\rho(X) = \rho(f - g) = \rho(f) - \rho(g) = f + g = Y$$

and

$$\rho(Y) = \rho(f + g) = \rho(f) + \rho(g) = f - g = X.$$

This definition naturally extends as a *homomorphism* to the Lie algebra generated by the vector fields  $f$  and  $g$  if we define

$$\rho([f, g]) = [\rho(f), \rho(g)]$$

and inductively extend this relation to higher-order Lie brackets. As before, we assume that  $\Omega$  is a simply connected region of  $\mathbb{R}^2$  and that  $f$  and  $g$  are linearly independent vector fields on  $\Omega$ . Thus, all higher-order Lie brackets of  $f$  and  $g$  can be expressed as linear combinations of  $f$  and  $g$  with coefficients that are smooth functions of  $x$ . Suppose  $[f, g](x) = \alpha(x)f(x) + \beta(x)g(x)$  and write

$$\rho([f, g]) = \rho(\alpha)\rho(f) + \rho(\beta)\rho(g).$$

The effects that an input symmetry has on the higher-order brackets and coordinate expressions can then easily be calculated through straightforward algebraic substitutions. We have that

$$[\rho(f), \rho(g)] = -[f, g] = -\alpha f - \beta g = -\alpha\rho(f) + \beta\rho(g)$$

and thus

$$\rho(\alpha) = -\alpha \quad \text{and} \quad \rho(\beta) = \beta. \quad (2.55)$$

Considering higher-order brackets, we arrive at analogous formulas for the Lie derivatives of  $\alpha$  and  $\beta$ :

$$\begin{aligned} [X, [f, g]] &= [X, \alpha f + \beta g] = L_X(\alpha)f + \alpha[X, f] + L_X(\beta)g + \beta[X, g] \\ &= L_X(\alpha)f + L_X(\beta)g + (\alpha + \beta)[f, g] \\ &= (L_X(\alpha) + (\alpha + \beta)\alpha)f + (L_X(\beta) + (\alpha + \beta)\beta)g, \end{aligned}$$

and analogously

$$[Y, [f, g]] = (L_Y(\alpha) - (\alpha - \beta)\alpha)f + (L_Y(\beta) - (\alpha - \beta)\beta)g.$$

Applying the input symmetry  $\rho$ , we have that

$$\begin{aligned} \rho([X, [f, g]]) &= [\rho(X), [\rho(f), \rho(g)]] = -[Y, [f, g]] \\ &= -(L_Y(\alpha) - (\alpha - \beta)\alpha)f - (L_Y(\beta) - (\alpha - \beta)\beta)g \\ &= (-L_Y(\alpha) + (\rho(\alpha) + \rho(\beta))\rho(\alpha))\rho(f) \\ &\quad + (L_Y(\beta) + (\rho(\alpha) + \rho(\beta))\rho(\beta))\rho(g), \end{aligned}$$

and therefore

$$\rho(L_X(\alpha)) = -L_Y(\alpha) \quad \text{and} \quad \rho(L_X(\beta)) = L_Y(\beta).$$



Since  $-L_Y(\alpha) = L_Y(-\alpha) = L_{\rho(X)}(\rho(\alpha))$ , this relation can succinctly be expressed in the form

$$\rho(L_X(\alpha)) = L_{\rho(X)}(\rho(\alpha)). \quad (2.56)$$

Analogously, it follows that

$$\rho(L_Y(\alpha)) = L_{\rho(Y)}(\rho(\alpha)) = -L_X(\alpha)$$

and

$$\rho(L_Y(\beta)) = L_{\rho(Y)}(\rho(\beta)) = L_X(\beta).$$

Similarly, for higher-order derivatives we have that

$$\rho(L_X^2(\alpha)) = L_{\rho(X)}(L_{\rho(X)}(\rho(\alpha))) = L_Y(L_Y(-\alpha)) = -L_Y^2(\alpha)$$

and

$$\rho(L_Y^2(\alpha)) = L_{\rho(Y)}(L_{\rho(Y)}(\rho(\alpha))) = L_X(L_X(-\alpha)) = -L_X^2(\alpha),$$

and so on. Once more, *the effects that an input symmetry has on the vector fields  $f$  and  $g$  and their Lie brackets are easily obtained through straightforward algebraic substitutions.*

We briefly reconsider the results of the previous section with this point of view. If  $\alpha$  is positive on some region  $\Omega$ , we have shown that optimal controlled trajectories that lie in  $\Omega$  have at most the structure  $XY$ . Applying the input symmetry  $\rho$  to the system changes the sign of  $\alpha$  and interchanges  $X$  with  $Y$ . Thus, it directly follows that optimal controlled trajectories are at most of type  $YX$  if  $\alpha$  is negative (see Proposition 2.9.1). On the other hand, in the codimension-1 situation (A1), the relevant conditions are all invariant under this input symmetry. For example, the singular arc is given by

$$S = f + \frac{L_X\alpha(x) + L_Y\alpha(x)}{L_X\alpha(x) - L_Y\alpha(x)}g = f + \frac{L_f\alpha}{L_g\alpha}g$$

and

$$\rho(S) = \rho(f) + \frac{\rho(L_f\alpha(x))}{\rho(L_g\alpha(x))}\rho(g) = f + \frac{-L_f\alpha(x)}{L_g\alpha(x)}(-g) = S.$$

Naturally, the strengthened Legendre–Clebsch condition (see Eq. (2.47),

$$\langle \lambda(t), [g, [f, g]](x(t)) \rangle = -L_g\alpha(x(t)),$$

is invariant under this input symmetry as well. In fact, the assumptions for each of the various codimension-1 cases considered in the last section are invariant under  $\rho$ . Still, this input symmetry is useful in the proof of Proposition 2.9.5, where we carried out the construction only for  $YXY$ -trajectories and merely claimed that the analogous construction excludes  $XYX$ -trajectories as well. Since  $\rho$  interchanges  $L_X(\alpha)$  and  $-L_Y(\alpha)$ ,

$$\rho(L_X(\alpha)) = -L_Y(\alpha) \quad \text{and} \quad \rho(L_Y(\alpha)) = -L_X(\alpha),$$

the assumptions of Proposition 2.9.5 are invariant under  $\rho$ , and thus, applying  $\rho$ , it immediately follows that  $XYX$ -trajectories cannot be optimal either. No further argument is necessary.

It is in the codimension-2 scenario, that input symmetries really become useful. We can limit our analysis to the case that one of the Lie derivatives of  $\alpha$  with respect to  $X$  or  $Y$  vanishes, and without loss of generality, we shall consider the case when

$$L_X(\alpha)(p) \neq 0 \quad \text{and} \quad L_Y(\alpha)(p) = 0, \quad \text{while} \quad L_Y^2(\alpha)(p) \neq 0.$$

Using a second symmetry that optimal trajectories possess, we can in addition normalize the sign for the second Lie derivative  $L_Y^2(\alpha)(p)$ . Time-optimal trajectories are also invariant under *time reversal*. If  $(x_*, u_*)$  is a time-optimal trajectory for the system  $\Sigma : \dot{x} = f(x) + ug(x)$ ,  $|u| \leq 1$ , defined over an interval  $[0, T]$  that steers a point  $q_1$  into  $q_2$ , then the pair  $(y_*, v_*)$  defined by  $y_*(t) = x_*(T - t)$  and  $v_*(t) = u_*(T - t)$  is a time-optimal trajectory that steers  $q_2$  into  $q_1$  for the system  $\check{\Sigma} : \dot{y} = \check{f}(y) + v\check{g}(y)$ ,  $|v| \leq 1$ , where time has been reversed. Since

$$\begin{aligned} \dot{y}_*(t) &= -\dot{x}_*(T - t) = -f(x_*(T - t)) - u_*(T - t)g(x_*(T - t)) \\ &= -f(y(t)) - v(t)g(y(t)), \end{aligned}$$

this property can be expressed in terms of a second input symmetry that reverses the signs of the vector fields  $f$  and  $g$ .

**Definition 2.10.3 (Time reversal).** For the system  $\Sigma : \dot{x} = f(x) + ug(x)$ ,  $|u| \leq 1$ , define time reversal  $\tau$  by  $\tau(f) = -f$  and  $\tau(g) = -g$ , or equivalently, by  $\tau(X) = -X$  and  $\tau(Y) = -Y$ .

As above, we extend this definition to the Lie algebra generated by  $f$  and  $g$  and then calculate the relations it implies on the coordinates with respect to the basis in terms of  $f$  and  $g$ . Simple computations verify that

$$\begin{aligned} \tau(\alpha) &= -\alpha, & \tau(\beta) &= -\beta, \\ \tau(L_X(\alpha)) &= L_X(\alpha), & \tau(L_Y(\alpha)) &= L_Y(\alpha), \\ \tau(L_X^2(\alpha)) &= -L_X^2(\alpha), & \tau(L_Y^2(\alpha)) &= -L_Y^2(\alpha), \end{aligned}$$

and it is the last relation that, without loss of generality, allows us to assume that  $L_Y^2(\alpha)$  is positive.

In a more abstract framework, the input symmetries generate a group  $\mathcal{G} = \{\text{id}, \rho, \tau, \tau \circ \rho\}$  of idempotent elements (i.e.,  $\rho \circ \rho = \text{id}$ , etc.) and using them, it is possible to reduce the number of codimension-2 scenarios by a factor of 4. It is to be expected that the mathematically more difficult scenarios arise when the Lie bracket configurations are invariant under this group of symmetries, and this will

again happen for nongeneric codimension-3 situations. The codimension-2 cases, however, essentially can be fully analyzed based on the earlier codimension-1 results of Sect. 2.9 and some additional geometric considerations.

### 2.10.2 Saturating Singular Arcs

We now assume that

- (A2) the vector fields  $f$  and  $g$  are linearly independent everywhere on  $\Omega \subset \mathbb{R}^2$  and there exists a point  $p \in \Omega$  with  $\alpha(p) = 0$ , but the Lie derivative of  $\alpha$  along  $X$  does not vanish on  $\Omega$ ,

$$\alpha(p) = 0, \quad L_X(\alpha)(x) \neq 0 \quad \text{for all } x \in \Omega;$$

furthermore, the Lie derivative of  $\alpha$  along  $Y$  vanishes at  $p$ , but the second Lie derivative of  $\alpha$  along  $Y$  is positive on  $\Omega$ ,

$$L_Y(\alpha)(p) = 0, \quad L_Y^2(\alpha)(x) > 0 \quad \text{for all } x \in \Omega.$$

Note that  $L_g\alpha(p) = \frac{1}{2}L_X(\alpha)(p) \neq 0$ , and thus there exists an  $X$ -aligned chart of coordinates (centered at  $p$ ),  $\psi: Q(\varepsilon) \subset \mathbb{R}^2 \rightarrow \Omega = \psi(Q(\varepsilon))$ ,  $(s, t) \mapsto \psi(s, t) = \Psi_t^X \circ \Psi_s^S(p)$ . As above, in these coordinates  $X \cong (0, 1)^T = \frac{\partial}{\partial t}$ , and we write  $Y \cong (a, b)^T$  for some differentiable functions  $a$  and  $b$ . Since  $X$  and  $Y$  are everywhere linearly independent on  $\Omega$ , the function  $a$  does not vanish on  $Q$ , and without loss of generality, we assume that  $a$  is positive on  $Q$ . (If  $a$  is negative, then simply change  $s$  in the definition of the coordinates to  $-s$ .) Assumption (A2) also implies that the integral curve  $\gamma$  of  $Y$  through the point  $p$  is tangent to the curve  $S = \{x \in \Omega : \alpha(x) = 0\}$  at  $p$  and that the order of contact is 1, i.e., for  $r$  near 0,

$$\begin{aligned} \alpha(\gamma(r)) &= \alpha(p) + L_Y\alpha(p)r + \frac{1}{2}L_Y^2\alpha(p)r^2 + o(r^2) \\ &= \frac{1}{2}L_Y^2\alpha(p)r^2 + o(r^2). \end{aligned}$$

Hence, except for the point  $p$ , the curve  $\gamma$  lies in  $\Omega_+ = \{x \in \Omega : \alpha(x) > 0\}$  and can be parameterized as the graph of a function of  $s$ . By choosing  $\varepsilon$  sufficiently small, we again can assume that this parameterization is defined on the full interval  $[-\varepsilon, \varepsilon]$ , say  $\gamma: [-\varepsilon, \varepsilon] \rightarrow Q(\varepsilon)$ ,  $s \mapsto (s, y(s))$ , and  $y'(0) = 0$ . The geometry is illustrated in Fig. 2.16.

The point  $p$  is the beginning or end point of an admissible singular arc. The singular control at  $p$  is given by

$$u_{\text{sing}}(p) = \frac{L_X\alpha(p) + L_Y\alpha(p)}{L_X\alpha(p) - L_Y\alpha(p)} = +1,$$



following specifications:

$$\begin{aligned} R_0 &= \{(s, t) \in Q : t > y(s)\}, \\ R_1 &= \{(s, t) \in Q : s < 0, t < y(s)\}, \end{aligned}$$

and

$$R_2 = \{(s, t) \in Q : s > 0, t < y(s)\}.$$

Thus  $R_0$  is the set above the integral curve  $Y$ , and the region below this curve is divided further into its components in  $\{s < 0\}$  and  $\{s > 0\}$  with the boundaries given by the trajectory  $Y$  and the negative  $t$ -axis,  $\{(s, t) \in Q : s = 0, t < 0\}$ . Since  $X \cong (0, 1)^T = \frac{\partial}{\partial t}$  is vertical,  $X$ -trajectories cross  $Y$  into  $R_0$ . Once there, since  $R_0 \subset \Omega_+$ , at most one switching from  $X$  to  $Y$  can occur in  $R_0$  and thus trajectories cannot leave  $R_0$  forward in time as long as they are contained in  $\Omega$ . If an optimal trajectory were to switch from  $X$  to  $Y$  on the curve  $Y$ , then another junction with  $X$  is possible only at  $p$  followed possibly by one more switch to  $Y$ . It will follow from our argument below that no prior switchings can exist in this case, and overall, such a trajectory is at most of type  $XYXY$ .

The switchings in  $R_1$  and  $R_2$  can be analyzed with the tools developed in the proof of Proposition 2.9.5. By choosing  $\varepsilon$  small enough, we can assume that the function  $\zeta$  constructed in Lemma 2.9.3 exists on  $Q(\varepsilon)$  with the properties specified there. It was shown in the proof of Proposition 2.9.5 that  $YXY$ -trajectories are not optimal if the component  $b$  in the vector field  $Y$  is negative over the neighborhood  $Q(\varepsilon)$ , but under assumption (A2) the function  $b$  vanishes at  $p$ , and we first need to analyze its zero set in  $Q$ .

We first show that under assumption (A2), we have that

$$b(0, 0) = 0 \quad \text{and} \quad \frac{\partial b}{\partial s}(s, 0) > 0 \quad \text{for all } s \in [-\varepsilon, \varepsilon].$$

For recall from the proof of Proposition 2.9.5 that  $L_X \alpha = \frac{\partial \alpha}{\partial t}$  and

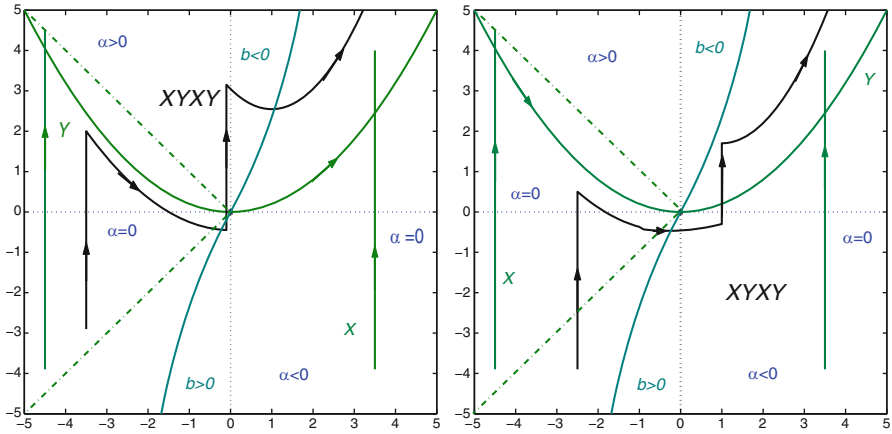
$$L_Y \alpha(s, t) = \frac{\partial \alpha}{\partial s}(s, t) \cdot a(s, t) + \frac{\partial \alpha}{\partial t}(s, t) \cdot b(s, t).$$

The singular curve  $\mathcal{S}$  is given by the  $s$ -axis,  $\alpha(s, 0) \equiv 0$ , and therefore  $\frac{\partial \alpha}{\partial s}(s, 0) \equiv 0$ . Hence we have along the  $s$ -axis that

$$L_Y \alpha(s, 0) = L_X \alpha(s, 0) \cdot b(s, 0), \tag{2.57}$$

and thus, under assumption (A2), it follows that  $b(0, 0) = 0$ . Differentiating Eq. (2.57) once more along the vector field  $Y$ , we get that

$$L_Y^2 \alpha(s, 0) = L_Y L_X \alpha(s, 0) \cdot b(s, 0) + L_X \alpha(s, 0) \cdot \left( \frac{\partial b}{\partial s}(s, 0) \cdot a(s, 0) + \frac{\partial b}{\partial t}(s, 0) \cdot b(s, 0) \right)$$



**Fig. 2.17** Optimal XYXY-trajectories for  $L_X \alpha > 0$

and therefore, upon evaluation at  $p \cong (0, 0)$ ,

$$L_Y^2 \alpha(0, 0) = L_X \alpha(0, 0) \cdot \frac{\partial b}{\partial s}(0, 0) \cdot a(0, 0).$$

Hence, with our normalization of  $a$  to be positive, we get that  $\frac{\partial b}{\partial s}(0, 0) > 0$ . In particular,  $b(s, 0)$  is negative for  $s < 0$  and positive for  $s > 0$ . It follows from the implicit function theorem that the equation  $b(s, t) = 0$  can be solved in terms of a differentiable function  $s = \sigma(t)$ ,  $\sigma(0) = 0$ , in a neighborhood of the origin. By making  $\varepsilon$  smaller, if necessary, we may assume that the function  $\sigma$  is defined over the full interval  $[-\varepsilon, \varepsilon]$ . Furthermore, the graph of  $\sigma$  is transversal to the integral curve  $Y$  of  $Y$  at  $p$  (see Figs. 2.16 and 2.17).

We also need to know the signs of the Lie derivative of the function  $\zeta$  along the vector field  $Y$ . By definition,

$$L_Y \zeta(s, t) = \frac{\partial \zeta}{\partial s}(s, t) a(s, t) + \frac{\partial \zeta}{\partial t}(s, t) b(s, t),$$

and it follows from Lemma 2.9.3 that  $\zeta(s, 0) \equiv 0$  and  $\frac{\partial \zeta}{\partial t}(s, 0) \equiv -1$ . In particular, all partial derivatives of  $\zeta$  with respect to  $s$  vanish along  $s = 0$ . Hence we have that  $L_Y \zeta(0, 0) = 0$  and for  $s < 0$ ,

$$L_Y \zeta(s, 0) = \frac{\partial \zeta}{\partial s}(s, 0) a(s, 0) + \frac{\partial \zeta}{\partial t}(s, 0) b(s, 0) = -b(s, 0) > 0.$$

Furthermore, with  $b(0, 0) = 0$ , the second Lie derivative with respect to  $Y$  at the origin simplifies to

$$L_Y^2 \zeta(0,0) = \frac{\partial \zeta}{\partial t}(0,0) \frac{\partial b}{\partial s}(0,0) a(0,0) = -\frac{\partial b}{\partial s}(0,0) < 0.$$

Thus the zero set  $Z = \{(s,t) : L_Y \zeta(s,t) = 0\}$  of the Lie derivative  $L_Y \zeta$  near the origin is a one-dimensional embedded submanifold that is transversal to the singular curve  $\mathcal{S} = \{(s,0) : |s| \leq \varepsilon\}$ .

**Lemma 2.10.1.** *YXY-trajectories that lie in the closure of  $R_1$  are not optimal.*

*Proof.* Since the zero sets of  $b$  and  $L_Y \zeta$  are transversal to  $Y$  at  $p$ , it follows that there exists an open sector  $V = \{(s,t) \in Q : s < 0, t < 2\omega|s|\}$  that lies entirely in the set  $\{(s,t) \in Q : b(s,t) < 0, L_Y \zeta(s,t) > 0\}$ . Since  $Y$  is tangent to the  $s$ -axis at  $p$ , by making  $\varepsilon$  smaller if necessary, we may assume that the curve  $Y$  for  $s < 0$  lies entirely in the smaller sector  $W = \{(s,t) \in Q : s < 0, t < \omega|s|\}$  (see Fig. 2.17). Now consider a  $YXY$ -trajectory that lies in  $R_1$  and suppose it has switchings at the points  $(\tilde{s}, \tilde{t})$  and  $(\tilde{s}, \tilde{t}')$ , respectively. If this trajectory is optimal, then the two junctions are  $g$ -dependent along  $X$  and we have that  $\tilde{t}' = \zeta(\tilde{s}, \tilde{t})$ . Since junctions of the type  $XY$  are optimal only in  $\Omega_+$ , we have that  $(\tilde{s}, \tilde{t}') \in W$ . It follows from Lemma 2.9.3 that

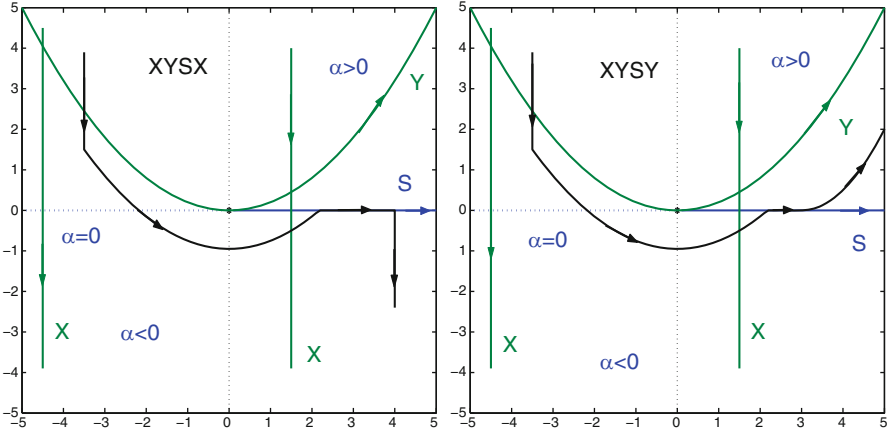
$$\tilde{t}' = \zeta(\tilde{s}, \tilde{t}) = \frac{\partial \zeta}{\partial t}(\tilde{s}, 0) \tilde{t} + o(\tilde{t}) = -\tilde{t} + o(\tilde{t}).$$

(We have  $\zeta(s,0) \equiv 0$ , and thus all derivatives in  $s$  vanish identically.) But then, for  $\varepsilon$  small enough, the first junction point  $(\tilde{s}, \tilde{t})$  still must lie in the larger sector  $V$  where the Lie derivative  $L_Y \zeta$  is positive, and by construction the second junction point lies in the region where  $b$  is negative. It follows from the remark following Lemma 2.9.4 that an envelope can be constructed, and thus this trajectory cannot be optimal.  $\square$

In particular, if there is an  $XY$ -junction on the curve  $Y$ , then there could not have been a previous  $YX$ -junction. Hence, as claimed earlier, such a trajectory can be at most of type  $XYXY$ .

The remainder of the argument follows from a direct geometric analysis of  $X$  and  $Y$  trajectories. It is possible that optimal trajectories are of the type  $XYXY$  in  $\{s < 0\}$ , but then the last switching must lie above  $Y$  in  $R_0$ , and overall such a trajectory cannot switch any more. Trajectories that do not cross  $Y$  in  $\{s < 0\}$  are at most concatenations of type  $XY$  in  $\{s < 0\}$ . If they switch to  $X$  at  $s = 0$ , then once more, only one additional switch to  $Y$  in  $\{t > 0\}$  is possible. If they cross  $\{s = 0\}$  along  $Y$ , then it is possible to have a switch to  $X$  in the fourth quadrant  $\{(s,t) \in Q : s > 0, t < 0\}$  and one more switch to  $Y$  in the first quadrant  $\{(s,t) \in Q : s > 0, t > 0\}$  (cf., Fig. 2.17). In any case, an optimal controlled bang-bang trajectory that lies in  $\Omega$  can have at most the concatenation sequence  $XYXY$ .  $\square$

**Proposition 2.10.2.** *Let  $\Omega$  be a domain on which condition (A2) is satisfied and suppose  $L_X(\alpha) = X\alpha$  is negative everywhere on  $\Omega$ . Then, for  $\Omega$  sufficiently small, optimal controlled trajectories that lie entirely in  $\Omega$  are at most concatenations of type  $XYSB$ .*



**Fig. 2.18** Optimal  $XYSB$ -trajectories for  $L_X \alpha < 0$

*Proof.* This is the easier case, and the result follows by a direct geometric reasoning from the codimension-1 scenarios. As above, consider an  $X$ -aligned chart of coordinates (centered at  $p$ ),  $\psi : Q(\varepsilon) \rightarrow \Omega = \psi(Q(\varepsilon))$ . In this case  $X$  and  $Y$  point to the same sides of the  $s$ -axis for  $s < 0$  and to opposite sides for  $s > 0$ . Furthermore, now  $\mathcal{S}_+ = \{(s, 0) : s > 0\}$  is a fast singular arc (that again saturates with  $u_{\text{sing}}(p) = +1$  at  $p$ ) and  $\mathcal{S}_- = \{(s, 0) : s < 0\}$  is inadmissible. By choosing  $Q$  small enough, it follows from Proposition 2.9.2 that optimal controlled trajectories that lie in  $Q_- = \{(s, t) : s < 0\}$ , the second and third quadrants, are at most of type  $XYX$ , and by Proposition 2.9.4, optimal controlled trajectories that lie in  $Q_+ = \{(s, t) : s > 0\}$ , the first and fourth quadrants, are at most of type  $BSB$ . However, overall, at most the concatenation sequence  $XYSB$  is possible. For only  $Y$ -trajectories can cross from  $Q_-$  into  $Q_+$ , and  $XY$ -junctions are optimal only in  $\Omega_+$ , the first and second quadrants, while  $YX$ -junctions are optimal only in  $\Omega_-$ , the third and fourth quadrants. Therefore, if a  $Y$ -trajectory crosses the  $t$ -axis at a positive value, then no further switching is possible, and such a trajectory can be at most of type  $XY$ . If it crosses for  $t = 0$  and does not switch at  $p$ , the same is true. If it switches at  $p$  to a singular arc, then only  $SB$  is possible afterward, and this limits the concatenation sequence to  $XYSB$ . If a switch to  $X$  occurs at  $p$ , then again no further switches are possible, and such a trajectory is at most of type  $XYX$ . Finally, if the crossing happens for  $t < 0$ , then it is possible to switch to  $X$  in the fourth quadrant (or also on the  $t$ -axis itself), and again in such a case we get at most  $XYX$ . If there is no switch to  $X$ , then the  $Y$ -trajectory may reach the singular arc and switch there, ending up with an  $SB$  concatenation. Overall, because of the directions of the vector fields  $X$  and  $Y$  near  $p$ , only concatenations of type  $XYSB$  can be optimal (see Fig. 2.18).  $\square$

Altogether, we have shown the following result:

**Theorem 2.10.1.** *Let  $p$  be a point where the vector fields  $f(p)$  and  $g(p)$  are linearly independent. Then, under generic conditions on the vector fields  $f$  and  $g$ , there exists*



a neighborhood  $\Omega$  of  $p$  such that optimal controlled trajectories that lie entirely in  $\Omega$  are concatenations of at most four pieces of either  $X = f - g$ ,  $Y = f + g$ , or the singular arc  $S$ . At most one of these pieces can be a singular arc, and if there are four segments, then it must be the second or third leg in the concatenation sequence. ■

In all the examples considered here, there is a very simple relation between the number of  $X$ ,  $Y$ , and singular segments in concatenation sequences that lie in a sufficiently small neighborhood  $\Omega$  of some reference point  $p$  and what is called the codimension of the *Lie-bracket configuration* of the system  $\Sigma = (f, g)$  at the point  $p$  that we still briefly want to point out. Loosely speaking, this Lie-bracket configuration consists of all the values of the vector fields  $f$  and  $g$  and their Lie brackets at  $p$ , and its *codimension* is given by the number of linearly independent “relevant” equality relations that hold between these vector fields at  $p$ . We are assuming that  $f$  and  $g$  are linearly independent on  $\Omega$  and thus always can express the Lie bracket as  $[f, g] = \alpha f + \beta g$  with some smooth functions  $\alpha$  and  $\beta$  defined on  $\Omega$ . In this case, the first “relevant” relation is that  $g$  and  $[f, g]$  are linearly dependent at  $p$ , characterized by  $\alpha(p) = 0$ . If  $\alpha$  does not vanish on  $\Omega$ , the codimension-0 case, optimal controls are simply bang-bang with one switching, and the sign of  $\alpha$  determines the order of the switchings. If  $\alpha$  does vanish at  $p$ , higher-order terms in the Taylor expansion of  $\alpha$  along the flows of  $X$  and  $Y$  at  $p$  matter, and depending on whether these Lie derivatives of  $\alpha$  vanish at  $p$ , more degenerate scenarios arise. In the codimension-1 cases, characterized by the fact that both Lie derivatives of  $\alpha$  along  $X$  and  $Y$  do not vanish at  $p$ , only three segments are possible. If we allow that one of the Lie derivatives vanishes, but again in a nondegenerate way, so that its second Lie derivative is nonzero, the codimension-2 case, this number increases to four. Overall, in each case we have the following simple relation:

$\Sigma_p$ : The maximum number of concatenations of  $X$ ,  $Y$ , and singular segments in time-optimal controlled trajectories that lie in a sufficiently small neighborhood  $\Omega$  of some reference point  $p$  is given by

$$2 + \text{codim}(\Sigma_p) = \dim \Omega + \text{codim}(\Sigma_p).$$

This relation has also been verified for numerous cases of low codimension in dimensions 3 and 4 (e.g., see [210, 211, 221]). For example, the possible concatenation sequences  $BBB$  and  $BSB$  that arise in the codimension-1 cases in the plane are precisely the time-optimal concatenation sequences in the codimension-0 three-dimensional case (see Sect. 7.3), and the optimal sequences  $BBBB$ ,  $BBSB$ , and  $BSBB$  for the codimension-2 case in the plane are the optimal sequences for the codimension-1 cases in  $\mathbb{R}^3$  (see Sect. 7.5) and the codimension-0 cases in  $\mathbb{R}^4$ . This is very much like the *unfolding of singularities* in the theory of differentiable mappings [108]. Thus, a *general classification of the concatenation sequences that optimal controlled trajectories for planar systems can have locally* in more degenerate cases based on Lie-theoretic conditions is not merely of intrinsic interest, but it also points to the structures of optimal solutions in higher dimensions. We shall return to this

topic in Chap. 7. In the next section, we shall analyze another classical optimal control problem in which the codimension of the Lie-bracket configuration becomes infinite, and indeed, optimal trajectories require an infinite number of switchings on a finite interval and thus are no longer piecewise continuous.

Examples of these correspondences abound not only for the time-optimal control problem, but in general. For example, in Sect. 6.2, we shall consider a three-dimensional optimal control problem for a mathematical model for tumor anti-angiogenesis [160] in which, because of the presence of optimal saturating singular controls, the solution is fully characterized by the concatenation sequences determined here for the codimension-2 scenario. Indeed, the optimal concatenation structures encountered for the time-optimal control problem in the plane that were analyzed in the last two sections consistently reappear in optimal solutions for general optimal control problems in increasing dimensions.

## 2.11 Chattering Arcs: The Fuller Problem

The Fuller problem has its origin in electronics, arising in communication across a nonlinear channel [34, 35, 94]. In this section, we give a solution to this problem, an innocent-looking problem whose optimal controlled trajectories are chattering arcs for which the controls switch infinitely often on an arbitrarily small interval as the switchings accumulate at the final time. In particular, optimal controls are no longer piecewise continuous, but lie in the class of Lebesgue measurable functions. The reason for this behavior lies in the presence of an optimal singular arc of order 2.

[Fuller] Given a point  $p \in \mathbb{R}^2$ , find a control (Lebesgue measurable function) with values in the interval  $[-1, 1]$  that steers  $p$  into the origin under the dynamics

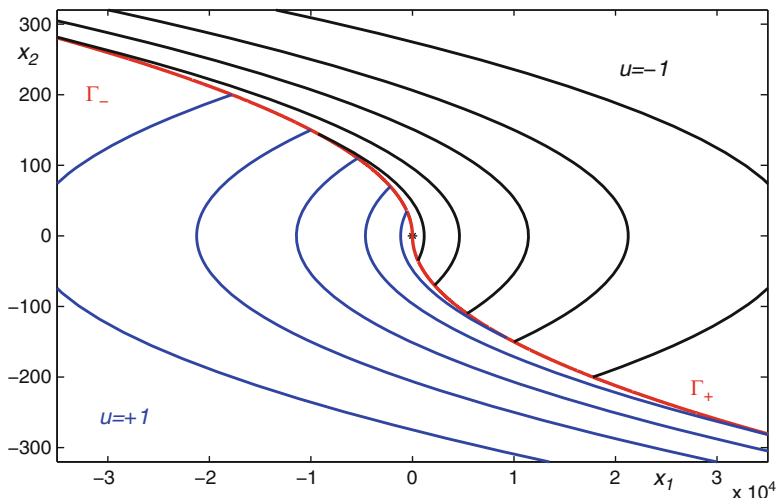
$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u,$$

and minimizes the objective

$$J(u) = \frac{1}{2} \int_0^T x_1^2(t) dt.$$

The time  $T$  of transfer is finite, but otherwise free. Since the problem is time-invariant, we can arbitrarily shift the interval of definition for the control, and for this problem it is more convenient to normalize the terminal time to be 0. We thus consider the controls and trajectories to be defined over intervals  $[-T, 0] \subset (-\infty, 0]$ .

**Theorem 2.11.1.** *Let  $\zeta = \sqrt{\frac{\sqrt{33}-1}{24}} = 0.4446236\dots$ , the unique positive root of the equation  $z^4 + \frac{1}{12}z^2 - \frac{1}{18} = 0$ , and define*



**Fig. 2.19** Optimal synthesis for the Fuller problem

$$\Gamma_+ = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = \zeta x_2^2, x_2 < 0\},$$

$$\Gamma_- = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -\zeta x_2^2, x_2 > 0\},$$

$$G_+ = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 < -\text{sgn}(x_2)\zeta x_2^2\},$$

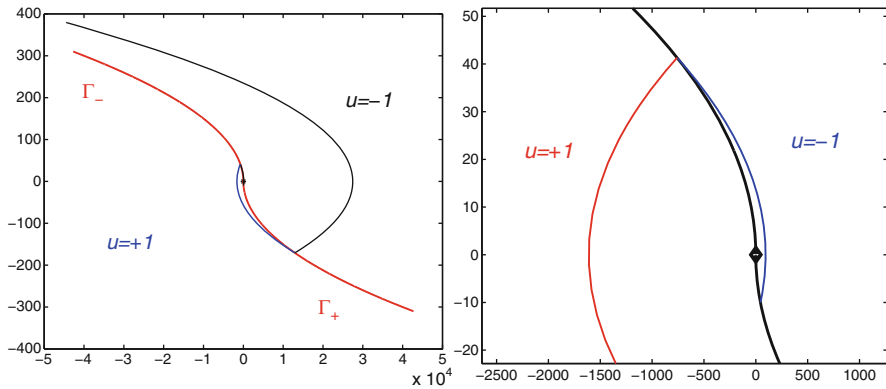
$$G_- = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > -\text{sgn}(x_2)\zeta x_2^2\}.$$

Then, the optimal control for the Fuller problem is given in feedback form as

$$u_*(x) = \begin{cases} +1 & \text{for } x \in G_+ \cup \Gamma_+, \\ -1 & \text{for } x \in G_- \cup \Gamma_-. \end{cases} \quad (2.58)$$

Corresponding trajectories cross the switching curves  $\Gamma_+$  and  $\Gamma_-$  transversally, changing from  $u = -1$  to  $u = +1$  at points on  $\Gamma_+$  and from  $u = +1$  to  $u = -1$  at points on  $\Gamma_-$ . These trajectories are chattering arcs with an infinite number of switchings that accumulate with a geometric progression at the final time  $T = 0$ .

Figures 2.18 and 2.19 depict the optimal synthesis for the Fuller problem. It looks very much like the synthesis for the double integrator, but with the significant difference that the switching curve  $\Gamma = \Gamma_+ \cup \{(0, 0)\} \cup \Gamma_-$  now is *not* a trajectory. Thus trajectories always cross  $\Gamma$  and cannot enter the origin along these curves.



**Fig. 2.20** An example of an optimal controlled trajectory (*left*) and a blowup near the final time (*right*)

### 2.11.1 The Fuller Problem as a Time-Optimal Control Problem in $\mathbb{R}^3$

The reason for the occurrence of the chattering controls is best understood if one embeds the Fuller problem into a time-optimal control problem of the form [NTOC] in  $\mathbb{R}^3$  by adding the objective as a third variable,  $\dot{x}_3 = \frac{1}{2}x_1^2$ , i.e., the drift vector field  $f$  and control vector field  $g$  are given by

$$f(x) = \begin{pmatrix} x_2 \\ 0 \\ \frac{1}{2}x_1^2 \end{pmatrix} \quad \text{and} \quad g(x) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

If one then considers the time-optimal control problem to the origin, the classical Fuller problem arises for initial conditions of the form  $p = (x_1^0, x_2^0, -J(x_1^0, x_2^0))$ , where  $J(x_1^0, x_2^0)$  is the optimal value for the Fuller problem with initial condition  $(x_1^0, x_2^0)$ . It will be seen that the solution to the Fuller problem is unique, and thus there exists exactly one control that steers  $p$  into the origin. Hence this control is the time-optimal one. In fact, the solutions to the Fuller problem are optimal abnormal extremals for this three-dimensional time-optimal control problem: the Hamiltonian for the Fuller problem is given by

$$H = \frac{1}{2}\lambda_0 x_1^2 + \lambda_1 x_2 + \lambda_2 u$$

with  $\lambda_0 \geq 0$ , while the Hamiltonian for the time-optimal control problem, where we change the notation for the multiplier to  $\psi$  in order to distinguish these two

formulations, is given by

$$H = \psi_0 + \psi_1 x_2 + \psi_2 u + \frac{1}{2} \psi_3 x_1^2.$$

We shall see below that extremals for the Fuller problem cannot be abnormal ( $\lambda_0 > 0$ ), and for the time-optimal control problem,  $\psi_3$  is a constant that cannot vanish if  $\psi_0 = 0$  with time-minimizing extremals corresponding to  $\psi_3 > 0$  and maximizing ones to  $\psi_3 < 0$ . Normalizing  $\psi_3 = 1$  and taking  $\psi_0 = 0$ , the conditions of the maximum principle for these two problems agree.

The Lie brackets of the vector fields  $f$  and  $g$  are easily computed as

$$[f, g](x) \equiv \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \quad [f, [f, g]](x) = \begin{pmatrix} 0 \\ 0 \\ x_1 \end{pmatrix}, \quad \text{and} \quad [g, [f, g]](x) \equiv 0.$$

Since  $[g, [f, g]]$  vanishes identically, so do the brackets  $[f, [g, [f, g]]]$  and  $[g, [g, [f, g]]]$ , and singular controls are of higher order. The other relevant fourth- and fifth-order brackets are

$$\text{ad}_f^3 g(x) = \begin{pmatrix} 0 \\ 0 \\ x_2 \end{pmatrix}, \quad \text{ad}_f^4 g(x) \equiv 0, \quad \text{and} \quad [g, \text{ad}_f^3 g](x) \equiv \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

In particular,

$$\langle \psi, [g, \text{ad}_f^3 g](x) \rangle = \psi_3 = 1 > 0,$$

and the Kelley condition for optimality of an order-2 singular arc is satisfied. The equation defining the singular control is

$$\Phi^{(4)}(t) = \psi \text{ad}_f^4 g(x) + u \psi [g, \text{ad}_f^3 g](x) = u$$

and thus the singular control is given by

$$u_{\text{sing}} \equiv 0.$$

The corresponding singular extremal  $\Gamma_F$  is therefore given by  $u \equiv 0$ ,  $x_1 \equiv x_2 \equiv 0$ , with multipliers  $\psi_0 = \psi_1 = \psi_2 = 0$  and  $\psi_3 \equiv 1$ . The classical Fuller problem can thus be interpreted as the problem of steering a point in  $\mathbb{R}^3$  time-optimally into an order-2 singular arc that satisfies the Kelley condition. By Proposition 2.8.5, the singular control cannot be concatenated with a constant bang control without violating the necessary conditions of the maximum principle. This can be accomplished only by means of a chattering control.

### 2.11.2 Elementary Properties of Extremals

We now construct an extremal synthesis for the Fuller problem following an argument of Kupka [143]. (The optimality of this synthesis will be verified in Sects. 5.1 and 5.2.3 by means of two completely different arguments.) Let  $(x, u)$  be an optimal controlled trajectory that transfers  $p$  into the origin and minimizes the integral  $\int_0^T x_1^2(t)dt$ . By Theorem 2.2.1, there exist a constant  $\lambda_0 \geq 0$  and an adjoint vector  $\lambda = (\lambda_1, \lambda_2)$  such that (i)  $(\lambda_0, \lambda_1, \lambda_2)$  do not vanish simultaneously, (ii)  $\dot{\lambda}_1 = -\lambda_0 x_1$ ,  $\dot{\lambda}_2 = -\lambda_1$ , and (iii) the control minimizes the Hamiltonian  $H = \frac{1}{2}\lambda_0 x_1^2 + \lambda_1 x_2 + \lambda_2 u$  over the interval  $[-1, 1]$  with the minimum value identically zero.

**Lemma 2.11.1.** *Extremals of optimal controlled trajectories are normal.*

*Proof.* Suppose  $\lambda_0 = 0$ . The switching function  $\Phi$  is given by the multiplier  $\lambda_2$ , and in this case  $\ddot{\lambda}_2 = 0$ . Hence the corresponding control  $u$  is bang-bang with at most one switching ending with either  $u = +1$  or  $u = -1$ . But this contradicts Proposition 2.8.5. For if we define a new control  $\check{u}$  by adding an interval  $[T, T + \varepsilon]$  with control  $\check{u}(t) \equiv 0$  on this interval, then the value of the objective does not change under this extension, and thus  $\check{u}$  is optimal as well. But the final segment with  $u = 0$  is a singular arc of order 2, and thus it cannot be concatenated optimally with a bang control. Contradiction.  $\square$

We henceforth normalize  $\lambda_0 = 1$ . Then the derivatives of the switching function  $\Phi = \lambda_2$  are given by

$$\dot{\Phi}(t) = -\lambda_1(t), \quad \ddot{\Phi}(t) = x_1(t), \quad \Phi^{(3)}(t) = x_2(t), \quad \Phi^{(4)}(t) = u(t),$$

and the minimum condition implies

$$u(t) = -\operatorname{sgn} \Phi(t).$$

In particular, the switching function is a solution to the nonsmooth differential equation  $\Phi^{(4)}(t) = -\operatorname{sgn} \Phi(t)$ . We start with some elementary properties of extremals.

**Lemma 2.11.2.** *Let  $u = \pm 1$ ; then the functions*

$$I_{1,\pm} = x_1 - \frac{1}{2}ux_2^2 \quad \text{and} \quad I_{2,\pm} = -\lambda_1 - ux_1x_2 + \frac{1}{3}x_2^3$$

*are first integrals for the extremals of the Fuller problem. That is, the functions  $I_{1,\pm}$  and  $I_{2,\pm}$  are constant along extremals for the controls  $u = \pm 1$ .*

*Proof.* This follows by direct differentiation from the system and adjoint equations

$$\dot{I}_{1,\pm} = \dot{x}_1 - ux_2\dot{x}_2 = x_2 - u^2x_2 = 0,$$

$$\dot{I}_{2,\pm} = -\dot{\lambda}_1 - u\dot{x}_1x_2 - ux_1\dot{x}_2 + x_2^2\dot{x}_2 = x_1 - ux_2^2 - u^2x_1 + x_2^2u = (1 - u^2)x_1 = 0.$$

□

**Lemma 2.11.3.** *Let  $\Gamma = ((x, u), \lambda)$  be an extremal defined over the interval  $[-T, 0]$ . If  $\tau < 0$  is a switching time, then  $\tau$  is an isolated zero of the switching function, and a bang-bang switch occurs at time  $\tau$ . This switch is from  $u = +1$  to  $u = -1$  if  $x_2(\tau) > 0$  and from  $u = -1$  to  $u = +1$  if  $x_2(\tau) < 0$ .*

*Proof.* Suppose  $\Phi(\tau) = \lambda_2(\tau) = 0$ . It is clear that a bang-bang switch occurs if  $\dot{\lambda}_2(\tau) = -\lambda_1(\tau)$  does not vanish. If  $\lambda_1(\tau) = 0$  as well, then the condition

$$0 = H(\tau) = \frac{1}{2}x_1^2(\tau) + \lambda_1(\tau)x_2(\tau) \quad (2.59)$$

implies that  $x_1(\tau) = 0$ , and thus, since the junction point is not the origin, we have  $x_2(\tau) \neq 0$ . But then  $\Phi(\tau) = \dot{\Phi}(\tau) = \ddot{\Phi}(\tau) = 0$  and

$$\Phi^{(3)}(\tau) = x_2(\tau) \neq 0.$$

Thus the switching function changes from negative to positive if  $x_2(\tau) > 0$  and from positive to negative if  $x_2(\tau) < 0$  and the corresponding bang-bang switch occurs. For the case  $\lambda_1(\tau) \neq 0$ , the same structure follows, since  $x_2(\tau)$  and  $\lambda_1(\tau)$  have opposite signs by (2.59). □

### 2.11.3 Symmetries of Extremals

The family of all extremals possesses two groups of symmetries, one continuous, the other discrete, which can be used very much to advantage in calculating the extremal synthesis. Without loss of generality, we define all extremals over the full interval  $(-\infty, 0]$  with the terminal time  $T$  normalized to be  $T = 0$ . Let  $\mathcal{G}_\alpha$  denote the multiplicative group of positive reals and define a 1-parameter group of scaling symmetries on  $(-\infty, 0] \times [-1, 1] \times \mathbb{R}^2 \times (\mathbb{R}^2)^*$  by

$$\mathcal{G}_\alpha : (t, u, x_1, x_2, \lambda_1, \lambda_2) \mapsto \left( \frac{t}{\alpha}, \alpha^0 u, \alpha^2 x_1, \alpha x_2, \alpha^3 \lambda_1, \alpha^4 \lambda_2 \right).$$

**Proposition 2.11.1.** *Given an extremal lift  $\Gamma = ((x, u), \lambda)$  for the Fuller problem and  $\alpha > 0$ , define  $\Gamma^\alpha = ((x^\alpha, u^\alpha), \lambda^\alpha)$  as the controlled trajectory  $(x^\alpha, u^\alpha)$  and corresponding adjoint vector  $\lambda^\alpha$  that are obtained under the action of the group  $\mathcal{G}_\alpha$  on the variables; that is, by*

$$u^\alpha(t) = u\left(\frac{t}{\alpha}\right), \quad x_1^\alpha(t) = \alpha^2 x_1\left(\frac{t}{\alpha}\right), \quad x_2^\alpha(t) = \alpha x_2\left(\frac{t}{\alpha}\right),$$

and

$$\lambda_1^\alpha(t) = \alpha^3 \lambda_1\left(\frac{t}{\alpha}\right), \quad \lambda_2^\alpha(t) = \alpha^4 \lambda_2\left(\frac{t}{\alpha}\right).$$

Then  $\Gamma^\alpha$  again is an extremal for the Fuller problem.

*Proof.* Consider the controlled trajectory  $(x, u)$  over the interval  $[-\bar{t}, 0]$  with initial condition at time  $-\bar{t}$  given by  $(\bar{x}_1, \bar{x}_2)$ . The rescaled control  $u^\alpha$ , restricted to  $[-\alpha\bar{t}, 0]$ , then steers  $(\bar{x}_1^\alpha, \bar{x}_2^\alpha) = (\alpha^2 \bar{x}_1, \alpha \bar{x}_2)$  into the origin with corresponding trajectory  $x^\alpha$ . A direct calculation verifies that the adjoint equation is invariant under this transformation as well,

$$\begin{aligned} \dot{\lambda}_1^\alpha(t) &= \alpha^3 \dot{\lambda}_1\left(\frac{t}{\alpha}\right) \frac{1}{\alpha} = -\alpha^2 x_1\left(\frac{t}{\alpha}\right) = -x_1^\alpha(t), \\ \dot{\lambda}_2^\alpha(t) &= \alpha^4 \dot{\lambda}_2\left(\frac{t}{\alpha}\right) \frac{1}{\alpha} = -\alpha^3 \lambda_1\left(\frac{t}{\alpha}\right) = -\lambda_1^\alpha(t), \end{aligned}$$

and also the Hamiltonian  $H$  remains unchanged:

$$\begin{aligned} H(\lambda^\alpha(t), x^\alpha(t), u^\alpha(t)) &= \frac{1}{2} x_1^\alpha(t)^2 + \lambda_1^\alpha(t) x_2^\alpha(t) + \lambda_2^\alpha(t) u^\alpha(t) \\ &= \frac{1}{2} \left[ \alpha^2 x_1\left(\frac{t}{\alpha}\right) \right]^2 + \alpha^3 \lambda_1\left(\frac{t}{\alpha}\right) \alpha x_2\left(\frac{t}{\alpha}\right) + \alpha^4 \lambda_2\left(\frac{t}{\alpha}\right) u\left(\frac{t}{\alpha}\right) \\ &= \alpha^4 H\left(\lambda\left(\frac{t}{\alpha}\right), x\left(\frac{t}{\alpha}\right), u\left(\frac{t}{\alpha}\right)\right) = 0. \end{aligned}$$

Furthermore, by construction, the minimum condition on the control carries over from the extremal lift  $\Gamma$ . Hence  $\Gamma^\alpha$  is an extremal as well.  $\square$

A second symmetry is given by reflecting controlled trajectories and their multipliers at the origin, in mathematical terms, by the action of the discrete group  $S_2$ . Let  $\mathcal{R}$  denote the reflection symmetry defined on  $(-\infty, 0] \times [-1, 1] \times \mathbb{R}^2 \times (\mathbb{R}^2)^*$  by

$$\mathcal{R} : (t, u, x_1, x_2, \lambda_1, \lambda_2) \mapsto (t, -u, -x_1, -x_2, -\lambda_1, -\lambda_2).$$

**Proposition 2.11.2.** *Given an extremal lift  $\Gamma = ((x, u), \lambda)$  for the Fuller problem, define  $\check{\Gamma} = ((\check{x}, \check{u}), \check{\lambda})$  as the controlled trajectory  $(\check{x}, \check{u})$  and corresponding adjoint vector  $\check{\lambda}$  that are obtained under the action of  $\mathcal{R}$ , that is, by*

$$\check{u}(t) = -u(t), \quad \check{x}_1(t) = -x_1(t), \quad \check{x}_2(t) = -x_2(t),$$



and

$$\check{\lambda}_1(t) = -\lambda_1(t), \quad \check{\lambda}_2(t) = -\lambda_2(t).$$

Then  $\check{\Gamma}$  again is an extremal for the Fuller problem.

*Proof.* It is clear that all the conditions of the maximum principle are invariant under this transformation.  $\square$

### 2.11.4 A Synthesis of Invariant Extremals

Whenever a mathematical problem exhibits symmetries, it is a good strategy to seek solutions that obey these symmetries. In fact, there is one extremal that is invariant under the action of all symmetries  $\mathcal{G}_\alpha$  for all  $\alpha > 0$  and  $\mathcal{R}$ , namely the trivial solution for  $u \equiv 0$  with  $x \equiv 0$  and  $\lambda \equiv 0$ . (The nontriviality condition is satisfied by  $\lambda_0 = 1$ .) In some sense, this is responsible for the special properties of trajectories that need to steer the system into the origin. But there also exists a specific value  $\alpha$  for which *all* extremals are invariant (as individual curves, not just as the whole family) under the actions of  $\mathcal{R}$  and  $\mathcal{G}_\alpha$ . These are the optimal controlled trajectories for the Fuller problem, and we now calculate this value.

Let  $\Gamma = ((x, u), \lambda)$  be an extremal for the Fuller problem and suppose  $t_0 < 0$  is a switching time where the control switches from  $u = +1$  to  $u = -1$ . Since the switchings are isolated, but must accumulate for  $T = 0$ , there exists a sequence  $\{t_n\}_{n \in \mathbb{Z}}$  of switching times that converges to 0 as  $n \rightarrow \infty$  and the control switches from  $u = +1$  to  $u = -1$  at even indices and from  $u = -1$  to  $u = +1$  at odd indices. Let  $\check{\Gamma}_\alpha = ((\check{x}^\alpha, \check{u}^\alpha), \check{\lambda}^\alpha)$  denote the image of the extremal  $\Gamma$  under the combined action  $\mathcal{A}_\alpha$  of the reflection  $\mathcal{R}$  and the group  $\mathcal{G}_\alpha$  for a fixed  $\alpha > 0$ , i.e., for all  $t \leq 0$ ,

$$\check{u}^\alpha(t) = -u\left(\frac{t}{\alpha}\right), \quad \check{x}_1^\alpha(t) = -\alpha^2 x_1\left(\frac{t}{\alpha}\right), \quad \check{x}_2^\alpha(t) = -\alpha x_2\left(\frac{t}{\alpha}\right),$$

and

$$\check{\lambda}_1^\alpha(t) = -\alpha^3 \lambda_1\left(\frac{t}{\alpha}\right), \quad \check{\lambda}_2^\alpha(t) = -\alpha^4 \lambda_2\left(\frac{t}{\alpha}\right).$$

By Propositions 2.11.1 and 2.11.2,  $\check{\Gamma}_\alpha = \mathcal{A}_\alpha(\Gamma)$  again is an extremal, but generally it will be different from  $\Gamma$ . If the extremals  $\Gamma$  and  $\check{\Gamma}_\alpha$  are the same, i.e., if  $\Gamma(t) = \check{\Gamma}_\alpha(t)$  for all  $t \leq 0$ , then the extremal is a *fixed point* under this transformation, and we say that it is invariant under this action. Note that if  $\Gamma$  is  $\mathcal{A}_\alpha$ -invariant, then it is also invariant under the action of any odd power  $\alpha^{2k+1}$  for all  $k \in \mathbb{Z}$ . But there always exists a smallest  $\alpha > 1$ , and this number will be called the generator.

**Proposition 2.11.3.** *let  $\Gamma = ((x, u), \lambda)$  be an extremal for the Fuller problem defined over the semi-infinite interval  $(-\infty, 0]$  with switching times  $\{t_n\}_{n \in \mathbb{Z}}$  and suppose the control switches from  $u = -1$  to  $u = +1$  for even indices. If the extremal  $\Gamma$  is invariant under the combined action  $\mathcal{A}_\alpha$  of the reflection  $\mathcal{R}$  and the group  $\mathcal{G}_\alpha$  with generator  $\alpha$ , i.e., if  $\Gamma(t) = \check{\Gamma}_\alpha(t)$  for all  $t \leq 0$ , then*

$$\alpha = \sqrt{\frac{1+2\zeta}{1-2\zeta}}, \quad \text{where} \quad \zeta = \sqrt{\frac{\sqrt{33}-1}{24}}.$$

The switching points lie on the curves

$$\Gamma_+ = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = \zeta x_2^2, x_2 < 0\}$$

and

$$\Gamma_- = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -\zeta x_2^2, x_2 > 0\},$$

and switchings are from  $X = f - g$  to  $Y = f + g$  at points on  $\Gamma_+$  and from  $Y$  to  $X$  at points on  $\Gamma_-$ .

*Proof.* The invariance condition and the choice of  $\alpha$  as the generator imply that the switching times  $t_i$  follow a geometric progression,  $t_{i-1} = \alpha t_i$ ,  $i \in \mathbb{Z}$ . Starting at the switching time  $t_0$  and integrating the control  $u = +1$  until the time  $t_1 = \frac{t_0}{\alpha}$ , using the first integral  $I_{1,+}$ , we obtain that

$$x_1(t_1) - \frac{1}{2}x_2^2(t_1) = x_1(t_0) - \frac{1}{2}x_2^2(t_0) \quad (2.60)$$

and thus

$$\frac{x_1(t_1)}{x_1(t_0)} - 1 = \frac{1}{2} \frac{x_2^2(t_1) - x_2^2(t_0)}{x_1(t_0)} = \frac{1}{2} \left( \frac{x_2^2(t_1)}{x_2^2(t_0)} - 1 \right) \frac{x_2^2(t_0)}{x_1(t_0)}.$$

It follows from the invariance of the trajectory under the action of  $\mathcal{G}_\alpha$  and  $\mathcal{R}$  that

$$x_1(t_0) = \check{x}_1^\alpha(t_0) = -\alpha^2 x_1\left(\frac{t_0}{\alpha}\right) = -\alpha^2 x_1(t_1)$$

and

$$x_2(t_0) = \check{x}_2^\alpha(t_0) = -\alpha x_2\left(\frac{t_0}{\alpha}\right) = -\alpha x_2(t_1). \quad (2.61)$$

In particular,  $x_i(t_1)$  and  $x_i(t_0)$  have opposite signs at consecutive switchings for both  $i = 1, 2$ . Hence

$$\frac{x_1(t_1)}{x_1(t_0)} = -\frac{1}{\alpha^2} \quad \text{and} \quad \frac{x_2(t_1)}{x_2(t_0)} = -\frac{1}{\alpha}. \quad (2.62)$$

But then we get for the  $XY$ -junction at time  $t_0$  that

$$-\frac{1}{\alpha^2} - 1 = \frac{1}{2} \left( \frac{1}{\alpha^2} - 1 \right) \frac{x_2^2(t_0)}{x_1(t_0)}$$

or equivalently,

$$x_1(t_0) = \frac{1}{2} \frac{\alpha^2 - 1}{\alpha^2 + 1} x_2^2(t_0).$$

Also, by Lemma 2.11.3,  $x_2(t_0)$  is negative.

Similarly, if we integrate  $u = -1$  between the switching times  $t_1$  and  $t_2$ , then we get from  $I_{1,-}$  that

$$x_1(t_2) + \frac{1}{2}x_2^2(t_2) = x_1(t_1) + \frac{1}{2}x_2^2(t_1),$$

which then leads to

$$\frac{x_1(t_2)}{x_1(t_1)} - 1 = \frac{1}{2} \frac{x_2^2(t_1) - x_2^2(t_2)}{x_1(t_1)} = \frac{1}{2} \left( 1 - \frac{x_2^2(t_2)}{x_2^2(t_1)} \right) \frac{x_2^2(t_1)}{x_1(t_1)}.$$

Analogous to Eq. (2.62), we also have that

$$\frac{x_1(t_2)}{x_1(t_1)} = -\frac{1}{\alpha^2} \quad \text{and} \quad \frac{x_2(t_2)}{x_2(t_1)} = -\frac{1}{\alpha},$$

and so it follows that

$$-\frac{1}{\alpha^2} - 1 = \frac{1}{2} \left( 1 - \frac{1}{\alpha^2} \right) \frac{x_2^2(t_1)}{x_1(t_1)}.$$

Hence

$$x_1(t_1) = -\frac{1}{2} \frac{\alpha^2 - 1}{\alpha^2 + 1} x_2^2(t_1),$$

and Lemma 2.11.3 now implies that  $x_2(t_0)$  is positive. Setting

$$\zeta = \frac{1}{2} \frac{\alpha^2 - 1}{\alpha^2 + 1} \in \left( 0, \frac{1}{2} \right), \quad (2.63)$$

the formulas for the switching curves follow.

It remains to calculate the value for  $\zeta$ . For the switching times  $t_0$  and  $t_1$  we have that

$$x_1(t_0) = \zeta x_2(t_0)^2 \quad \text{and} \quad x_1(t_1) = -\zeta x_2(t_1)^2$$

and thus, once more using the first integral  $I_{1,+}$ , we get from Eq. (2.60) that

$$I_{1,+}(t_1) = \left( \zeta - \frac{1}{2} \right) x_2^2(t_1) = \left( -\zeta - \frac{1}{2} \right) x_2^2(t_0)$$

or equivalently, by Eq. (2.62),

$$\left( \zeta - \frac{1}{2} \right) = \left( -\zeta - \frac{1}{2} \right) \alpha^2. \quad (2.64)$$

Similarly, using the first integral  $I_{2,+}$ , we also have that

$$-\lambda_1(t_1) - x_1(t_1)x_2(t_1) + \frac{1}{3}x_2^3(t_1) = -\lambda_1(t_0) - x_1(t_0)x_2(t_0) + \frac{1}{3}x_2^3(t_0). \quad (2.65)$$

It follows from the condition  $H(t) \equiv 0$  that at every switching time  $t_i$  we have that

$$0 = \frac{1}{2}x_1^2(t_i) + \lambda_1(t_i)x_2(t_i)$$

and thus

$$\lambda_1(t_i) = -\frac{1}{2} \frac{x_1^2(t_i)}{x_2(t_i)} = -\frac{1}{2} \zeta^2 x_2^3(t_i).$$

Hence Eq. (2.65) becomes

$$\left(\frac{1}{2}\zeta^2 - \zeta + \frac{1}{3}\right)x_2^3(t_1) = \left(\frac{1}{2}\zeta^2 + \zeta + \frac{1}{3}\right)x_2^3(t_0),$$

and thus, again by Eq. (2.62),

$$\left(\frac{1}{2}\zeta^2 - \zeta + \frac{1}{3}\right) = -\left(\frac{1}{2}\zeta^2 + \zeta + \frac{1}{3}\right)\alpha^3. \quad (2.66)$$

Solving Eqs. (2.64) and (2.66) for  $\alpha$  and equating the resulting expressions gives the following relation on  $\zeta$ :

$$\frac{\left(\frac{1}{2}\zeta^2 - \zeta + \frac{1}{3}\right)^2}{\left(\zeta - \frac{1}{2}\right)^3} = \frac{\left(\frac{1}{2}\zeta^2 + \zeta + \frac{1}{3}\right)^2}{\left(-\zeta - \frac{1}{2}\right)^3}. \quad (2.67)$$

This expressions simplifies to the equation

$$\zeta^4 + \frac{1}{12}\zeta^2 - \frac{1}{18} = 0,$$

which has a unique positive solution given by

$$\zeta = \sqrt{\frac{\sqrt{33}-1}{24}}.$$

The formula for  $\alpha$  follows from Eq. (2.63). □

These calculations prove that if there exist extremal controlled trajectories that are invariant under the combined action  $\mathcal{A}$  defined by the composition of the group actions  $\mathcal{R}$  and  $\mathcal{G}_\alpha$  for some  $\alpha$ , then the generator is given by

$$\alpha = \sqrt{\frac{1+2\zeta}{1-2\zeta}} = 4.1301599\dots, \quad (2.68)$$

and the trajectories are those corresponding to the synthesis defined in Theorem 2.11.1. It is not difficult to reverse these computations and show that this construction indeed gives rise to a family of  $\mathcal{A}_\alpha$ -invariant extremals.

**Proposition 2.11.4.** *The synthesis  $\mathcal{F}$  defined in Theorem 2.11.1 generates a family of  $\mathcal{A}_\alpha$ -invariant extremals.*

*Proof.* Let  $p > 0$  and consider the point  $\gamma(p) = (\zeta p^2, -p)$  on the switching curve  $\Gamma_+$ . We first calculate the total time  $T_p$  it takes for the controlled trajectory of the synthesis  $\mathcal{F}$  that starts at the point  $\gamma(p)$  to reach the origin. If we take  $t_0 = -T_p$  as initial time  $t_0$  for the trajectory, then the time of the next switching is  $t_1 = \frac{t_0}{\alpha}$ , and  $\frac{x_2(t_1)}{x_2(t_0)} = -\frac{1}{\alpha}$ . Since  $\dot{x}_2 = 1$  over  $[t_0, t_1]$ , we have that

$$x_2(t_1) - x_2(t_0) = t_1 - t_0 = \left(\frac{1}{\alpha} - 1\right)t_0$$

and thus, dividing by  $-x_2(t_0)$ ,

$$\left(1 - \frac{1}{\alpha}\right) \frac{t_0}{(-p)} = 1 - \frac{x_2(t_1)}{x_2(t_0)} = 1 + \frac{1}{\alpha},$$

which gives

$$T_p = -t_0 = \frac{1 + \frac{1}{\alpha}}{1 - \frac{1}{\alpha}} p = \frac{\alpha + 1}{\alpha - 1} p.$$

Given  $\gamma(p)$ , define a control  $u_p$  over the infinite interval  $(-\infty, 0)$  to have the switching times  $\{t_n\}_{n \in \mathbb{Z}}$  given by  $t_0 = \frac{1+\alpha}{1-\alpha} p < 0$  and  $t_i = \alpha^{-i} t_0$  with the controls alternating between  $+1$  and  $-1$  at the switching times and  $u_p \equiv +1$  on the interval  $(t_0, t_1)$ . Let  $x_p = (x_1, x_2)^T$  be the corresponding trajectory. This is the controlled trajectory generated by the synthesis  $\mathcal{F}$  through the point  $\gamma(p)$ . Define a solution  $\lambda_p = (\lambda_1, \lambda_2)$  of the corresponding adjoint equation by taking as initial conditions at time  $t_0$  the values

$$\lambda_1(t_0) = -\frac{1}{2} \frac{x_1^2(t_0)}{x_2(t_0)} = -\frac{1}{2} \zeta^2 p^3 \quad \text{and} \quad \lambda_2(t_0) = 0.$$

We claim that this defines an  $\mathcal{A}$ -invariant extremal  $\Gamma_p = ((x_p, u_p), \lambda_p)$ . This is fairly obvious by construction. Clearly, the control  $u_p$  is  $\mathcal{A}$ -invariant, and calculations invoking the first integral  $I_1$  analogous to those carried out in the proof of Lemma 2.11.3 verify that the corresponding trajectory  $x_p$  is invariant as well. We have taken care to choose the correct initial condition for the multiplier, and the  $\mathcal{A}$ -invariance of the adjoint vector can be verified using the other first integral  $I_2$ . Finally, the fact that the Hamiltonian is identically zero simply follows from the fact that

$$\begin{aligned}
H(t_0) &= \frac{1}{2}x_1^2(t_0) + \lambda_1(t_0)x_2(t_0) + \lambda_2(t_0)u(t_0) \\
&= -\frac{1}{2}(\zeta p^2)^2 - \frac{1}{2}\zeta^2 p^3(-p) + 0 = 0
\end{aligned}$$

and  $\frac{d}{dt}H(t)$  vanishes, since  $\lambda$  is a solution to the corresponding adjoint equation. Since every controlled trajectory generated by the synthesis  $\mathcal{F}$  is of this form, this proves the proposition.  $\square$

It is easy to define a “patch”  $\Xi_0$  of controlled trajectories that generates this synthesis. Simply take the value  $p = 1$  and consider the point  $(\zeta, -1) \in \Gamma_+$ . The first return of this trajectory to the curve  $\Gamma_+$  then is at

$$\bar{x}_2 = x_2(t_2) = \frac{1}{\alpha^2}x_2(t_0) = -\frac{1}{\alpha^2}.$$

Define the function  $t_0 : [0, \infty) \rightarrow (-\infty, 0]$  by

$$t_0(p) = \frac{1 + \alpha}{1 - \alpha}p$$

and let

$$D_0 = \left\{ (t, p) : \frac{1}{\alpha^2} < p \leq 1, t_0(p) \leq t < t_1(p) = \frac{t_0(p)}{\alpha} \right\}$$

be the domain for a parametrization of the controlled trajectories of the Fuller synthesis  $\mathcal{F}$ ,

$$\Xi_0 : D_0 \rightarrow \mathbb{R}^2 \setminus \{(0, 0)\}, \quad (t, p) \mapsto (x_1(t, p), x_2(t, p)).$$

Then the iterates  $\Xi_n$  under the action defined by  $\mathcal{A}$ ,

$$\Xi_n : D_0 \rightarrow \mathbb{R}^2 \setminus \{(0, 0)\}, \quad (t, p) \mapsto (-1)^n \begin{pmatrix} x_1^{\alpha^n}(t, p) \\ x_2^{\alpha^n}(t, p) \end{pmatrix},$$

for all  $n \in \mathbb{Z}$  cover the full state space, except for the origin.

We used invariance properties of the extremals to give a rather elegant and short construction of an extremal synthesis. By itself, however, this does not guarantee optimality. The optimality of this field will be verified in Sect. 5.2.3. In fact, there we shall give a rather elementary constructive argument that proves the optimality of this synthesis based on the parameterization of the patch  $\Xi_0$ .

It is also true that the extremals constructed here are the *only* extremals possible, but this argument is quite a bit more technical and involved (for example, see [34, 35]). Coupled with a standard result that guarantees the existence of optimal solutions for the Fuller problem, this indeed then proves the optimality of the synthesis constructed. But here we are interested rather in illustrating the use of invariance properties, a tremendously powerful tool in the analysis of nonlinear

systems with symmetries. Our presentation here is based on ideas and arguments of Kupka. While this problem with its solution given by chattering arcs was considered an aberration for a long time, in his paper [143], Kupka has shown that this is far from the truth and that chattering extremals indeed are a generic phenomenon, i.e., are in some very precise mathematical sense “typical” in higher state-space dimensions.

Another important point that is made with this problem is that optimal controls in general need not be piecewise continuous for even the simplest-looking real analytic system. It is easy to see that an arbitrary measurable control  $u$  can be the solution of a time-optimal control problem for a system of the form  $\dot{x} = f(x) + ug(x)$  with control set  $U = [-1, 1]$  and some sufficiently “weird” smooth vector fields  $f, g \in C^\infty$ . But whether optimal controls can be that general if the vector fields are real analytic, or whether they then do have some regularity properties, as might be expected, still is an open problem for which only partial results exist. While the structure of these optimal chattering controls still is rather simple, nevertheless these are not piecewise continuous, but only Lebesgue measurable controls. And this is the correct class of controls to consider in any optimal control problem, since it allows for a reasonable theory of existence of optimal solutions (e.g., [33]). The main aim of this chapter was to illustrate how the conditions of the maximum principle can be used to solve problems, and for this the class of piecewise continuous controls is mostly adequate. But in order to proceed with the deeper theory, even if we shall not concern ourselves with existence theory, we shall need to allow for Lebesgue measurable controls. We shall see next that even for linear systems this is indispensable.

## 2.12 Notes

Linear-quadratic optimal control is a classical design principle in automatic control and is at the heart of many actual control schemes including autopilots on commercial aircraft, process control in chemical engineering, and many other regulation processes. There exist many excellent engineering textbooks that are fully devoted to this subject and its extensions, both as deterministic systems and in a stochastic (noisy) environment. For this reason, we included only the most fundamental results on this topic. We highly recommend the classical text by Kwakernaak and Sivan [144] to the interested reader. We used the textbook by Knowles [139] as a source for the introductory one-dimensional examples that allow for explicit integrations of the solutions.

Time-optimal control for linear systems also is a classical topic treated in depth in many of the textbooks from the 1960s and 1970s such as those by Lee and Marcus [147] and Athans and Falb [25]. We shall take up this topic in some more detail next.

The necessary conditions for optimality of singular controls presented in Sects. 2.8.4 and 2.8.5 represent only the culmination of the classical research on this topic that was carried out in the 1960s, e.g., [31, 104, 107, 121, 122, 131, 132, 169, 173, 178, 201]. We shall prove these results in Chap. 4, but using very different computational methods. Also, the lecture notes by H.W. Knobloch [137] provide an alternative approach to

many of these conditions. A treatment of singular trajectories that proceeds beyond these classical developments and takes into account conjugate points is given by Bonnard and Kupka [48, 49], and for an in-depth analysis of singular trajectories, we highly recommend the monograph by Bonnard and Chyba [44]. Genericity properties of singular trajectories are developed in the work by Chitour, Jean, and Trélat [71–73].

There do not exist many textbooks that provide the differential-geometric framework that we employ in our treatment of optimal control. In fact, the early texts that give some of these foundations are in engineering, such as those by Isidori [120] and Nijmeijer and van der Schaft [176], but these texts focus on concepts from automatic control such as regulation and disturbance decoupling and do not address optimal control. The textbooks by Sontag [225] and Jurdjevic [126] address a more mathematical audience. While focused on the foundations of nonlinear systems theory (e.g., reachability and controllability, integral manifolds), these texts also include an introduction to optimal control problems, however largely motivated by linear-quadratic control problems.

The results that are included in the later sections of this chapter were for a long time only scattered in the research literature or some edited volumes such as [1, 6] and [4]. It is only more recently that some specialized monographs have been published that include these issues, such as those by Bonnard and Chyba [44], Boscain and Piccoli [51], and Bressan and Piccoli [56]. Among these, the book by Boscain and Piccoli is fully devoted to optimal control problems in the plane. We refer the reader to this text and Sussmann's original paper [236] for a complete analysis of generic systems. In his papers, Sussmann carries this analysis further, analyzing all cases of positive codimension for a nondegenerate dynamical system with smooth vector fields  $f$  and  $g$  in  $C^\infty(\Omega)$  [236] and arbitrary real analytic vector fields  $f$  and  $g$  in  $C^\omega(\Omega)$  [237]. In [238], it is then shown how these local results combine to provide a global solution to the problem in terms of a *regular synthesis*. These results are developed further by Boscain and Piccoli, who, more generally, analyze the time-optimal control problem and the structure of its optimal syntheses for systems on two-dimensional manifolds [51]. Much less is known in dimension three, and we shall pick up this topic in Chap. 7.

The Fuller problem is another classical optimal control problem. For quite some time, the structure of its solution was considered an aberration until I.A.K. Kupka showed that indeed this is a common phenomenon in higher dimensions [143]. As in the Fuller problem, it arises naturally if controlled trajectories need to follow or leave a locally optimal singular arc that is of order 2 and the singular controls take values in the interior of the control set. While this, in principle, is not a generic scenario, there are many interesting practical problems in which this happens. For example, in mathematical models for tumor anti-angiogenic treatments (see Sect. 6.2), there exists an optimal singular arc of order 1 that on addition of pharmacokinetic models for the drug action becomes of order 2, leading to optimal chattering connections [165]. Similarly, these phenomena arise in the control of autonomous underwater robots [74, 75]. The most comprehensive treatment of chattering arcs so far is given in the monograph by Zelikin and Borisov [262].



Geometric Optimal Control  
Theory, Methods and Examples  
Schättler, H.; Ledzewicz, U.  
2012, XX, 640 p., Hardcover  
ISBN: 978-1-4614-3833-5