Technische Universität München
Zentrum Mathematik
Wissenschaftliches Rechnen

Local methods for global and stochastic problems in optimal control

Michael Christoph Kratzer

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Tim N. Hoffmann
Prüfer der Dissertation:	1.	Prof. Dr. Oliver Junge
	2.	Prof. Dr. Sina Ober-Blöbaum University of Oxford, UK (schriftliche Beurteilung)
	3.	Prof. Roberto Ferretti Università Roma Tre, Italien

Die Dissertation wurde am 11.07.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 31.10.2016 angenommen.

Abstract

Global and stochastic problems in optimal control usually require the computation of the value function on the entire state space, respectively to ensure that the global optimum is found and because Brownian motion is unbounded. In high dimensions, this is impractical as the computational effort grows exponentially. In this thesis, local methods which approximately solve global resp. stochastic problems and only require the solution of ordinary differential equations are developed.

Zusammenfassung

Globale und stochastische Probleme in der Optimalsteuerung erfordern in der Regel die Berechnung der Wertefunktion auf dem gesamten Zustandsraum, um sicherzustellen, dass das globale Optimum gefunden wird, bzw. da die brownsche Bewegung unbeschränkt ist. In hohen Dimensionen ist dies auf Grund des exponentiell wachsenden Rechenaufwands nicht durchführbar. In dieser Arbeit werden zur näherungsweisen Lösung von globalen und stochastischen Problemen lokale Methoden entwickelt, die lediglich die Lösung von gewöhnlichen Differentialgleichungen erfordern.

Acknowledgements

First and foremost, I would like to thank Oliver Junge, my supervisor, for his continuous guidance, support and encouragement over many years.

I am grateful to Andreas Bittracher, Andreas Denner, Daniel Karrasch and Alex Schreiber for interesting discussions and helpful comments on this thesis.

The members of the M3 research unit at TU München have created a friendly and constructive atmosphere and it has been a pleasure to work with them.

I would like to thank Studienstiftung des deutschen Volkes for financial support and miscellaneous opportunities.

Last but not least I am indebted to my parents for their invaluable support. Without them, this work would not have been possible.

Contents

Notation

We will frequently suppress arguments of functions. This includes in particular the time t in ODEs. Linearizing an ODE around a point z or a trajectory $z(t)$ usually involves evaluating a lot of derivatives at z resp. $z(t)$ and in such situations we often omit z as an argument of those derivatives.

We also follow the conventions of probability theory in omitting the dependence of random variables $X = X(\omega)$ on the outcome $\omega \in \Omega$ and abbreviating the notation of sets of outcomes, e.g. $\{X > 0\} := \{\omega \in \Omega : X(\omega) > 0\}$.

To avoid ambiguities we introduce some further conventions:

When a function might be confused with its evaluations, we use bold letters for the function itself. E.g. $c(x, u, t) := c\,(x(t), u(t), t)$, but $L(\mathbf{x}, \mathbf{u}) = \int_0^T c(x, u, t)$ is a functional which takes the whole functions $\mathbf{x} = x(\cdot)$ and $\mathbf{u} = u(\cdot)$ as its arguments.

When factors in a multiplication are lengthy or might be confused with arguments, we separate them with $\cdot$ ($\backslash cdot$). E.g. $Df(x, u)$ is the derivative of f taken at the point (x, u) but $Df \cdot (x, u)$ is the derivative taken at a point given by the context and multiplied by (x, u).

The derivative of functions between Banach spaces alway means the Fréchet derivative, i.e. we say that $g : U \to W$, with V, W Banach spaces and an open set $U \subseteq V$, has the (Fréchet) derivative $Dg(x) : V \to W$ at $x \in U$ if $Dg(x)$ is bounded and

$$\lim_{h \to 0} \frac{\|g(x + h) - g(x) - Dg(x) \cdot h\|_W}{\|h\|_V} = 0.$$

Finally, the difference between a derivative and a gradient is merely typographical in this thesis: for a function g with several arguments, we understand by ∇g the derivative Dg written as a column instead of a row vector in order to save horizontal space.

Introduction

The optimal control problem

By a control system we understand a continuous[1] dynamical system on a state space $\mathcal{X} = \mathbb{R}^n$, which can be influenced by a control chosen from a space $\mathcal{U} = \mathbb{R}^m$ such that the dynamics of the state $x \in \mathcal{X}$ are given by the ODE

$$\dot{x} = f(x, u, t). \tag{0.1}$$

A pair of functions $(\mathbf{x}, \mathbf{u})$ fulfilling (0.1) is called a *trajectory* of the control system.

In optimal control, the objective is to choose the control function $\mathbf{u}$ such that a given cost functional $L(\mathbf{x}, \mathbf{u})$ of the trajectory is minimized.[2]

In this thesis, we will consider two types of optimal control problems (OCPs): One with finite and the other with infinite time horizon. In this introduction, we consider only the first type, which is set on a finite time interval $[t_0, T]$, has a given initial condition $x(t_0) = x_0$ and a cost functional

$$L(\mathbf{x}, \mathbf{u}) := \int_{t_0}^{T} c(x, u, t)\,\mathrm{d}t + \varphi(x(T)).$$

As $\mathbf{x}$ is governed by (0.1), it is completely determined by $\mathbf{u}$ and the initial condition and hence one can formulate the OCP as

$$\min_{\mathbf{u}} L(\mathbf{u}) := L(\mathbf{x}(\mathbf{u}), \mathbf{u}). \tag{0.2}$$

Global and local methods

Solution methods for optimal control problems can be divided into global and local approaches.

Global methods

Bellman introduced the solution of optimal control problems via *dynamic programming*, in which one solves (0.2) for *all* possible initial conditions and times. Hence, one aims to find the value function

$$V(x_0, t_0) := \inf_{\mathbf{u}} \left\{ \int_{t_0}^{T} c(x(t), u(t), t)\,\mathrm{d}t + \varphi(x(T)) : \dot{x} = f(x, u, t), x(t_0) = x_0 \right\}.$$

[1] Discrete systems can be treated in a very similar way but are not a subject of this thesis.

[2] Maximization can be achieved analogously or by simply changing the sign.

The crucial insight is that the OCP can be split into two minimization problems as, for $\tau > 0$,

$$V(x_0, t_0) = \inf_{\mathbf{u}} \left\{ \int_{t_0}^{t_0+\tau} c(x, u, t)\,\mathrm{d}t + V(x_0(t_0+\tau), t_0+\tau) \right\}.$$

In the limit $\tau \to 0$, one obtains the Hamilton-Jacobi-Bellman (HJB) equation ([Bel54])

$$V_t(x,t) + \inf_u \left(f(x,u,t) \cdot V_x(x,t) + c(x,u,t)\right) = 0,$$

which can be solved backwards in time from the terminal condition

$$V(\cdot, T) = \varphi.$$

The HJB approach can also be extended to cover stochastic optimal control problems, e.g. with

$$\mathrm{d}x = f(x,u,t)\,\mathrm{d}t + \Sigma\Sigma^\top\,\mathrm{d}W,$$

where W is an n-dimensional Brownian motion and $\Sigma\Sigma^\top \in \mathbb{R}^{n\times n}$ a symmetric positive (but not necessarily definite) matrix. The goal is to minimizes the expected value $\mathbb{E}[L]$.

The HJB equation for this stochastic problem contains an additional diffusion term and reads

$$V_t(x,t) + \inf_u \left(f(x,u,t) \cdot V_x x(x,t) + c(x,u,t)\right) + \frac{1}{2}\operatorname{tr}(\Sigma\Sigma^\top V_{xx}) = 0.$$

Methods which directly solve the HJB equation are *global methods* as one considers all initial conditions and (implicitly) all possible controls, and is guaranteed to find the globally optimal solution.

For low-dimensional state spaces, efficient solution methods exist ([FF94], [GJ05], [Grü97], [JO04], [JS15]). In high-dimensional problems however, these methods suffer from the *curse of dimension*, the exponential growth of computational complexity with the dimension of the state space. This issue is exacerbated by the fact that value functions are often non-smooth. Consequently, specialized approaches for high dimensions, e.g. sparse grids ([BG04]), fail to be fully effective.

Local methods

An alternative to the HJB equation is the use of *local methods* in which one computes only a trajectory *in* the state space instead of the value function *on* the state space. Thereby one avoids the exponential scaling in the dimension.

Examples are gradient descent methods or approaches based on Pontryagin's minimum principle ([Pon87]) where one seeks solutions to a system of ODEs (cf. Theorem 2.1) which are a necessary condition for a minimum.

The downsides of local methods include that they find only local minima without a guarantee of global optimality and that they are not applicable to stochastic problems.

Contribution and structure of the thesis

In this thesis, we develop two methods which are local, i.e. which only solve ODEs instead of computing the entire value function or any other object of comparable complexity. Nevertheless, they still manage to approximately solve, respectively, a global and stochastic optimal control problem.

The main body of this thesis can be divided into three parts:

In the first part (chapters 1 and 2) we present known results about optimal control from a functional analytic point of view. We consider the OCP as an abstract constrained optimization problem and provide some general results for constrained problems in Chapter 1. These results are used in Chapter 2 to derive the necessary conditions known as Pontryagin's minimum principle. In addition, we show the existence of minimizers by treating the OCP as a problem in the calculus of variations with a cost function taking values in $\mathbb{R} \cup \{+\infty\}$.

In the second part (chapters 3 to 5) we develop a method for finding global optima. In Chapter 3, we provide a brief review of bifurcation theory, which is used in Chapter 4 to determine the global structure of the set of all local extrema for all initial conditions. We also show that bifurcations in this set are responsible for the loss of smoothness in the value function. In Chapter 5, we finally present an algorithm for systematically finding local minima and establish conditions under which this algorithm finds the global minimum.

In the third part (chapters 6 and 7), we treat the stochastic optimal control problem. Here we face the inherent difficulty that Brownian motion is unbounded and hence the expected value necessarily depends on the entire domain. To circumvent this, we do not minimize the expected value but a quantile, i.e. the value under which the cost stays with a given probability, say 95%.

Our approach to this problem is an extension of the First Order Reliability Method (FORM). It was originally used in structural engineering ([HL74, Rac76]) to approximately compute (but not optimize) the probability whether some function will exceed a given value. In Chapter 6, we review the FORM and show how its extension to optimization can be interpreted as a game with the antagonist representing the random influence. In Chapter 7, we apply the underlying idea to the dynamic setting of the stochastic optimal control problem. We obtain a linear approximation for the quantile optimization problem and provide error bounds.

The methods developed in this thesis are primarily conceptual: they merely require that certain equations be solved without specifying how to actually do this. We have deliberately omitted implementational details from the main body of the thesis. In fact, the best numerical method may depend on the concrete optimal control problem.

For convenience, we give in Appendix A a description of our prototype implementation. Newton's method for solving nonlinear equations is an important component of this implementation. We present in Appendix B results which show that certain step size strategies are particularly good at detecting singularities which indicate that the equation to be solved may actually not have any solutions.

1. Constrained optimization

Consider a control system

$$\dot{x} = f(x, u, t),\ x \in \mathcal{X},\ u \in \mathcal{U}$$

on some (possibly infinite) interval $[0, T)$ where w.l.o.g. we have set $t_0 := 0$ the *state space* $\mathcal{X}$ and the *control space* $\mathcal{U}$ are for now arbitrary Banach spaces. As in the introduction, we assume that we can freely choose the *control function* $u : [0, T) \to \mathcal{U}$ and wish to do so in a manner that minimizes some *cost functional*

$$L(\mathbf{x}, \mathbf{u}),$$

which depends on the *trajectory* consisting of $\mathbf{x} = x(\cdot)$ and $\mathbf{u} = u(\cdot)$. The initial condition $x(0) = x_0$ is given. $\mathbf{x}$ and $\mathbf{u}$ are elements of suitable Banach spaces $\mathbf{X}$ and $\mathbf{U}$ of functions $[0, T) \to \mathcal{X}$ resp. $[0, T) \to \mathcal{U}$.

As mentioned in the introduction, $\mathbf{x}$ is determined by $\mathbf{u}$ and x_0, and so one could attempt the minimization

$$\min_{\mathbf{u}} L(\mathbf{u}) := L(\mathbf{x}(\mathbf{u}), \mathbf{u}) \tag{0.2}$$

However, for $t_2 > t_1$, $x(t_2)$ depends on $u(t_1)$ in a rather complicated way and it is easier to consider instead the equivalent *constrained optimization problem*

$$\begin{aligned} &\min_{(\mathbf{x},\mathbf{u})} L(\mathbf{x}, \mathbf{u}) \\ &\text{s.t. } \dot{\mathbf{x}} - f(\mathbf{x}, \mathbf{u}) \equiv 0, \end{aligned} \tag{1.1}$$

in which the constraint relates $\mathbf{x}$ and $\mathbf{u}$ only locally in time.

This can also be written as

$$\begin{aligned} &\min_{(\mathbf{x},\mathbf{u})} L(\mathbf{x}, \mathbf{u}) \\ &\text{s.t. } (\mathbf{x}, \mathbf{u}) \in \mathcal{A} := \{(\mathbf{x}, \mathbf{u}) : g(\mathbf{x}, \mathbf{u}) = 0\}, \end{aligned} \tag{1.2}$$

or, with $\mathcal{Z} := \mathbf{X} \times \mathbf{U}$ and $z = (\mathbf{x}, \mathbf{u})$, as

$$\begin{aligned} &\min_{z \in \mathcal{Z}} L(z) \\ &\text{s.t. } z \in \mathcal{A} = \{z : g(z) = 0\}, \end{aligned} \tag{1.3}$$

where $\mathcal{Y}$ is a Banach space and $g : \mathcal{Z} \to \mathcal{Y}$ a suitable function that will be given explicitly in the next chapter.

In this chapter, we will develop some results for the more general problems (1.2) resp. (1.3) before applying them to optimal control problems in Chapter 2.

The *tangent space* $T_{\mathcal{A}}(z^*) \subseteq \mathcal{Z}$ of $\mathcal{A}$ at $z^* \in \mathcal{A}$ is defined as

$$T_{\mathcal{A}}(z^*) := \{z \in \mathcal{Z} : \exists \phi \in \mathcal{C}^1([-1,1], \mathcal{A}) : \phi(0) = z^*, \phi'(0) = z\}$$

For a constrained optimization problem of the form (1.3), we have the geometric extremality condition $DL(z^*) \perp T_{\mathcal{A}}(z^*)$. , i.e. [1]

$$DL(z^*) \in T_{\mathcal{A}}(z^*)^{\perp}.$$

At generic points, the tangent space $T_{\mathcal{A}}$ is equal to the linearized tangent space[2]

$$\{z \in \mathcal{Z} : Dg(z^*) \cdot z = 0\}. \tag{1.4}$$

For an operator $A : \mathcal{Z} \to \mathcal{Y}$ and its adjoint $A^* : \mathcal{Y}^* \to \mathcal{Z}^*$, defined by $\langle Az, y\rangle = \langle z, A^*y\rangle$ for all $z \in \mathcal{Z}$ and $y \in \mathcal{Y}^*$, we have the Fredholm alternative $\mathcal{N}(A)^{\perp} = \mathcal{R}(A^*)$.

Assuming (1.4) we conclude that an extremum z^* necessarily fulfills

$$\exists \lambda \in \mathcal{Y}^* : Dg^*(z^*) \cdot \lambda + DL(z^*) = 0. \tag{1.5}$$

The arguments sketched above can be made rigorous to give the following theorem:

Theorem 1.1 ([Zei95, Chapter 4.14]). *Let $L : \mathcal{Z} \to \mathbb{R}$ and $g : \mathcal{Z} \to \mathcal{Y}$ be continuously (with respect to the operator norm) Fréchet-differentiable on an open neighborhood of $z \in \mathcal{Z}$, where $\mathcal{Y}$ is a real Banach space.*

If z is a solution of (1.3) *and $Dg(z) : \mathcal{Z} \to \mathcal{Y}$ is surjective, then there exists a functional $\lambda \in \mathcal{Y}^*$ such that* (1.5) *holds true.*

Consequently, returning to $\mathbf{x}$ and $\mathbf{u}$, we say that

Definition 1.2. $(\mathbf{x}^*, \mathbf{u}^*)$ *is a* critical point *of* (1.2) *if and only if there exists a λ such that*

$$F(\mathbf{x}^*, \mathbf{u}^*, \lambda) := \begin{pmatrix} \nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{u}^*) + D_{\mathbf{x}} g^*(\mathbf{x}^*, \mathbf{u}^*) \cdot \lambda \\ \nabla_{\mathbf{u}} L(\mathbf{x}^*, \mathbf{u}^*) + D_{\mathbf{u}} g^*(\mathbf{x}^*, \mathbf{u}^*) \cdot \lambda \\ g(\mathbf{x}^*, \mathbf{u}^*) \end{pmatrix} = 0. \tag{1.6}$$

Introducing

$$H(\mathbf{x}, \mathbf{u}, \lambda) := L(\mathbf{x}, \mathbf{u}) + \langle \lambda, g(\mathbf{x}, \mathbf{u})\rangle$$

allows us to write F more compactly as

$$F(\mathbf{x}^*, \mathbf{u}^*, \lambda) := \begin{pmatrix} H_x(\mathbf{x}^*, \mathbf{u}^*, \lambda) \\ H_u(\mathbf{x}^*, \mathbf{u}^*, \lambda) \\ g(\mathbf{x}^*, \mathbf{u}^*) \end{pmatrix}.$$

[1] Note that we do not in general deal with Hilbert spaces, so for $\Omega \subseteq \mathcal{Z}$, $\Omega^{\perp} \subseteq \mathcal{Z}^*$ is not the orthogonal complement but the annihilator $\Omega^{\perp} := \{v \in \mathcal{Z}^* : \Omega \subseteq \mathcal{N}(v)\}$.

[2] In optimization, a condition ensuring this is called a *constraint qualification*. In Theorem 1.1, the constraint qualification is that Dg is surjective.

Remark 1.3. *As the Lagrange multiplier λ fulfills* (1.5)*, which involves the adjoint operator Dg^*, it is also called the* adjoint variable *of the optimization problem.*

1.1. Interpretation of H

The results in this section will allow us to relate H to the original problem (0.2) from a different viewpoint.

For suitable λ, $H(\cdot, \lambda)$ can, at least locally, be understood as an extension of L to inadmissible points. Let $P : \mathcal{Z} \to \mathcal{A}$ be a (possibly nonlinear) projection onto the admissible set $\mathcal{A}$ which is differentiable on $\mathcal{A}$. We can turn the constrained problem

$$\begin{aligned} &\min_{z \in \mathcal{Z}} L(z) \\ &\text{s.t. } z \in \mathcal{A} = \{z : g(z) = 0\} \end{aligned} \tag{1.3}$$

into an unconstrained one by applying P before evaluating L. To this end, we define

$$\widehat{H}(z, P) := L(P(z)) = L(z + (P(z) - z)),$$

where $P(z) - z$ is the correction required to move z into $\mathcal{A}$. With this definition, $\widehat{H}(\mathcal{Z}, P) = L(\mathcal{A})$ and so

$$\min_{z \in \mathcal{Z}} \widehat{H}(z, P) = \min_{z \in \mathcal{A}} L(z).$$

The following holds for any problem of the type (1.3), which is slightly more general than (1.2) because no decomposition of $\mathcal{Z}$ is given a priori. Hence we will consider general decompositions of $\mathcal{Z}$ before returning to $\mathcal{Z} = \mathbf{X} \times \mathbf{U}$ as a special case.

Now we linearize the problem at a point z_0. $DP = DP(z_0)$ is a linear projection with $\mathcal{R}(DP) = \mathcal{N}(Id - DP) = T_{\mathcal{A}} = \mathcal{N}(Dg)$ and is uniquely determined by $\mathcal{Z}_1 := \mathcal{N}(DP) = \mathcal{R}(Id - DP)$, where $\mathcal{Z}_1 \subseteq \mathcal{Z}$ is a subspace complementing $\mathcal{N}(Dg)$ (in the usual sense that $\mathcal{Z} = \mathcal{Z}_1 \oplus \mathcal{N}(Dg)$).

There is a unique pseudoinverse $Dg^+ : \mathcal{R}(Dg) \to \mathcal{Z}_1$ such that $Dg^+ Dg = Id - DP$ and hence

$$D\widehat{H} = DL + DL(DP - Id) = DL - DL \cdot Dg^+ Dg.$$

With $\lambda := -DL \cdot Dg^+$ we find that locally $\widehat{H}$ coincides with H to first order, i.e.

$$\widehat{H} \doteq L + \langle \lambda, g \rangle = H.$$

Let $\mathcal{Z} = \mathcal{Z}_1 \oplus \mathcal{Z}_2$ be another decomposition and accordingly $z = z_1 + z_2$. $\mathcal{Z}_2$ and $T_{\mathcal{A}}$ are both complements of $\mathcal{Z}_1 = \mathcal{N}(DP)$ and so the projection $DP : \mathcal{Z}_2 \to T_{\mathcal{A}}$ of $\mathcal{Z}_2$ onto $T_{\mathcal{A}}$ along $\mathcal{Z}_1$ is a bijection: It can be seen that z_2 parametrizes $T_{\mathcal{A}}$ as $DP \cdot (z_1 + z_2) = DPz_2 = -Dg^+ Dg \cdot z_2 + z_2 =: \tilde{z}_1(z_2) + z_2$ depends only on z_2. In this situation, the total

derivative of L with respect to z_2 is given by the partial derivative of H:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z_2}L &:= \frac{\mathrm{d}}{\mathrm{d}z_2}L\left(\tilde{z}_1(z_2)+z_2\right)\Big|_{z_1+z_2=z_0} = DL\cdot DP|_{\mathcal{Z}_2} \\ &= \frac{\mathrm{d}}{\mathrm{d}z_2}\widehat{H}(z_1+z_2)\Big|_{z_1+z_2=z_0} =: \frac{\partial}{\partial z_2}\widehat{H} = \frac{\partial}{\partial z_2}H\end{aligned} \tag{1.7}$$

Remark 1.4. *If z_0 is a critical point of (1.2), then λ is independent of P.*

Proof. Let P_1 and P_2 be two projections onto $\mathcal{A}$ leading to pseudoinverses Dg^+ resp. $Dg^\dagger$ and λ_1 resp. λ_2 as described above. We have $\mathcal{R}(Dg^+ - Dg^\dagger) \subseteq \mathcal{N}(Dg) \subseteq \mathcal{N}(DL)$, as z_0 is a critical point, and so $\lambda_1 - \lambda_2 = -DL(Dg^+ - Dg^\dagger) = 0$. □

In optimal control problems, we can use (1.7) to return to the original viewpoint of regarding $\mathbf{x}$ as a function of $\mathbf{u}$ and $L(\mathbf{u}) := L(\mathbf{x}(\mathbf{u}), \mathbf{u})$ as a function of $\mathbf{u}$ only, and obtain the total derivative $\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}}L = \frac{\mathrm{d}}{\mathrm{d}\mathbf{u}}L(\mathbf{x}(\mathbf{u}), \mathbf{u})$.

We always assume that g_x is invertible and that there is a unique solution $\mathbf{x}(\mathbf{u})$ of $g(\mathbf{x}(\mathbf{u}), \mathbf{u}) = 0$ for any $\mathbf{u} \in \mathbf{U}$ (which we will verify for optimal control problems in the next chapter).

With the decomposition of $\mathcal{Z} = \mathbf{X} \times \mathbf{U}$ into $\mathcal{Z}_1 = \mathbf{X} \times \{0\}$ and $\mathcal{Z}_2 = \{0\} \times \mathbf{U}$, we have $Dg^+ = \begin{pmatrix} g_x^{-1} \\ 0 \end{pmatrix}$ and so

$$\lambda = -\frac{\partial}{\partial \mathbf{x}}L\left(\frac{\partial}{\partial \mathbf{x}}g\right)^{-1} \text{ and} \tag{1.8}$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}}L = \frac{\partial}{\partial \mathbf{u}}H. \tag{1.9}$$

For this decomposition, $\widehat{H}$ can be understood as ignoring whatever value for $\mathbf{x}$ it is given and replacing it with $\mathbf{x}(\mathbf{u})$, obtained from the projection onto $\mathcal{A}$, before passing it on to L.

Relation to original definition

Summarizing some results from above, we have the following Theorem:

Theorem 1.5. *Consider a point $z_0 \in \mathcal{Z}$ at which all functions are evaluated. Let $H = L + \langle \lambda, g \rangle$ with $\lambda \in \mathcal{R}(Dg^*)$, $\mathcal{Z} = \mathcal{Z}_1 \oplus \mathcal{N}(Dg)$ and DP a projection onto $\mathcal{N}(Dg)$ with $\mathcal{N}(DP) = \mathcal{Z}_1$. Then the following are equivalent:*

(i) $DH = DL \cdot DP$,

(ii) $\frac{\partial}{\partial z_1}H = DH|_{\mathcal{Z}_1} = 0$,

(iii) $\lambda = -DL \cdot Dg^+$, *where* $Dg^+ : \mathcal{R}(Dg) \to \mathcal{Z}_1$ *is a pseudoinverse of* Dg.

Proof. $(i) \implies (ii)$: Let $dz_1 \in \mathcal{Z}_1$. Then $DHdz_1 = DL \cdot DPdz_1 = DL \cdot 0 = 0$.
$(ii) \implies (iii)$: Let $dz_1 \in \mathcal{Z}_1$. Then $0 = DHdz_1 = DLdz_1 + \langle \lambda, Dg \cdot dz_1 \rangle$ and $Dg^+Dg \cdot dz_1 = dz_1$. Hence $0 = (DL \cdot Dg^+ + \lambda \cdot DgDg^+)Dg \cdot dz_1 = (DL \cdot Dg^+ + \lambda)Dg \cdot dz_1$. Since this holds for any $dz_1 \in \mathcal{Z}_1$ and $Dg\mathcal{Z}_1 = \mathcal{R}(Dg)$, it follows that $\lambda = -DL \cdot Dg^+$.
$(iii) \implies (i)$: $DH = DL + \lambda \cdot Dg = DL(Id - Dg^+Dg) = DL \cdot DP$ □

This can be used to relate critical points of (1.2) and (0.2).

Lemma 1.6. *Let* $\mathbf{u}$ *be a critical point of* (0.2). *Set* $\mathbf{x} := \mathbf{x}(\mathbf{u})$ *and* $\lambda := -L_x(\mathbf{x}, \mathbf{u})(g_x(\mathbf{x}, \mathbf{u}))^{-1}$. *Then* $F(\mathbf{x}, \mathbf{u}, \lambda) = 0$, *i.e.* $(\mathbf{x}, \mathbf{u})$ *is a critical point of* (1.2).

Proof. We use again $\mathcal{Z}_1 = \mathbf{X} \times \{0\}$ and $\mathcal{Z}_2 = \{0\} \times \mathbf{U}$. (iii) of Theorem 1.5 is fulfilled by definition, so we have (ii), i.e. $H_x = 0$. $H_u = 0$ follows from (1.9) and $g = 0$ by the definition of $\mathbf{x}$. □

Lemma 1.7. *Let* $(\mathbf{x}, \mathbf{u})$ *be a critical point of* (1.2). *Then* $\mathbf{u}$ *is a critical point of* (0.2).

Proof. There is a λ such that $F(\mathbf{x}, \mathbf{u}, \lambda) = 0$, i.e. $H_x = 0$, $H_u = 0$ and $\mathbf{x} = \mathbf{x}(\mathbf{u})$ as $g = 0$. We have $\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} g(\mathbf{x}(\mathbf{u}), \mathbf{u}) = 0$ and so

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} L = \frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} H(\mathbf{x}(\mathbf{u}, \mathbf{u}, \lambda)) = H_x \frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} \mathbf{x}(\mathbf{u}) + H_u = 0.$$

□

1.2. Second derivative

The second total derivative $\frac{d^2}{du^2} L$ determines whether critical points of (0.2) are isolated and also which types of extrema they are. The projection onto $\mathcal{A}$ that is implicitly included in H can be used to compute $\frac{d^2}{du^2} L$ without having to compute an admissible point to second order.

Consider $z_0 = (\mathbf{x}_0, \mathbf{u}_0) \in \mathcal{A}$, $z = (\mathbf{x}_0 + d\mathbf{x}, \mathbf{u}_0 + d\mathbf{u}) =: z_0 + dz$ where $d\mathbf{x}$ is chosen such that $g(z) = 0$, i.e. $\mathbf{x}_0 + d\mathbf{x} = \mathbf{x}(\mathbf{u}_0 + d\mathbf{u})$, and λ_0 such that $H_x(z_0, \lambda_0) = 0$, i.e. *(ii)* of Theorem 1.5 is fulfilled. Instead of L we can consider H as $g(z) = g(z_0) = 0$ and so $H(z, \lambda) = L(z)$, $H(z_0, \lambda) = L(z_0)$ for any λ. Then, with the argument λ omitted,

$$H(z) - H(z_0) = \begin{pmatrix} H_x \\ H_u \end{pmatrix} \cdot \begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{pmatrix} \cdot \left(\begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \end{pmatrix}, \begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \end{pmatrix} \right) + \mathcal{O}\left(\|dz\|^2 \right)$$

and, from $g(\mathbf{x}_0 + d\mathbf{x}, \mathbf{u}_0 + d\mathbf{u}) = 0$,

$$d\mathbf{x} = -g_x^{-1} g_u d\mathbf{u} + \mathcal{O}\left(\|d\mathbf{u}\|^2 \right).$$

As we have arranged for $H_x = 0$, the second order term in $d\mathbf{x}$ does not contribute and we have

$$\frac{d^2}{d\mathbf{u}^2} L = H_{uu} - H_{ux}(\cdot, g_x^{-1} g_u \cdot) - H_{xu}(g_x^{-1} g_u \cdot, \cdot) + H_{xx}(g_x^{-1} g_u \cdot, g_x^{-1} g_u \cdot). \tag{1.10}$$

Note that for this result, H mainly serves to simplify calculations, as for comparison, one can directly compute

$$\begin{aligned}\frac{d}{d\mathbf{u}}L =& L_u - L_x g_x^{-1} g_u \\ \frac{d^2}{d\mathbf{u}^2}L =& L_{uu} + L_{ux}(\cdot, -g_x^{-1}g_u\cdot) \\ & - L_{xu}(g_x^{-1}g_u\cdot, \cdot) - L_{xx}(g_x^{-1}g_u\cdot, -g_x^{-1}g_u\cdot) \\ & - L_x\left(-g_x^{-1}g_{xu}(g_x^{-1}g_u\cdot, \cdot)\right) - L_x\left(-g_x^{-1}g_{xx}(g_x^{-1}g_u\cdot, -g_x^{-1}g_u\cdot)\right) \\ & - L_x g_x^{-1} g_{uu} - L_x g_x^{-1} g_{ux}(\cdot, -g_x^{-1}g_u\cdot),\end{aligned}$$

which with $\lambda = -L_x g_x^{-1}$ (from (1.8)) gives the same result after gathering terms.

1.3. Shadow prices

It is interesting to know how the optimal value L^* changes when we vary the constraint, e.g. by introducing an additional force to the dynamics of a control system or changing the initial condition. The latter is of particular interest for us, as it will yield the derivative $\frac{\partial}{\partial x_0}V = \frac{\mathrm{d}}{\mathrm{d}x_0}L^*$ of value functions of optimal control problems in later chapters. The sensitivities with respect to variations in the constraints are known as *shadow prices* and we will show that they are given by the adjoint λ.

Theorem 1.8 (Shadow Price Theorem)**.** *Let $g(z) = g(z,\mu) = g_0(z) + \mu$ with $g_0 : \mathcal{Z} \to \mathcal{Y}$ contain an additional parameter $\mu \in \mathcal{Y}$. Then the value $L^* = L(z^*)$ of a critical point $z^* = z^*(\mu)$ of (1.2) has the derivative*

$$\frac{\mathrm{d}L^*}{\mathrm{d}\mu} = \lambda^*,$$

where λ^* is the Lagrange multiplier corresponding to z^*.

Proof. Let $Dg_0 := Dg_0(z^*)$ and $DL := DL(z^*)$. We consider the variation $\mathrm{d}z^*$ of the critical point z^*. Let $\mathcal{Z} = \mathcal{Z}_1 \oplus \mathcal{N}(Dg_0)$ and $Dg_0^+ : \mathcal{R}(Dg_0) \to \mathcal{Z}_1$ as in Theorem 1.5, and decompose accordingly $\mathrm{d}z^* := \mathrm{d}z^*|_{\mathcal{Z}_1} + \mathrm{d}z^*|_{\mathcal{N}(Dg)}$.

As z^* always fulfills $g(z^*) = 0$, we have $Dg_0\,\mathrm{d}z^* + \mathrm{d}\mu = 0$ and so $Dg_0\,\mathrm{d}z^* = Dg_0\,\mathrm{d}z^*|_{\mathcal{Z}_1} = -\,\mathrm{d}\mu$, i.e. $-Dg_0^+\,\mathrm{d}\mu = \mathrm{d}z^*|_{\mathcal{Z}_1}$. Since z^* is a critical point, we have $DL\,\mathrm{d}z^*|_{\mathcal{N}(Dg_0)} = 0$ and finally $\mathrm{d}L^* = DL\,\mathrm{d}z^* = DL\,\mathrm{d}z^*|_{\mathcal{Z}_1} + DL\,\mathrm{d}z^*|_{\mathcal{N}(Dg_0)} = -DLDg_0^+\,\mathrm{d}\mu = \lambda \cdot \mathrm{d}\mu$. □

1.4. Derivative of F

To consider variations of critical points we need to invert the derivative of the function F defined in (1.6). We have

$$DF\begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \\ d\lambda \end{pmatrix} = \begin{pmatrix} H_{xx} & H_{xu} & g_x^* \\ H_{ux} & H_{uu} & g_u^* \\ g_x & g_u & 0 \end{pmatrix}\begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \\ d\lambda \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} dH_x \\ dH_u \\ dg \end{pmatrix}. \tag{1.11}$$

Successively solving the third, first and second equation yields

$$\begin{aligned} d\mathbf{x} &= g_x^{-1}(dg - g_u d\mathbf{u}) \\ d\lambda &= (g_x^*)^{-1}(dH_x - H_{xx}d\mathbf{x} - H_{xu}d\mathbf{u}) \\ &= (g_x^*)^{-1}\left(dH_x + H_{xx}g_x^{-1}(g_u d\mathbf{u} - dg) - H_{xu}d\mathbf{u}\right) \\ dH_u &= H_{ux}g_x^{-1}(dg - g_u d\mathbf{u}) + H_{uu}d\mathbf{u} \\ &\quad + g_u^*(g_x^*)^{-1}\left(dH_x + H_{xx}g_x^{-1}(g_u d\mathbf{u} - dg) - H_{xu}d\mathbf{u}\right) \end{aligned}$$

$$\begin{aligned} \stackrel{(1.10)}{\Longrightarrow} -\frac{\mathrm{d}^2 L}{\mathrm{d}\mathbf{u}^2}d\mathbf{u} &= \left(H_{ux}g_x^{-1}g_u - H_{uu} - g_u^*(g_x^*)^{-1}H_{xx}g_x^{-1}g_u + g_u^*(g_x^*)^{-1}H_{xu}\right)d\mathbf{u} \\ &= -dH_u + H_{ux}g_x^{-1}dg + g_u^*(g_x^*)^{-1}(dH_x - H_{xx}g_x^{-1}dg). \end{aligned} \tag{1.12}$$

Lemma 1.9. *Assume that the second derivatives appearing in* (1.11) *exist and are bounded linear operators. Then DF is invertible if and only if g_x and $\frac{\mathrm{d}^2 L}{\mathrm{d}\mathbf{u}^2}$ are invertible.*

Proof. " $\Longleftarrow$ ": If g_x is invertible, then so is g_x^* and the calculation in (1.12) gives the inverse of DF.

" $\Longrightarrow$ ": If DF is invertible, define $(d\mathbf{x}, d\mathbf{u}, d\lambda)$ as the solution of (1.11). One obtains $g_x^{-1}dg = d\mathbf{x}$ from (1.12) by setting $d\mathbf{u} := 0$ and then $\frac{\mathrm{d}^2 L}{\mathrm{d}\mathbf{u}^2}dH_u = d\mathbf{u}$ by setting $dg = 0$. □

For optimal control problems, g_x will always be invertible, but $\frac{\mathrm{d}^2 L}{\mathrm{d}\mathbf{u}^2}$ will, in many problems, be singular for some solutions. We will discuss this in Chapter 4.

2. Optimal control

In this chapter, we will formulate optimal control problems on finite-dimensional state and control spaces $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{U} = \mathbb{R}^m$ for both finite and infinite time horizons. W.l.o.g. we assume $t_0 = 0$.

We will use $\mathcal{X}^*$ in places where the dual space would appear in a more general setting, although $\mathcal{X}^* = \mathcal{X}$ for $\mathcal{X} = \mathbb{R}^n$.

2.1. Finite horizon

Let $[0, T]$ be a finite interval and $f : \mathcal{X} \times \mathcal{U} \times [0, T] \to \mathcal{X}$.

To fulfill the assumptions of Theorem 1.1, the spaces $\mathbf{X}$, $\mathbf{U}$ and $\mathcal{Y}$ need to be chosen so that Dg becomes a bounded operator. g involves the differentiation $\mathbf{x} \mapsto \dot{\mathbf{x}}$, which is bounded as a mapping $\mathcal{W}^{1,p} \to \mathcal{L}^p$. To make the pointwise operation $(x, u) \to f(x, u, t)$ well-behaved, we need control over the maxima of $\|x\|$ and $\|u\|$, and so we choose $p = \infty$. The relation $\dot{x} = f(x, u, t)$ also suggests that $\mathbf{u}$ should have the same regularity as $\dot{\mathbf{x}}$ and hence we use the spaces

$$\mathbf{X} := \mathcal{W}^{1,\infty}([0, T]; \mathcal{X}), \ \mathbf{U} := \mathcal{L}^\infty([0, T]; \mathcal{U}).$$

We consider a problem with a running cost c and a cost φ on the terminal state $x(T)$, but no terminal constraint. The total cost is

$$L(\mathbf{x}, \mathbf{u}) = \int_0^T c(x(t), u(t), t) \, \mathrm{d}t + \varphi(x(T)) \tag{2.1}$$

with $c : \mathcal{X} \times \mathcal{U} \times [0, T] \to \mathbb{R}$ and $\varphi : \mathcal{X} \to \mathbb{R}$.

The regularity conditions on f, c and φ in x and u will vary for different propositions. We always assume that f and c are measurable in t.

We enforce the dynamics

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

and the boundary condition

$$x(0) = x_0 \in \mathcal{X}$$

with a single constraint. To this end we define[1]

$$\mathcal{Y} := (\mathcal{L}^\infty([0, T]; \mathcal{X}) + \mathcal{X}\delta_0) \subset \mathcal{W}^{-1,\infty}([0, T]; \mathcal{X})$$

[1] Note that the precise choice of $\mathcal{Y}$ is important because we need Dg to be surjective for Theorem 1.1.

and $g : \mathbf{X} \times \mathbf{U} \to \mathcal{Y}$ by

$$g(\mathbf{x}, \mathbf{u}) := f(x, u, \cdot) - \frac{d}{dt}\mathbf{x} + (x_0 - x(0))\delta_0, \tag{2.2}$$

i.e. $\langle g(\mathbf{x}, \mathbf{u}), \psi \rangle = \int_0^T (f(x, u, t) - \dot{x}(t))\, \psi(t)\, \mathrm{d}t + (x_0 - x(0))\psi(0)$ for $\psi \in \mathcal{W}^{1,\infty}([0, T]; \mathcal{X}^*)$.

We will write the operator $\mathbf{x} \mapsto x(0)\delta_0$ as $\delta_0\delta_0^*$, resembling the projection operator $vv^\top$ onto a normalized vector v in a Hilbert space.

The operator $(\frac{\mathrm{d}}{\mathrm{d}t} + \delta_0\delta_0^*)$ can be interpreted as taking the derivative of a function which starts at 0 and has an impulse-valued derivative at $t = 0$. This rather uncommon approach is useful when computing the adjoint operator as we have

$$\begin{aligned} \left\langle \left(\frac{\mathrm{d}}{\mathrm{d}t} + \delta_0\delta_0^* \right) \mathbf{x}, \psi \right\rangle &= \int_0^T \dot{x}(t)\psi(t)\, \mathrm{d}t + x(0)\psi(0) \\ &= -\int_0^T x(t)\dot{\psi}(t)\, \mathrm{d}t + x(T)\psi(T) \\ &= \left\langle \mathbf{x}, \left(-\frac{\mathrm{d}}{\mathrm{d}t} + \delta_T\delta_T^* \right) \psi \right\rangle. \end{aligned} \tag{2.3}$$

2.1.1. Necessary conditions

The derivative of L is given by

$$DL(\mathbf{x}, \mathbf{u}) \begin{pmatrix} d\mathbf{x} \\ d\mathbf{u} \end{pmatrix} = \int_0^T c_x(x, u, t) \cdot dx(t) + c_u(x, u, t) \cdot du(t)\, \mathrm{d}t + \varphi'(x(T)) \cdot dx(T),$$

for which we write

$$\nabla L = \begin{pmatrix} c_x + \varphi' \cdot \delta_T \\ c_u \end{pmatrix} \in (\mathcal{L}^\infty([0, T]; \mathcal{X}^*) + \mathcal{X}^*\delta_T) \times \mathbf{U}^* \subset \mathbf{X}^* \times \mathbf{U}^*.$$

Similiarly,

$$\nabla g = \begin{pmatrix} g_x \\ g_u \end{pmatrix} = \begin{pmatrix} -\frac{d}{dt} + f_x - \delta_0\delta_0^* \\ f_u \end{pmatrix} : \mathbf{X} \times \mathbf{U} \to \mathcal{W}^{-1,\infty}([0, T]; \mathcal{X})$$

and with (2.3) we obtain

$$Dg^* = \begin{pmatrix} g_x^* \\ g_u^* \end{pmatrix} = \begin{pmatrix} \frac{d}{dt} + f_x^\top - \delta_T\delta_T^* \\ f_u^\top \end{pmatrix} : \mathcal{W}^{1,\infty}([0, T]; \mathcal{X}^*) \to \mathbf{X}^* \times \mathbf{U}^*.$$

To show that g_x is invertible, we consider the equation $-g_x\mathbf{x} = \mathbf{y} + y_0\delta_0 \in \mathcal{Y}$ with $\mathbf{y} \in \mathcal{L}^\infty([0, T]; \mathcal{X})$, $y_0 \in \mathcal{X}$. It encodes the differential equation

$$\begin{aligned} \dot{x} &= f_x \cdot x + y \text{ on } [0, T] \\ x(0) &= y_0, \end{aligned} \tag{2.4}$$

which is just the linearization of the dynamics and, if f_x is Lipschitz-continuous in t, has a unique solution $x \in \mathcal{W}^{1,\infty}([0,T];\mathcal{X})$. It follows that g_x is invertible on $\mathcal{Y} = \mathcal{L}^\infty([0,T];\mathcal{X}) + \mathcal{X}\delta_0$.

Similarly, $-g_x^*\lambda = \mathbf{y} + y_T\delta_T$, $\mathbf{y} \in \mathcal{L}^\infty([0,T];\mathcal{X}^*)$, $y_T \in \mathcal{X}^*$ corresponds to

$$\begin{aligned} -\dot\lambda &= f_x^\top \lambda + y \text{ on } [0,T] \\ \lambda(T) &= y_T, \end{aligned} \tag{2.5}$$

which is again a linear differential equation.

We now have the necessary conditions for a minimum:

Theorem 2.1. *Let $f(\cdot,\cdot,t)$, $c(\cdot,\cdot,t)$, and φ be continuously differentiable for all $t \in [0,T]$ and let f_x, f_u, c_x and c_u be uniformly in t continuous in (x,u) on compact sets in $\mathcal{X}\times\mathcal{U}$.*

If $(\mathbf{x},\mathbf{u}) \in \mathbf{X}\times\mathbf{U}$ is a minimum of

$$\begin{aligned} &\min_{\substack{\mathbf{x}\in\mathbf{X}\\ \mathbf{u}\in\mathbf{U}}} L(\mathbf{x},\mathbf{u}) \\ &\text{s.t. } \dot x = f \text{ weakly} \\ &\qquad x(0) = x_0, \end{aligned} \tag{2.6}$$

then there exits $\lambda \in \mathcal{W}^{1,\infty}([0,T],\mathcal{X}^*)$ such that

$$-\dot\lambda = f_x^\top \lambda + c_x, \tag{2.7a}$$

$$c_u = -f_u^\top \lambda \ , \tag{2.7b}$$

$$\dot x = f \tag{2.7c}$$

weakly on $[0,T]$ and

$$\lambda(T) = \varphi'(x(T)), \tag{2.7d}$$

$$x(0) = x_0. \tag{2.7e}$$

Proof. As $\mathbf{x}$ and $\mathbf{u}$ are in $\mathcal{L}^\infty$, $(x(t),u(t),t)$ is in some compact set for almost all t. It follows that e.g. for the operator norm of $f_x : \mathcal{L}^\infty([0,T];\mathcal{X}) \to \mathcal{L}^\infty([0,T];\mathcal{X})$, defined as the pointwise in t multiplication with $f_x(x(t),u(t),t)$, we have

$$\|f_x\|_{\mathcal{L}^\infty\to\mathcal{L}^\infty} \le \operatorname*{ess\,sup}_t \|f_x(x(t),u(t),t)\|_{\mathcal{X}} < \infty.$$

From this and similar bounds we see that L and g have continuous Fréchet-derivatives. $Dg(\mathbf{x},\mathbf{u}) : \mathbf{X}\times\mathbf{U} \to \mathcal{Y}$ is surjective, as g_x is invertible and $Dg(\mathbf{x},\mathbf{u})\cdot(g_x^{-1}\mathbf{y},0) = \mathbf{y}$ for all $\mathbf{y} \in \mathcal{Y}$. Hence we can apply Theorem 1.1 and obtain a $\lambda \in \mathcal{Y}^*$ fulfilling (1.6). Rewriting the operators as ODEs yields (2.7) and as λ solves the ODE (2.7a), we also have $\lambda \in \mathcal{W}^{1,\infty}$. □

Remark 2.2. *If c is strongly convex in u and f_u is independent of u, then* (2.7b) *uniquely*

determines u given x and λ. In that case we can eliminate u from (2.7a) *and* (2.7c) *to obtain an ODE for (x, λ). We will rely on this ODE to study variations of extremals in Chapter 4.*

Remark 2.3. *Equations* (2.7a), (2.7d) *and* (2.7b) *are a slightly weakened form of* Pontryagin's minimum principle *([Pon87]). The classical form requires that $c + f^\top \lambda$ is minimal in u, which implies* (2.7b)*, whereas conversely* (2.7b) *only implies that u is critical. If c is convex in u, both statements become equivalent.*

Theorem 2.4. *Let $f(\cdot,\cdot,t)$, $c(\cdot,\cdot,t)$, and φ be twice differentiable for all $t \in [0,T]$ such that the second derivatives are uniformly bounded on compact sets in $\mathcal{X} \times \mathcal{U} \times [0,T]$. Let $\mathbf{x} \in \mathbf{X}$, $\mathbf{u} \in \mathbf{U}$ and $\lambda \in \mathcal{L}^\infty([0,T]; \mathcal{X}^*)$.*

Then $DF(\mathbf{x}, \mathbf{u}, \lambda)$ is invertible if and only if $\mathrm{d}^2 L / \mathrm{d}u^2(\mathbf{x}, \mathbf{u}, \lambda)$ is invertible.

Proof. The boundedness of the second derivatives of H follows as in the proof of Theorem 2.1. Then the claim is a direct consequence of Lemma 1.9. □

2.1.2. Existence of minimizers

We will prove the existence of minimizers (under certain assumptions to be stated later) not directly for the optimal control problem, but first for the calculus of variations, which deals with cost functions $c(x, \dot{x}, t)$ involving the derivative $\dot{x}$ instead of the control u. Optimal control theory includes the calculus of variations as a special case with $f(x, u, t) = u$ and we will show how conversely OCPs can be translated to the CoV.

Minimizers in the calculus of variations

Let $1 < p < \infty$ and consider the problem

$$\begin{aligned} L(\mathbf{x}) &:= \int_0^T c(x, \dot{x}, t)\,\mathrm{d}t + \varphi(x(T)) = \min_{\mathcal{A}}!, \\ \mathcal{A} &:= \left\{ \mathbf{x} \in \mathcal{W}^{1,p}([0,T], \mathcal{X}) : x(0) = x_0 \right\}. \end{aligned} \tag{2.8}$$

In classical analysis, a function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is guaranteed to have a minimizer if Ω is compact and f is continuous, as any minimizing sequence will have an accumulation point, which is a minimizer. The following analysis of (2.8) is quite analogous: (weak) compactness can be achieved by demanding that L is *coercive* so that the search for a minimizer can be restricted to a bounded set. Continuity can be replaced by the weaker notion of lower semicontinuity, which requires c to be *convex* in $\dot{x}$. We will discuss both requirements and sketch the underlying ideas before proceeding to the detailed proof.

Coercivity and compactness

Definition 2.5. *A functional $L : \Omega \subseteq \mathcal{W}^{1,p} \to \mathbb{R}$ is called* coercive *on Ω, if $L(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\|_{\mathcal{W}^{1,p}} \to \infty$.*

It is easily seen that the functional L in (2.8) is coercive on the affine subspace $\mathcal{A}$ if there are constants $\gamma_1 > 0$, $\gamma_2 \in \mathbb{R}$, such that

$$\begin{aligned} c(x, \dot{x}, t) &\geq \gamma_1 \|\dot{x}\|^p + \gamma_2 \qquad \forall x, \dot{x}, t, \\ \varphi(x) &\geq \gamma_2 \qquad \forall x. \end{aligned}$$

(Recall that $\mathcal{A}$ includes the boundary condition for $\mathbf{x}$.)

Unfortunately the supremum $\|\dot{x}\|_{\mathcal{L}^\infty}$ cannot be controlled by the integral of c and so we typically get coercivity only for some $p < \infty$. Consequently we will prove at first only the existence of a minimizer in $\mathcal{W}^{1,p}$ and show later that it is also in $\mathcal{W}^{1,\infty}$.

We also need to exclude $p = 1$ because in $\mathcal{W}^{1,1}$ closed and bounded does not imply weakly compact. The following is an example of a problem without a minimizer.

Example 2.6. *Let* $\mathbf{x} : [0,1] \to \mathbb{R}$, $x_0 := 1$, $c(x, \dot{x}, t) := 2|x| + |\dot{x}|$, $\varphi(t) := 0$. *We have* $L(\mathbf{x}) \geq 2 \inf_t |x(t)| + (x_0 - \inf_t x(t)) \geq \inf_t |x(t)| + x_0 \geq 1$ *with equality if and only if* $\inf_t x(t) = 0$. $L(\mathbf{x}) = 1$ *is not achievable for any* $\mathbf{x} \in \mathcal{W}^{1,1}$ *as* $\inf_t x(t) = 0$ *would imply* $\int_0^1 |\dot{x}| \geq 1$ *and hence* $x = 0$ *a.e., a contradiction to* $x(0) = 1$. *On the other hand, the sequence*

$$x_i(t) := \begin{cases} 1 - it & , t \in [0, 1/i) \\ 0 & , t \in (1/i, 1] \end{cases}$$

achieves $L(\mathbf{x}_i) = 1 + \frac{1}{i} \xrightarrow[i\to\infty]{} 1$, *so* 1 *is the infimum but not the minimum of* L.

Convexity

Now let $1 < p < \infty$. We are still faced with the problem that the unit ball of the infinite dimensional space $\mathcal{W}^{1,p}$ is only weakly compact. For an argument completely analogous to the finite dimensional case to succeed, we would need L to be weakly continuous, which in general it is not.

Let us sketch what we can deduce so far: by assuming coercivity any minimizing sequence is bounded. Because bounded sets of $\mathcal{W}^{1,p}$ are weakly compact, every such sequence has a weak accumulation point $\bar{\mathbf{x}}$, so we can pass to a weakly convergent subsequence $\mathbf{x}_i \rightharpoonup \bar{\mathbf{x}}$. In detail this means $\mathbf{x}_i \to \bar{\mathbf{x}}$ almost uniformly (see the proof of Theorem 2.8), but $\dot{\mathbf{x}}_i \rightharpoonup \dot{\bar{\mathbf{x}}}$ only weakly in $\mathcal{L}^p$, i.e. the derivatives $\dot{\mathbf{x}}_i$ converge to $\dot{\bar{\mathbf{x}}}$ in average (w.r.t. to integration against functions in the dual space).

That $\dot{\bar{\mathbf{x}}}$ is an average in this sense suggests a way to ensure that $\bar{\mathbf{x}}$ is a minimizer: for convex functions f we have Jensen's inequality $f\left(⨍ g(t)\,\mathrm{d}t\right) \leq ⨍ f(g(t))\,\mathrm{d}t$, i.e. "value of average $\leq$ average value"). We will show that if c is convex in $\dot{x}$, then L is lower semicontinuous, i.e.

$$L(\bar{\mathbf{x}}) \leq \liminf_{i\to\infty} L(\mathbf{x}_i) = \inf L,$$

which is in fact enough, as the other inequality $L(\bar{\mathbf{x}}) \geq \inf L$ holds by definition.

The need for convexity is illustrated by an example due to Bolza [Bol02].

Example 2.7. *Let* $\mathbf{x} : [0,1] \to \mathbb{R}$, $x_0 := 0$, $c(x, \dot{x}, t) := x^2 + (\dot{x}^2 - 1)^2$, $\varphi(t) := 0$. *Obviously* $L(\mathbf{x}) \geq 0$ *and equality is not possible as it would require both* $x(t) = 0$ *and*

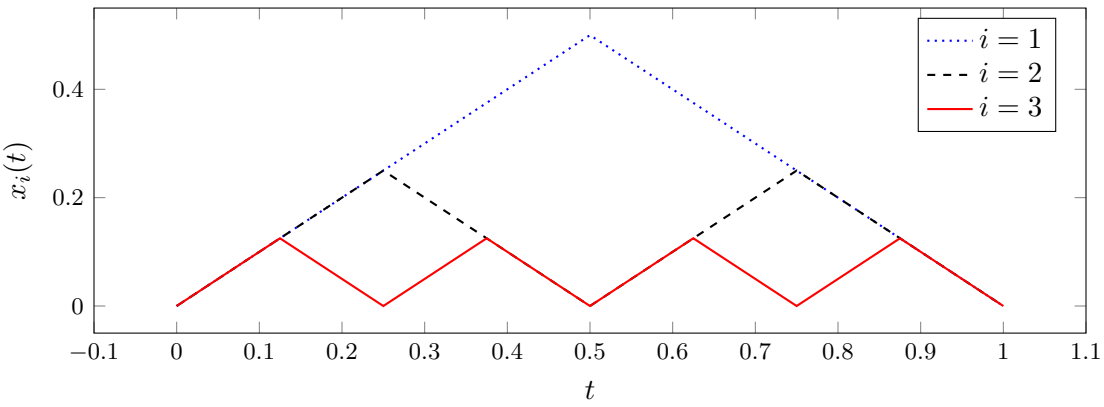

Figure 2.1.: Functions $\mathbf{x}_i$ in the Bolza example.

$\dot{x}(t) = \pm 1$ *to hold almost everywhere. Let* $\phi : \mathbb{R} \to \mathbb{R}$ *be the periodic sawtooth function*

$$\phi(t) := \begin{cases} s & , t = s + 2k, s \in [0,1), k \in \mathbb{Z} \\ 1 - s & , t = s + 2k + 1, s \in [0,1), k \in \mathbb{Z} \end{cases}$$

and $x_i(t) := 2^{-i}\phi(2^i t)$. *Then* $L(\mathbf{x}_i) \xrightarrow[i\to\infty]{} 0$ *and so* L *has* 0 *as an infimum but not as a minimum.*

More generally one can show, by applying L to functions similar to the $\mathbf{x}_i$ in Example 2.7, that the convexity of c in $\dot{x}$ is a necessary condition for L to be lower semicontinuous, cf. [Dac07, Theorem 3.15].

We will now prove the above claim that convexity of c in $\dot{x}$ implies the weak lower semicontinuity of L.

Theorem 2.8. *Let* φ *be lower semicontinuous. Let* $c : \mathbb{R}^n \times \mathbb{R}^n \times [0,T] \to \mathbb{R} \cup \{+\infty\}$ *be lower semicontinuous in* $(x, \dot{x})$ *and bounded from below, and let* $\dot{x} \mapsto c(x, \dot{x}, t)$ *be convex for all* $x \in \mathbb{R}^n$, $t \in [0,T]$. *Then* $L : \mathcal{W}^{1,p}([0,T], \mathbb{R}^n) \to \mathbb{R}$ *is weakly lower semicontinuous for any* $1 < p < \infty$.

Proof. Let $\{\mathbf{x}_i\}_{i=1}^{\infty} \subset \mathcal{W}^{1,p}$ be a sequence with $\mathbf{x}_i \rightharpoonup \bar{\mathbf{x}}$ weakly as $i \to \infty$. We have to show $L(\bar{\mathbf{x}}) \leq \liminf_{i\to\infty} L(\mathbf{x}_i) =: l$.

As $[0,T]$ is one-dimensional, the point evaluations $\mathbf{x} \mapsto x(t)$ are elements of the dual space of $\mathcal{W}^{1,p}$ and so $\mathbf{x}_i \to \mathbf{x}$ pointwise. By Egoroff's Theorem there exists, for any $\epsilon > 0$, a set $E_\epsilon \subset [0,T]$ with $\mu([0,T] \backslash E_\epsilon) \leq \epsilon$, where μ is the Lebesgue measure, such that $\mathbf{x}_i \to \bar{\mathbf{x}}$ uniformly on E_ϵ. Set $\delta_{i,\epsilon} := \sup_{t \in E_\epsilon, j \geq i} \|x_j(t) - \bar{x}(t)\|$ and note that $\delta_{i,\epsilon} \searrow 0$ as $i \to \infty$.

By assumption we have $c \geq \gamma$ for some $\gamma \in \mathbb{R}$. Hence

$$\begin{aligned} L(\mathbf{x}_i) &= \int_0^T c(x_i, \dot{x}_i, t)\,\mathrm{d}t + \varphi(x_i(T)) \\ &\geq \int_{E_\epsilon} c(x_i, \dot{x}_i, t)\,\mathrm{d}t + \varphi(x_i(T)) + \gamma\mu([0,T]\backslash E_\epsilon). \end{aligned} \tag{2.9}$$

We are now faced with the technical difficulty that, although $c(x_i, \dot{x}_i, t)$ is finite for almost all t if i is large enough (because $L(\mathbf{x}_i)$ will be finite), other expressions such as $c(x, \dot{x}_i, t)$ may be infinite (cf. Remark 2.11). We therefore define the relaxation $c^{(i)}(\dot{x}, t)$ as[2]

$$c^{(i)}(\cdot, t) := \operatorname{conv}\inf\{c(y, \cdot, t) : y \in B_{\delta_{i,\epsilon}}(\bar{x}(t))\}$$

and note that $c(x_i(t), \dot{x}_i(t), t) \geq c^{(i)}(\dot{x}_i(t), t)$ for $t \in E_\epsilon$ as $x_i(t) \in B_{\delta_{i,\epsilon}}(\bar{x}(t))$. Similarly, $c^{(i)}$ is monotonously increasing in i by virtue of $B_{\delta_{j,\epsilon}}(\bar{x}(t)) \subseteq B_{\delta_{i,\epsilon}}(\bar{x}(t))$ for $j \geq i$. Since c is lower semicontinuous in x, the infimum in the definition of $c^{(i)}$ is attained and we have the convex combination

$$c^{(i)}(\dot{x}, t) = \sum_{k=1}^{n+1} \alpha_k^{(i)}(\dot{x}, t) c\left(y_k^{(i)}(\dot{x}, t), \dot{x}, t\right)$$

with $\alpha_k^{(i)}(\dot{x}, t) \in [0,1]$, $\sum_{k=1}^{n+1} \alpha_k^{(i)}(\dot{x}, t) = 1$ and $y_k^{(i)}(\dot{x}, t) \in B_{\delta_{i,\epsilon}}(\bar{x}(t))$, in particular $y_k^{(i)}(\dot{x}, t) \xrightarrow[i\to\infty]{} \bar{x}(t)$ for all $k, \dot{x}, t$. Using again the lower semicontinuity of c, we see that

$$\liminf_{i\to\infty} c^{(i)}(\dot{x}, t) \geq \sum_{k=1}^{n+1} \left(\alpha_k^{(i)}(\dot{x}, t) \liminf_{i\to\infty} c\left(y_k^{(i)}(\dot{x}, t), \dot{x}, t\right)\right) \geq c(\bar{x}, \dot{x}, t).$$

Together with $c(\bar{x}, \dot{x}, t) \geq c^{(i)}(\dot{x}, t)$ (since $\bar{x}(t) \in B_{\delta_{i,\epsilon}}(\bar{x}(t))$) and the monotonicity of $c^{(i)}$ in i this gives

$$\lim_{i\to\infty} c^{(i)}(\dot{x}, t) = \liminf_{i\to\infty} c^{(i)}(\dot{x}, t) = c(\bar{x}, \dot{x}, t). \tag{2.10}$$

We need to establish that $c^{(i)}(\dot{\bar{x}}, t)$ is finite for almost all t. Clearly, $c^{(i)}(\cdot, t)$ is finite on $C_i(t) := \overline{\operatorname{conv}\{\dot{x}_j(t) : j \geq i\}}$ for almost all t, so assume by way of contradiction, that $\dot{\bar{x}}(t) \notin C_i(t)$ for t in some set G of non-vanishing measure. As $C_i(t)$ is convex, there exists $v(t) \in \mathbb{R}^n$ and $\alpha(t) \in \mathbb{R}$ for all $t \in G$ such that $v(t) \cdot \dot{\bar{x}}(t) > \alpha(t)$ and $v(t) \cdot \dot{\bar{x}}_j(t) \leq \alpha(t)$ for all $j \geq i$. We set $v(t) := 0$ and $\alpha(t) = 0$ for $t \notin G$ and obtain the contradiction $\int_0^T v \cdot \dot{\bar{x}} > \int_0^T \alpha \geq \int_0^T v \cdot \dot{x}_j \xrightarrow[j\to\infty]{} \int_0^T v \cdot \dot{\bar{x}}$.

We let i be fixed for now. The convex function $c^{(i)}(\cdot, t)$ has a non-empty set of subgradients at $\dot{\bar{x}}(t)$ whenever $c^{(i)}(\dot{\bar{x}}, t) < \infty$. As we have just demonstrated, this holds for almost all t. On these t, we define $c_{\dot{x}}^{(i)}(\dot{\bar{x}}, t)$ to be any of those subgradients. For $\tilde{\epsilon} > 0$,

[2] conv means the lower semicontinuous envelope w.r.t. the argument $\dot{x}$.

let $F_\epsilon^{i,\tilde\epsilon} := E_\epsilon \cap \left\{t \in [0,T] : \left\| c_{\dot x}^{(i)}(\dot{\bar x}, t)\right\| \le \frac{1}{\tilde\epsilon}\right\}$. Then $\mu(E_\epsilon \backslash F_\epsilon^{i,\tilde\epsilon}) \to 0$ as $\tilde\epsilon \to 0$. Due to convexity we can estimate

$$\int_{F_\epsilon^{i,\tilde\epsilon}} c^{(i)}(\dot x_j, t)\,\mathrm{d}t \ge \int_{F_\epsilon^{i,\tilde\epsilon}} c^{(i)}(\dot{\bar x}, t)\,\mathrm{d}t + \int_{F_\epsilon^{i,\tilde\epsilon}} c_{\dot x}^{(i)}(\dot{\bar x}, t)\cdot(\dot x_j - \dot{\bar x})\,\mathrm{d}t.$$

$\chi_{F_\epsilon^{i,\tilde\epsilon}} c_{\dot x}^{(i)}(\dot{\bar x}, t)$ is in the dual space of $\mathcal{W}^{1,p}$ and so the weak convergence $\mathbf{x}_j \rightharpoonup \bar{\mathbf{x}}$ as $j \to \infty$ implies

$$\lim_{j\to\infty} \int_{F_\epsilon^{i,\tilde\epsilon}} c_{\dot x}^{(i)}(\dot{\bar x}, t)(\dot x_j - \dot{\bar x})\,\mathrm{d}t = 0.$$

Recall that $c(x_j, \dot x_j, t) \ge c^{(j)}(\dot x_j, t) \ge c^{(i)}(\dot x_j, t)$ for $j \ge i$. Now we have

$$\liminf_{j\to\infty} \int_{F_\epsilon^{i,\tilde\epsilon}} c(x_j, \dot x_j, t)\,\mathrm{d}t \ge \liminf_{j\to\infty} \int_{F_\epsilon^{i,\tilde\epsilon}} c^{(i)}(\dot x_j, t)\,\mathrm{d}t \ge \int_{F_\epsilon^{i,\tilde\epsilon}} c^{(i)}(\dot{\bar x}, t)\,\mathrm{d}t$$

and with $\tilde\epsilon \to 0$

$$\liminf_{j\to\infty} \int_{E_\epsilon} c(x_j, \dot x_j, t)\,\mathrm{d}t \ge \int_{E_\epsilon} c^{(i)}(\dot{\bar x}, t)\,\mathrm{d}t.$$

Taking the limit $i \to \infty$, the monotone convergence (2.10) yields

$$\liminf_{j\to\infty} \int_{E_\epsilon} c(x_j, \dot x_j, t)\,\mathrm{d}t \ge \int_{E_\epsilon} c(\bar x, \dot{\bar x}, t)\,\mathrm{d}t.$$

Returning to (2.9), the above equation and the lower semicontinuity of φ imply that

$$l = \liminf_{i\to\infty} L(\mathbf{x}_i) \ge \int_{E_\epsilon} c(\bar x, \dot{\bar x}, t)\,\mathrm{d}t + \varphi(\bar x(T)) + \gamma\mu([0,T]\backslash E_\epsilon)$$

for any $\epsilon > 0$. With $\epsilon \to 0$ and the Monotone Convergence Theorem we obtain that $l \ge L(\bar{\mathbf{x}})$ as claimed. □

We can now show the existence of a minimizer.

Theorem 2.9. *Let $L : \mathcal{W}^{1,p} \to \mathbb{R}$, $1 < p < \infty$, be coercive on $\mathcal{A}$ and let φ be lower semicontinuous. Let c be continuous and bounded from below, and let $\dot x \mapsto c(x, \dot x, t)$ be convex for all $x \in \mathbb{R}^n$, $t \in [0,T]$.*

Then there exists at least one minimizer of

$$L(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{A}}!$$

Proof. Set $l := \inf_{\mathbf{x}\in\mathcal{A}} L(\mathbf{x})$. If $l = \infty$, every $\mathbf{x}$ is a minimizer, so we only need to consider $l < \infty$. Let $\{\mathbf{x}_i\}_{i=1}^\infty$ be a minimizing sequence, i.e. $L(\mathbf{x}_i) \to l$ as $i \to \infty$.

As L is coercive, we have $\sup_i \|\mathbf{x}_i\|_{\mathcal{W}^{1,p}} < \infty$ and consequently $\{\mathbf{x}_i\}$ has a weak accumulation point, i.e. there is $\bar{\mathbf{x}} \in \mathcal{W}^{1,p}$ and a subsequence $\{\mathbf{x}_{i_j}\}$ such that $\mathbf{x}_{i_j} \rightharpoonup \bar{\mathbf{x}}$ weakly in $\mathcal{W}^{1,p}$.

Since the functional $\mathbf{x} \mapsto x(0)$ is in the dual space of $\mathcal{W}^{1,p}$, we have $\bar{x}(0) = \lim_{j\to\infty} x_{i_j}(0) = x_0$ and so $\bar{\mathbf{x}} \in \mathcal{A}$. Finally, it follows with Theorem 2.8 that $l \geq L(\bar{\mathbf{x}}) \geq \inf_{\mathbf{x}\in\mathcal{A}} L(\mathbf{x}) = l$ and we see that $\bar{\mathbf{x}}$ is a minimizer. □

Minimizers in optimal control

In order to apply Theorem 2.9 to the optimal control problem, we define

$$\widetilde{c}(x, \dot{x}, t) := \inf_u \{c(x, u, t) : f(x, u, t) = \dot{x}\}, \tag{2.11}$$

with $\inf \emptyset := \infty$ and

$$\widetilde{L}(\mathbf{x}) := \int_0^T \widetilde{c}(x(t), \dot{x}(t), t)\,\mathrm{d}t + \varphi(x(T)), \qquad \mathbf{x} \in \mathcal{A}.$$

Under the assumption that the infimum in (2.11) is attained, the corresponding minimization problem is equivalent to the optimal control problem

$$\begin{aligned} &\min_{\substack{\mathbf{x}\in\mathcal{W}^{1,p}\\ \mathbf{u}\in\mathcal{L}^{1,p}}} L(\mathbf{x}, \mathbf{u}) \\ &\quad \text{s.t. } \dot{x} = f(x, u, t) \text{ weakly} \\ &\qquad x(0) = x_0 \end{aligned} \tag{2.12}$$

Lemma 2.10. *Assume that the infimum in* (2.11) *is attained whenever* $\widetilde{c} < \infty$. *Let* L^* *be the minimal values of the OCP* (2.12) *and*

$$\widetilde{L}^* := \inf_{\mathbf{x}\in\mathcal{A}} \widetilde{L}(\mathbf{x}). \tag{2.13}$$

Then $L^* = \widetilde{L}^*$ *and further, if* (2.13) *has a minimizer* $\widetilde{\mathbf{x}}^*$, *then* $\left(\widetilde{\mathbf{x}}^*, \widetilde{\mathbf{u}}\left(\dot{\widetilde{\mathbf{x}}}^*\right)\right)$, *with* $\widetilde{u}$ *a minimizer in* (2.11), *is a minimizer of* (2.12).

Proof. Let $(\mathbf{x}, \mathbf{u})$ be admissible for (2.12). Then $f(x, u, t) = \dot{x}$ weakly and consequently $\widetilde{c}(x, \dot{x}, t) \leq c(x, u, t)$ for almost all t. It follows that $\widetilde{L}(\mathbf{x}) \leq L(\mathbf{x}, \mathbf{u})$ and hence $\widetilde{L}^* \leq L^*$.

For the reverse implication let $\widetilde{L}(\mathbf{x}) < \infty$. (If $\widetilde{L}^* = \infty$, the remaining claims hold trivially.) This implies that, for almost all t, $\widetilde{c}(x, \dot{x}, t) < \infty$ holds, $\widetilde{u} = \widetilde{u}(x, \dot{x}, t)$ defined as above exists and $\dot{x} = f(x, \widetilde{u}, t)$. Consequently, $(\mathbf{x}, \widetilde{\mathbf{u}})$ is admissible and $L(\mathbf{x}, \widetilde{\mathbf{u}}) = \widetilde{L}(\mathbf{x})$. This proves $\widetilde{L}^* \geq L^*$ and, with $\mathbf{x} = \widetilde{\mathbf{x}}^*$, the final claim. □

Remark 2.11. *If* $\dim \mathcal{U} < \dim \mathcal{X}$, *then* $\widetilde{c} = \inf \emptyset = \infty$ *almost everywhere. This necessitates the relaxation to the* $c^{(i)}$ *in the proof of Theorem 2.9.*

It remains to give some conditions when Theorem 2.9 is applicable to (2.13). Note first that convexity of $\widetilde{c}$ in $\dot{x}$ requires that the set $\mathcal{V}(x, t) := \{f(x, u, t) : u \in \mathcal{U}\}$ of possible velocities is convex. $\mathcal{V}(x, t)$ is parametrized by the linear space $\mathcal{U}$. Unless some degeneracy occurs, $\mathcal{V}(x, t)$ is a manifold, which has to be flat to be convex. It is

therefore not excessively restrictive to assume that $\mathcal{V}(x,t)$ is a linear space with a simple parametrization, leading us to the following defininiton:

Definition 2.12 ([Cla13, 23.8]). *A control system $f : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}$ is* finitely generated *if f has the form*

$$f(x,u,t) = v_0(x,t) + \sum_{j=1}^{m} v_j(x,t)u_j,$$

where $v_j : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$, $j = 0,\dots,m$.

Lemma 2.13. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}^n$ be finitely generated with v_j continuous in x and let $c : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}$ be convex in u, continuous in (x,u), and fulfill the coercivity condition*

$$c(x,u,t) \geq \gamma_1 \|u\|^p + \gamma_2 \ \forall x,t$$

for some $p > 1, \gamma_1 > 0, \gamma_2 \in \mathbb{R}$.

Then $\widetilde{c}$ is convex in $\dot{x}$, lower semicontinuous in $(x,\dot{x})$, and on any bounded set $\Omega \subset \mathbb{R}^n$ fulfills

$$\widetilde{c}(x,\dot{x},t) \geq \widetilde{\gamma}_1 \|\dot{x}\|^p + \widetilde{\gamma}_2 \ \forall x \in \Omega, t$$

for some $\widetilde{\gamma}_1 > 0, \widetilde{\gamma}_2 \in \mathbb{R}$ depending on Ω. Further, the infimum in (2.11) is attained whenever $\widetilde{c} < \infty$.

Proof. To show convexity in $\dot{x}$, fix x and t and set $A := (v_1(x,t),\dots,v_m(x,t)) \in \mathbb{R}^{n\times m}$. We can assume $v_0 = 0$ by replacing $\dot{x}$ with $\dot{x} - v_0$ without affecting convexity. Obviously, $\widetilde{c}(x,\dot{x},t) = \infty$ for $\dot{x} \notin \mathcal{R}(A)$, so we need to show convexity on $\mathcal{R}(A)$. Let A^+ be a pseudoinverse of A. Then, for $\dot{x} \in \mathcal{R}(A)$,

$$\tilde{c}(x,\dot{x},t) = \inf_{w \in \mathcal{N}(A)} c(x, A^+\dot{x} + w, t)$$

and the infimum is attained because $c(x, A^+\dot{x} + \cdot, t)$ is a convex function. Now let $\lambda \in [0,1]$; $\dot{x}_1, \dot{x}_2 \in \mathcal{R}(A)$ and let w_1, w_2 be corresponding minimizers. We have

$$\begin{aligned}
&\widetilde{c}(x, \lambda\dot{x}_1 + (1-\lambda)\dot{x}_2, t) \\
&\leq c\left(x, \lambda(A^+\dot{x}_1 + w_1) + (1-\lambda)(A^+\dot{x}_2 + w_2), t\right) \\
&\leq \lambda c\left(x, A^+\dot{x}_1 + w_1, t\right) + (1-\lambda) c\left(x, A^+\dot{x}_2 + w_2, t\right) \\
&= \lambda\widetilde{c}(x,\dot{x}_1,t) + (1-\lambda)\widetilde{c}(x,\dot{x}_2,t),
\end{aligned}$$

proving the convexity in $\dot{x}$.

$A(x,t)$and $v_0(x,t)$ are continuous and so on the bounded set $\Omega \times [0,T]$ we have $\|A(x,t)\| \leq \alpha$ and $\|v_0(x,t)\| \leq \alpha$ for some $\alpha \in \mathbb{R}$. Consequently we have, for some $\widetilde{\gamma}_1 > 0, \widetilde{\gamma}_2 \in \mathbb{R}$,

$$\widetilde{c}(x,\dot{x},t) \geq \gamma_1 \|u\|^p + \gamma_2 \geq \gamma_1 \left(\frac{\|\dot{x} - v_0(x,t)\|}{\|A(x,t)\|}\right)^p + \gamma_2 \geq \widetilde{\gamma}_1 \|\dot{x}\|^p + \widetilde{\gamma}_2,$$

where u is again the minimizer in (2.11). (If $\|A(x,t)\| = 0$, then either $\dot{x} = v_0(x,t)$ or $\widetilde{c}(x,\dot{x},t) = \infty$, so in any case the claim remains true.)

It remains to show the lower semicontinuity in $(x,\dot{x})$. Let t be fixed and $(x_i,\dot{x}_i) \xrightarrow[i\to\infty]{} (x,\dot{x})$. We need to show $\widetilde{c}(x,\dot{x},t) \le \liminf_{i\to\infty} \widetilde{c}(x_i,\dot{x}_i,t) =: l$. If $l = \infty$, we are done. Otherwise, we can assume w.l.o.g. that $\widetilde{c}(x_i,\dot{x}_i,t) < \infty$ for all i and, by passing to a subsequence, that $\lim_{i\to\infty} \widetilde{c}(x_i,\dot{x}_i,t) = l$. Let u_i be the corresponding minimizers. Due to the coercivity of c, there is an accumulation point u of $\{u_i\}_{i=1}^{\infty}$. By continuity, we have $A(x,t)u + v_0(x,t) = \dot{x}$ and $c(x,u,t) = l$. It follows that $\widetilde{c}(x,\dot{x},t) \le l$ as claimed. □

Theorem 2.14. *Let φ be lower semicontinuous and bounded from below. Let f and c fulfill the conditions of Lemma 2.13. If v_j, $j = 0,\dots,m$ grows at most linearly, i.e. if there is a $\nu \in \mathbb{R}$ such that*

$$\|v_j(x,t)\| \le \nu(1 + \|x\|) \ \forall j,x,t,$$

then (2.12) *has a minimizer.*

Proof. We need to rectify that Lemma 2.13 gives the coercivity of $\widetilde{c}$ in $\dot{x}$ only for x in bounded sets. We will use that the minimizer remains in a bounded set and that we can consequently increase $\widehat{c}$ outside this set without changing the minimum or the minimizer.

As before, we can assume that the optimal value L^* of (2.12) is finite. The coercivity of c and the lower bound on φ imply (using $\mathcal{L}^1([0,T]) \subset \mathcal{L}^p([0,T])$ and the equivalence of the finite dimensional norms) that there is some constant $\kappa \in \mathbb{R}$ such that, for all admissible $(\mathbf{x},\mathbf{u})$,

$$L(\mathbf{x},\mathbf{u}) \le L^* + 1 \implies \int_0^T \|u(t)\|_1 \,\mathrm{d}t \le \kappa. \tag{2.14}$$

By the assumptions on the v_j, we have $\dot{x}(t) \le \nu(1+\|x\|)(1+\|u\|_1)$ and so it follows from Gronwall's inequality that

$$\|x(t)\| \le (\|x_0\| + \nu T + \nu\kappa)\exp(\nu(T+\kappa)) =: M \ \forall t \in [0,T]$$

for all admissible $(\mathbf{x},\mathbf{u})$ with $L(\mathbf{x},\mathbf{u}) \le L^* + 1$.

Define $\widehat{c}(x,\dot{x},t) := \widetilde{c}(x,\dot{x},t) + (1 - \chi_{[-M,M]}(\|x\|))\,\|\dot{x}\|^p$ and the corresponding value functional $\widehat{L}$ as $\widetilde{L}$ with $\widehat{c}$ in the place of $\widetilde{c}$. Obviously $\widetilde{L} \le \widehat{L}$. For any $1 > \epsilon > 0$ there exists an admissible pair $(\mathbf{x}_\epsilon,\mathbf{u}_\epsilon)$ with $L(\mathbf{x},\mathbf{u}) \le L^* + \epsilon$ and $\|x(t)\| \le M$ for all t. From the definition of $\widehat{c}$ and the proof of Lemma 2.10 we see that $\widehat{L}(\mathbf{x}_\epsilon) = \widetilde{L}(\mathbf{x}_\epsilon) \le L(\mathbf{x}_\epsilon,\mathbf{u}_\epsilon) \le L^* + \epsilon$. With Lemma 2.10 and $\epsilon \to 0$ it follows that $\widehat{L}^* = \widetilde{L}^* = L^*$.

$\widehat{c}$ fulfills the conditions of Lemma 2.13 and so, by Theorem 2.9, $\widehat{L}$ has a minimizer $\widehat{\mathbf{x}}$, which is also a minimizer of $\widetilde{L}$ since

$$\widetilde{L}^* \le \widetilde{L}(\widehat{\mathbf{x}}) \le \widehat{L}(\widehat{\mathbf{x}}) = \widehat{L}^* = \widetilde{L}^*.$$

With Lemma 2.10 we finally conclude that (2.12) has a minimzer. □

We cannot conclude that this minimizer fulfills the necessary conditions unless we show that it is also in $\mathcal{W}^{1,\infty} \times \mathcal{L}^\infty$. This is not generally true as the following example shows:

Example 2.15. *Let* $\mathbf{x}, \mathbf{u} : [0,1] \to \mathbb{R}$, $f(x,u,t) := u$, $c(x,u,t) := \frac{1}{2}(u-1)^2 - \frac{1}{3}\left(x - \frac{3}{2}t^{\frac{2}{3}}\right)u^4$, $x_0 := 0$. *Then* $x^*(t) = \frac{3}{2}t^{\frac{2}{3}}$, $u^*(t) = t^{-\frac{1}{3}}$, $\lambda^* = 1 - t^{-\frac{1}{3}}$ *is an extremal trajectory with* $u^* \to \infty$ *as* $t \to 0$.

In the following Theorem, we therefore require assumptions on f and c which prevent the adjoint equation (2.7a) from blowing up. Note that the assumption on f is fulfilled if f is finitely generated.

Theorem 2.16. *If, in addition to the assumptions of Theorem 2.14,*

- φ *is continuously differentiable,*
- f *is differentiable with* $\|f_x(x,u,t)\| \leq \nu(1 + \|u\|)$ $\forall x, u, t$ *for some* $\nu \in \mathbb{R}$, *and*
- $\|c_x(x,u,t)\| \leq \gamma(x)(1 + \|u\|^p)$ $\forall x, u, t$ *for some continuous function* γ,

then $\mathbf{x}^* \in \mathcal{W}^{1,\infty}([0,T];\mathcal{X})$ *and* $\mathbf{u}^* \in \mathcal{L}^\infty([0,T];\mathcal{U})$.

Proof. Note first that $\mathbf{x}^* \in \mathcal{W}^{1,p}([0,T])$ implies $\mathbf{x}^* \in \mathcal{L}^\infty([0,T])$. Therefore $\dot{x}^* = f(x^*,u^*) \to \infty$ can only occur for $u^* \to \infty$ and we only need to verify that $\mathbf{u}^* \in \mathcal{L}^\infty([0,T])$.

Assume, by way of contradiction, that $\mathbf{u}^* \notin \mathcal{L}^\infty([0,T])$. We will show that such a $(\mathbf{x}^*, \mathbf{u}^*)$ is not a minimizer since L can be reduced by removing the peaks of $\mathbf{u}^*$.

For $0 < \epsilon \leq 1$ set $E_\epsilon := \{t \in [0,T] : \|\dot{x}\| > \frac{1}{\epsilon}\}$. We have $\mu(E_\epsilon) =: \delta_\epsilon > 0$. Now define $(\widehat{\mathbf{x}}, \widehat{\mathbf{u}})$ by setting $\widehat{u}(t) := \chi_{[0,1/\epsilon]}(\|u^*(t)\|)u^*(t)$, the cut-off control, and $\dot{\widehat{x}}(t) = f(\widehat{x}, \widehat{u}, t)$, $\widehat{x}(0) = x_0$, the resulting trajectory according to the dynamics. Set $\kappa := \int_{E_\epsilon} \|\dot{x}^*(t)\| \,\mathrm{d}t$. Note that $\kappa < \infty$ as $\dot{\mathbf{x}} \in \mathcal{L}^p([0,T])$. We have $\delta_\epsilon = \epsilon \int_{E_\epsilon} \frac{1}{\epsilon} \,\mathrm{d}t \leq \epsilon\kappa$ and

$$\int_{E_\epsilon} \|f(x^*,u^*,t) - f(\widehat{x},\widehat{u},t)\| \,\mathrm{d}t \leq \int_{E_\epsilon} \|\dot{x}^*(t)\| \,\mathrm{d}t + \delta_\epsilon \sup_t \|f(\widehat{x},0,t)\| \leq (1 + \epsilon\alpha_1)\kappa$$

with $\alpha_1 := \sup_t \|f(\widehat{x},0,t)\|$.

Set $\beta := T + \int_0^T \|u^*(t)\| \,\mathrm{d}t < \infty$. By Gronwall's inequality and the bound on $\|f_x\|$ it follows that

$$\sup_{t\in[0,T]} \|\widehat{x}(t) - x^*(t)\| \leq (1 + \epsilon\alpha_1)\kappa e^{\nu\beta}.$$

Let $L_{E_\epsilon}(\mathbf{x},\mathbf{u}) := \int_{E_\epsilon} c(x,u,t) \,\mathrm{d}t$ and $L_{E_\epsilon^c} := L - L_{E_\epsilon}$. As $\kappa \leq \int_0^T \|\dot{x}^*(t)\| \,\mathrm{d}t < \infty$ is bounded independently of ϵ, there are also bounds $\alpha_2 := \sup_{\epsilon,t} \|\widehat{x}(t)\| < \infty$, $\alpha_3 := \sup_{\|x\|<\alpha_2} \gamma(x)$ and $\alpha_4 := \sup_{\|x\|<\alpha_2} \|\varphi'(x)\|$. Together with the bound on c_x we have

$$L_{E_\epsilon^c}(\widehat{\mathbf{x}},\widehat{\mathbf{u}}) \leq L_{E_\epsilon^c}(\mathbf{x}^*,\mathbf{u}^*) + (1 + \epsilon\alpha_1)\kappa e^{\nu\beta}\left(\alpha_3(T + \|\mathbf{u}^*\|_{\mathcal{L}^p}) + \alpha_4\right). \tag{2.15}$$

We also have constants m and $M(\epsilon)$ such that

$$c(\widehat{x}(t),0,t) \le m \ \forall t, \qquad \widetilde{c}(x^*,\dot{x}^*,t) \ge M(\epsilon)\,\|\dot{x}^*\| + \widetilde{\gamma}_2 \ \forall t \in E_\epsilon,$$

with $M(\epsilon) \to \infty$ as $\epsilon \to 0$, which follows from the bounds on $\widehat{x}$ and c_x, resp. the coercivity of $\widetilde{c}$. We estimate

$$L_{E_\epsilon}(\widehat{\mathbf{x}},\widehat{\mathbf{u}}) \le L_{E_\epsilon}(\mathbf{x}^*,\mathbf{u}^*) + (m+\widetilde{\gamma}_2)\epsilon\kappa - M(\epsilon)\kappa. \tag{2.16}$$

Adding (2.15) and (2.16) we see that

$$L(\widehat{\mathbf{x}},\widehat{\mathbf{u}}) \le L(\mathbf{x}^*,\mathbf{u}^*) + \kappa\,(\alpha_5 + \alpha_6\epsilon - M(\epsilon)))$$

with $\alpha_5 := e^{\nu\beta}\,(\alpha_3(T + \|\mathbf{u}^*\|_{\mathcal{L}^p}) + \alpha_4)$, $\alpha_6 := \alpha_1\alpha_5 + m - \widetilde{\gamma}_2$. For $\epsilon > 0$ small enough, we obtain $L(\widehat{\mathbf{x}},\widehat{\mathbf{u}}) < L(\mathbf{x}^*,\mathbf{u}^*)$ in contradiction to $(\mathbf{x}^*,\mathbf{u}^*)$ being a minimizer. □

Bibliographical notes

Using lower semicontinuity and compactness in a suitable topology is a standard approach to proving existence of minimizers. Our Theorems 2.8 and 2.9 are adapted from [Eva10, Chapter 8.2], which however requires the cost function c to be smooth. It is possible to approach the optimal control problem directly without the detour through the calculus of variations. For example, [Cla13, Theorem 23.11] shows directly the existence of minimizers for finitely generated control systems. Our approach, on the other hand, has the advantage of yielding reusable intermediate results. Theorem 2.8, for example, can also be used for systems with bounded control sets as long as the set of possible velocities is convex.

2.2. Infinite horizon

In this section, we consider problems on the infinite time horizon $[0,\infty)$. The spaces are now

$$\mathbf{X} = \mathcal{W}^{1,\infty}([0,\infty);\mathcal{X}),\ \mathbf{U} = \mathcal{L}^\infty([0,\infty);\mathcal{U}).$$

We impose some restrictions by allowing only autonomous dynamics and limiting the dependence of c on the time t to exponential discounting with a factor $\mu \ge 0$. Without a terminal time, we also do not have a cost of the terminal state. Hence, the problem we consider is

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}}\ & L(\mathbf{x},\mathbf{u}) := \int_0^\infty \mathrm{e}^{-\mu t} c(x(t),u(t))\,\mathrm{d}t \\ \text{s.t. }\ & \dot{x} = f(x,u) \text{ weakly} \\ & x(0) = x_0. \end{aligned} \tag{2.17}$$

Note that the total cost L may easily become infinite. Even in problems where the optimal trajectory has finite cost (e.g. one often wants to compute the optimal way to

reach a stable equilibrium $f(x^*, u^*) = 0$ and then chooses c such that $c(x^*, u^*) = 0$), trajectories arising as intermediate iterates during its computation may have $L = \infty$ and gradients ∇L which are not bounded functionals.

To deal with this complication, we introduce the partial costs up to time T,

$$L^T(\mathbf{x}, \mathbf{u}) := \int_0^T \mathrm{e}^{-\mu t} c(x(t), u(t))\,\mathrm{d}t,$$

and use a notion of optimality which loosely speaking requires that trajectory is not dominated:

Definition 2.17. *An admissible pair* $(\mathbf{x}^*, \mathbf{u}^*)$ *is* weakly (locally) overtaking minimal*, if (there exists* $\delta > 0$ *such that) for any admissible pair* $(\mathbf{x}, \mathbf{u})$ *(with* $\|\mathbf{u}^* - \mathbf{u}\| < \delta$*), any* $\epsilon > 0$ *and any* $T > 0$ *one can find* $T' \geq T$ *such that*

$$L^{T'}(\mathbf{x}, \mathbf{u}) \geq L^{T'}(\mathbf{x}^*, \mathbf{u}^*) - \epsilon.$$

The non-local version of this concept (without the restriction $\|\mathbf{u}^* - \mathbf{u}\| < \delta$) was defined by Carlson and Haurie in [CH87].

Concerning the dynamics, we now also need to state explicitly that we allow only compactly supported test functions for the weak derivative and define $g : \mathcal{L}^\infty([0,T];\mathcal{X}) \times \mathbf{U} \to (\mathcal{L}^\infty([0,T];\mathcal{X}) + \mathcal{X}\delta_0) \subset \left(\mathcal{W}_c^{1,\infty}([0,\infty);\mathcal{X})\right)^*$ by

$$g(\mathbf{x}, \mathbf{u}) := f(x, u, \cdot) - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} + (x_0 - x(0))\delta_0,$$

As a further assumption we will require that the linearization around the (w.l.o.) optimal trajectory fulfills the asymptotic stability condition

$$\int_t^\infty \left\|\Psi^{s\leftarrow t}\right\|_{\mathcal{X}\to\mathcal{X}} \mathrm{d}s \leq \beta < \infty \;\forall t \geq 0, \tag{2.18}$$

where $\Psi^{s\leftarrow t}$ is the fundamental solution of $\dot{x} = f_x x$. W.l.o.g. it also suffices if the trajectory is asymptotically stabilizable, i.e. if there is a matrix-valued function $B(t)$ such that (2.18) is fulfilled for $\dot{x} = (f_x + f_u B)x$. In that case, one can change to new coordinates $u' = u + Bx$.

It is convenient to introduce the notation

$$\widehat{c}(x, u, t) := \mathrm{e}^{-\mu t} c(x, u)$$

for the discounted cost.

We will now show that the naive extrapolation

$$\begin{pmatrix} \widehat{c}_x \\ \widehat{c}_u \end{pmatrix} + Dg^* \cdot \lambda \stackrel{!}{=} 0$$

of (1.5) is a necessary condition for a w.l.o. minimal trajectory.

Note that $\begin{pmatrix}\widehat{c}_x\\ \widehat{c}_u\end{pmatrix} + Dg^* \cdot \lambda \in \mathcal{L}^\infty([0,\infty), \mathcal{X}^* \times \mathcal{U}^*)$ but due to the infinite horizon $\mathcal{L}^\infty([0,\infty), \mathcal{X}^* \times \mathcal{U}^*) \not\subseteq (\mathcal{L}^\infty([0,\infty), \mathcal{X} \times \mathcal{U}))^*$ and so a gradient ∇L with respect to arbitrary variations does generally not exist. However, $\mathcal{L}^\infty([0,\infty), \mathcal{X}^* \times \mathcal{U}^*) \subseteq (\mathcal{L}^\infty_c([0,\infty), \mathcal{X} \times \mathcal{U}))^*$, i.e. there is a gradient w.r.t. compactly supported variations. Its relation to the gradients of the finite horizon problems for the L^T will be used in the following proof.

Theorem 2.18. *Let $(\mathbf{x}^*, \mathbf{u}^*)$ be a w.l.o. minimal trajectory. Then there exists $\lambda \in \mathcal{L}^\infty([0,\infty), \mathcal{X}^*)$ such that*

$$\begin{pmatrix}\widehat{c}_x\\ \widehat{c}_u\end{pmatrix} + Dg^*\lambda = 0.$$

Proof. The proof proceeds in three steps. We begin by proving that the first component of the equation, $\widehat{c}_x + g_x^*\lambda = 0$, already determines a unique candidate for λ. We will then show that on subintervals we have $\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} L^T \approx \widehat{c}_u + g_u^*\lambda$ for this λ. Finally we conclude that if $\frac{\mathrm{d}}{\mathrm{d}\mathbf{u}} L^T \approx \widehat{c}_u + g_u^*\lambda \neq 0$, then an improvement on intervals $[0,T]$ would be possible and violate the w.l.o. minimality of $(\mathbf{x}^*, \mathbf{u}^*)$.

Consider first g_x^*. Proceeding as for finite horizons, we find that $-g_x^*\lambda = \mathbf{y}$,$\mathbf{y} \in \mathcal{L}^\infty([0,\infty), \mathcal{X}^*)$ corresponds to the differential equation

$$-\dot{\lambda} = f_x^\top \lambda + y \text{ on } [0,\infty)$$

and that we lack a boundary condition. However, the requirement that λ be bounded makes $\lambda(t) := \int_t^\infty \Phi^{t\leftarrow s} y(s)\,\mathrm{d}s$, where $\Phi^{s\leftarrow t}$ is the fundamental solution of $-\dot{\lambda} = f_x^\top\lambda$, the unique solution (cf. e.g. [Cop78]). To see this note first that $\lambda(t) = \int_t^\infty \left(\Psi^{s\leftarrow t}\right)^* y(s)\,\mathrm{d}s$ and hence by (2.18), the λ so defined is bounded. It is easily verified that it is indeed a solution. Furthermore, if $\tilde{\lambda} \neq \lambda$ is an additional solution, then $\Delta\lambda := \tilde{\lambda} - \lambda$ fulfills $\Delta\dot{\lambda} = -f_x^\top \Delta\lambda$ and $\Delta\lambda(0) \neq 0$. (2.18) implies $\left\|\Psi^{s\leftarrow 0}\right\|_{\mathcal{X}\to\mathcal{X}} = \left\|\Phi^{0\leftarrow s}\right\|_{\mathcal{X}^*\to\mathcal{X}^*} \xrightarrow[s\to\infty]{} 0$ and hence $\left\|\Phi^{s\leftarrow 0}\right\|_{\mathcal{X}^*\to\mathcal{X}^*} \xrightarrow[s\to\infty]{} \infty$, which means that $\tilde{\lambda}$ cannot be bounded.

Having established the existence of $(g_x^*)^{-1}$, it follows that the $\mathbf{X}$-component of the claimed equality can only be fulfilled by $\lambda = -(g_x^*)^{-1}\widehat{c}_x$ and we will show by contradiction that this λ also fulfills the $\mathbf{U}$-component. To this end, assume for the rest of the proof that $\widehat{c}_u + g_u^* \cdot \lambda = \widehat{c}_u + g_u^* \cdot (-g_x^*)^{-1}\widehat{c}_x \neq 0$ with $\|\widehat{c}_u + g_u^* \cdot \lambda\|_{\mathbf{U}^*} = \alpha > 0$. Then there is $\tau > 0$ such that $\left\|\chi_{[0,\tau]}\left(\widehat{c}_u + g_u^* \cdot \lambda\right)\right\|_{\mathbf{U}^*} > \alpha/2$.

We will now show that $\chi_{[0,\tau]} \frac{d}{d\mathbf{u}} L^T \xrightarrow[T\to\infty]{} \chi_{[0,\tau]}\left(\widehat{c}_u + g_u^*\lambda\right)$ (in the $\mathbf{U}^*$-norm). For the finite horizon subproblem, $\frac{d}{d\mathbf{u}} L^T = \widehat{c}_u^T + (g_u^T)^* \cdot \lambda$ with $\lambda^T = -(g_x^{T*})^{-1}\widehat{c}_x^T$. Obviously $\chi_{[0,\tau]}\widehat{c}_u^T = \chi_{[0,\tau]}\widehat{c}_u$ for $T \geq \tau$. The multiplication operator $g_u^* = f_u^\top$ is bounded and $\chi_{[0,\tau]} g_u^{T*} z = \chi_{[0,\tau]} g_u^* z\ \forall z \in \mathbf{X}^*, T \geq \tau$. Hence it suffices to show $\chi_{[0,\tau]}\lambda^T \xrightarrow[T\to\infty]{} \chi_{[0,\tau]}\lambda$. We have that

$$\lambda(t) = \Phi^{t\leftarrow 0} \int_t^\infty \left(\Psi^{s\leftarrow 0}\right)^* \widehat{c}_x(s)\,\mathrm{d}s$$

and similarly one obtains

$$\lambda^T(t) = \Phi^{t\leftarrow 0} \int_t^T \left(\Psi^{s\leftarrow 0}\right)^* \widehat{c}_x(s)\,\mathrm{d}s$$

$(t \le \tau \le T)$. With (2.18),

$$\left\|\chi_{[0,\tau]}(\lambda - \lambda^T)\right\|_{\mathcal{L}^\infty} \le \int_0^\tau \left\|\Phi^{t\leftarrow 0}\right\|_{\mathcal{X}^*\to\mathcal{X}^*} \mathrm{d}t \cdot \int_T^\infty \left\|\left(\Psi^{s\leftarrow 0}\right)^*\right\|_{\mathcal{X}^*\to\mathcal{X}^*} \mathrm{d}s \cdot \|\widehat{c}_x\|_{\mathcal{L}^\infty} \xrightarrow[T\to\infty]{} 0$$

gives the desired convergence.

We will now construct a $\mathbf{u}$ with lower cost on intervals $[0,T]$. One can find a $T' \ge \tau$, such that $\left\|\chi_{[0,\tau]}\left(\frac{d}{d\mathbf{u}}L^T - \widehat{c}_u - g_u^*\lambda\right)\right\|_{\mathbf{U}^*} \le \frac{\alpha}{8}$ for all $T \ge T'$. Next, consider $\frac{d^2}{d\mathbf{u}^2}L^T(\chi_{[0,\tau]}\cdot, \chi_{[0,\tau]}\cdot)$ and refer to (1.10). It can be seen that $\frac{d^2}{d\mathbf{u}^2}L^T(\chi_{[0,\tau]}\cdot, \chi_{[0,\tau]}\cdot)$ depends on T only through λ^T because all other terms are just restrictions of their infinite horizon counterparts and we already restrict to $[0,\tau]$. With the convergence result for λ^T this yields

$$\left\|\frac{d^2}{d\mathbf{u}^2}L^T(\chi_{[0,\tau]}\cdot, \chi_{[0,\tau]}\cdot)\right\|_{\mathbf{U}\to\mathbf{U}^*} < const$$

independent of T in some neighborhood of $(\mathbf{x}^*, \mathbf{u}^*)$.

Now choose $\Delta\mathbf{u}$, $\mathrm{supp}(\Delta\mathbf{u}) \subseteq [0,\tau]$, $\|\Delta\mathbf{u}\|_{\mathbf{U}} = 1$ such that $\langle \widehat{c}_u + g_u^* \cdot \lambda, \Delta\mathbf{u}\rangle > \frac{\alpha}{4}$ and hence[3] $\left\langle \frac{d}{d\mathbf{u}}L^T, \Delta\mathbf{u}\right\rangle > \frac{\alpha}{8}$ $\forall T \ge T'$. Together with the bound on the second derivative this implies that for any $\delta > 0$ there are sufficiently small $\epsilon > 0$, $\delta > h > 0$ such that $L^T(\mathbf{x}(\mathbf{u}^* + h\Delta\mathbf{u}), \mathbf{u}^* + h\Delta\mathbf{u}) < L^T(\mathbf{x}^*, \mathbf{u}^*) - \epsilon$ for all $T \ge T'$, contradicting the assumption that $(\mathbf{x}^*, \mathbf{u}^*)$ is w.l.o. minimal and completing the proof. □

Remark 2.19. *From the above proof we also see that* $\lambda(t) = \int_t^\infty \Phi^{t\leftarrow s}\widehat{c}_x(s)\,\mathrm{d}s$. *If* $(\mathbf{x},\mathbf{u})$ *converges to a minimum* (x^*, u^*) *of* c, *then* $\widehat{c}_x \to 0$ *and*

$$\lim_{t\to\infty} \lambda(t) = 0.$$

Combining the above Theorem with the admissibility condition, we find that the slight modification

$$F(\mathbf{x}^*, \mathbf{u}^*, \lambda) := \begin{pmatrix} \widehat{c}_x + D_{\mathbf{x}} g^*(\mathbf{x}^*, \mathbf{u}^*) \cdot \lambda \\ \widehat{c}_u + D_{\mathbf{u}} g^*(\mathbf{x}^*, \mathbf{u}^*) \cdot \lambda \\ g(\mathbf{x}^*, \mathbf{u}^*) \end{pmatrix} = 0. \tag{2.19}$$

of (1.6) also characterizes candidates for w.l.o. minimal trajectories although $(\widehat{c}_x, \widehat{c}_u)$ can no longer be understood as the derivative of L in the same sense. The calculations of section 1.4 still hold, if L_x is replaced with $\widehat{c}_x$ and so on.

[3] Restriction to the proper interval is understood to be implicitly done where necessary.

Existence of minimizers

We show the existence of minimizers for $c \geq 0$ and $L^* < \infty$. In this situation, weak overtaking minimality is equivalent to minimality in the normal sense:

Lemma 2.20. *Let $c \geq 0$ and assume $L^* < \infty$. Then $(\mathbf{x}^*, \mathbf{u}^*)$ is weakly overtaking minimal if and only if it is a minimizer.*

Proof. Note first that for $S \leq T$ we have $L^S(\mathbf{x}, \mathbf{u}) \leq L^T(\mathbf{x}, \mathbf{u}) \leq L(\mathbf{x}, \mathbf{u})$ and that $L^T(\mathbf{x}, \mathbf{u}) \xrightarrow[T\to\infty]{} L(\mathbf{x}, \mathbf{u})$ for all $(\mathbf{x}, \mathbf{u})$.

Assume that $(\mathbf{x}^*, \mathbf{u}^*)$ is w.o. minimal and, by way of contradiction, that there exists $(\mathbf{x}, \mathbf{u})$ with $L(\mathbf{x}, \mathbf{u}) < L(\mathbf{x}^*, \mathbf{u}^*) - \epsilon$ for some $\epsilon > 0$. Choose T such that $L^T(\mathbf{x}^*, \mathbf{u}^*) > L(\mathbf{x}^*, \mathbf{u}^*) - \epsilon/2$. As $(\mathbf{x}^*, \mathbf{u}^*)$ is w.o. minimal, there exists $T' \geq T$ such that $L^{T'}(\mathbf{x}, \mathbf{u}) > L^{T'}(\mathbf{x}^*, \mathbf{u}^*) - \epsilon/2$. It follows that

$$L(\mathbf{x}^*, \mathbf{u}^*) - \epsilon > L(\mathbf{x}, \mathbf{u}) \geq L^{T'}(\mathbf{x}, \mathbf{u}) > L^{T'}(\mathbf{x}^*, \mathbf{u}^*) - \epsilon/2 > L(\mathbf{x}^*, \mathbf{u}^*) - \epsilon,$$

a contradiction.

For the reverse implication, assume that $(\mathbf{x}^*, \mathbf{u}^*)$ is a minimizer. Given any $(\mathbf{x}, \mathbf{u})$, $\epsilon > 0$ and $T > 0$, there is a $T' \geq T$ such that

$$L^{T'}(\mathbf{x}, \mathbf{u}) > L(\mathbf{x}, \mathbf{u}) - \epsilon \geq L(\mathbf{x}^*, \mathbf{u}^*) - \epsilon.$$

□

By also requiring the assumptions of Theorem 2.14, we can now show the existence of a minimizer in $\mathcal{W}^{1,p}_{loc} \times \mathcal{L}^{1,p}_{loc}$.

Theorem 2.21. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be finitely generated with v_j continuous and of linear growth*

$$\|v_j(x)\| \leq \nu(1 + \|x\|) \ \forall j, x,$$

in x for some $\nu > 0$ and let $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $c \geq 0$ be convex in u, continuous in (x, u) and fulfill the coercivity condition

$$c(x, u) \geq \gamma_1 \|u\|^p + \gamma_2 \ \forall x, u$$

for some $p > 1$, $\gamma_1 > 0$ and $\gamma_2 \in \mathbb{R}$.

Then (2.17) has a minimizer $(\mathbf{x}^*, \mathbf{u}^*) \in \mathcal{W}^{1,p}_{loc}([0, \infty)) \times \mathcal{L}^p_{loc}([0, \infty))$ if $L^* < \infty$.

Proof. From $L^S \leq L^T \leq L$ for $S \leq T$ it follows for the respective minima that $L^{S,*} \leq L^{T,*} \leq L^*$ and $L^{T,*} \nearrow \lim_{T\to\infty} L^{T,*} =: L^{\infty,*} \leq L^*$ as $T \to \infty$. (In fact, $L^{\infty,*} = L^*$, but this is not proven yet.)

For $i \in \mathbb{N}$, let $(\mathbf{x}_i, \mathbf{u}_i)$ be the minimizer of L^i, which exists according to Theorem 2.14. We will consider a chain of iteratively defined subsequences of $(\mathbf{x}_i, \mathbf{u}_i)$. Let the sequence itself be the 0th subsequence. For $k \in \mathbb{N}$, select the kth subsequence from the $(k-1)$th subsequence such that its restrictions to $[0, k]$ converge weakly in $\mathcal{W}^{1,p} \times \mathcal{L}^p$ and define

$(\mathbf{x}^*, \mathbf{u}^*)|_{[0,k]}$ to be this limit. This is possible because $L^k(\mathbf{x}_i, \mathbf{u}_i) \leq L^i(\mathbf{x}_i, \mathbf{u}_i) < L^*$ for $i \geq k$ and L^k is coercive. The limit agrees with the previous definition of $(\mathbf{x}^*, \mathbf{u}^*)$ on $[0, k-1]$ because the subsequences are nested.

For $i \geq j$, we have $L^j(\mathbf{x}_i, \mathbf{u}_i) \leq L^i(\mathbf{x}_i, \mathbf{u}_i) \leq L^{i,*} \leq L^{\infty,*}$ and hence, by lower semicontinuity of L^j, that $L^j(\mathbf{x}^*, \mathbf{u}^*) \leq L^{\infty,*} \leq L^*$ for all j. As $L^j(\mathbf{x}^*, \mathbf{u}^*) \xrightarrow[j\to\infty]{} L(\mathbf{x}^*, \mathbf{u}^*)$, we have $L(\mathbf{x}^*, \mathbf{u}^*) \leq L^*$.

□

The result for $p = \infty$ similarly carries over from the finite horizon case.

Theorem 2.22. *If, in addition to the assumptions of Theorem 2.21,*

- *f is differentiable with $\|f_x(x,u,t)\| \leq \nu(1 + \|u\|)\ \forall x, u, t$ for some $\nu \in \mathbb{R}$, and*
- *$\|c_x(x,u,t)\| \leq \gamma(x)(1 + \|u\|^p)\ \forall x, u, t$ for some continuous function γ,*

then $\mathbf{x}^* \in \mathcal{W}^{1,\infty}_{loc}([0,\infty); \mathcal{X})$ and $\mathbf{u}^* \in \mathcal{L}^\infty_{loc}([0,\infty); \mathcal{U})$.

Proof. Theorem 2.21 yields minimizers of L^T in $\mathcal{W}^{1,\infty} \times \mathcal{L}^\infty$. We can then proceed as in the proof of Theorem 2.21 with weak convergence replaced by weak-* convergence as $\mathcal{L}^\infty$ is not reflexive. □

Getting a globally bounded minimizer, however, is more difficult.

In the infinite horizon case, trajectories with finite cost may still be unbounded if the discount factor decreases faster than c increases, e.g. for the system $x_0 = 1$, $f(x,u) = f(x) = x$, $c(x,u) = x^2 + u^2$ with any $\mu > 1$. Even without discounting, we require coercivity in x to prevent unbounded solutions, as the following counterexample shows:

Example 2.23. *For $x, u \in \mathbb{R}$ let $c(x,u) = e^{-x} + u^2$, $f(x,u) = ux$, $x_0 = 1$ and $\mu = 0$. $x(t) := t + 1$, $u(t) := \frac{1}{t+1}$ is admissible and has a finite cost, so $L^* < \infty$. It follows that any minimizer would have $c \to 0$ and consequently $|x| \to \infty$ as $t \to \infty$.*

With coercivity in x and no discount, we can obtain global bounds:

Theorem 2.24. *Assume that, in addition to the assumptions of Theorem 2.22,*

$$c(x,u) \geq \gamma_3 \|x\| + \gamma_4\ \forall x, u$$

for some $\gamma_3 > 0, \gamma_4 \in \mathbb{R}$.

Then the undiscounted problem (2.17) with $\mu = 0$ has a minimizer $(\mathbf{x}^*, \mathbf{u}^*) \in \mathcal{W}^{1,p}_{loc}([0,\infty)) \times \mathcal{L}^p_{loc}([0,\infty))$ if $L^* < \infty$.

Proof. We first bound $\|x^*(t)\|$.

Fix $d > 1$ and assume $\|x^*(t_1)\| \geq d$ for some t_1. Set $t_2 := t_1 + \frac{\ln d}{8\nu\sqrt{d}}$. If $\int_{t_1}^{t_2} \|u^*(t)\|_1 \,\mathrm{d}t \leq \frac{\ln d}{8\nu}$, we set $y(t) = \|x^*(t)\|$ and observe that $y(t_1) = d$ and $\dot{y} \geq -2\nu y - 2\nu \|u^*(t)\|_1 y$ as long as $y(t) \geq 1$. With Gronwall's lemma, it follows that

$$y(t_2) \geq y(t_1) \exp(-2\nu(t_2 - t_1) - 2\nu \int_{t_1}^{t_2} \|u^*(t)\|_1 \,\mathrm{d}t) \geq \sqrt{d}$$

and hence, in this case, $L(\mathbf{x},\mathbf{u}) \geq \frac{\ln d}{8\nu}\left(\gamma_3 + \frac{\gamma_4}{\sqrt{d}}\right)$ by coercivity in x and using only contributions from $[t_1, t_2]$.

Due to $p > 1$ and the equivalence of finite-dimensional norms, we also have

$$c(x,u) \geq \gamma_5 \|u\|_1 + \gamma_6 \ \forall x, u$$

for some constants $\gamma_5 > 0$, $\gamma_6 \in \mathbb{R}$. If $\int_{t_1}^{t_2} \|u^*(t)\|_1 \,\mathrm{d}t > \frac{\ln d}{8\nu}$, we use this coercivity to obtain $L(\mathbf{x},\mathbf{u}) \geq \frac{\ln d}{8\nu}\left(\gamma_5 + \frac{\gamma_6}{\sqrt{d}}\right)$.

It follows that $L(\mathbf{x}^*,\mathbf{u}^*) \geq \min\left\{\frac{\ln d}{8\nu}\left(\gamma_3 + \frac{\gamma_4}{\sqrt{d}}\right), \frac{\ln d}{8\nu}\left(\gamma_5 + \frac{\gamma_6}{\sqrt{d}}\right)\right\} \xrightarrow[d\to\infty]{} \infty$ and so $x^*(t)$ has to be bounded if $L^* < \infty$.

The bound on $\|u^*(t)\|$ (and thereby $\|\dot{x}^*(t)\|$) is analogous to Theorem 2.16. The only difference is that $\|\widehat{x}(t) - x^*(t)\|$ can no longer be bounded by Gronwall's inequality, so instead we use (2.18) to obtain

$$\begin{aligned}
&\int_0^\infty \|\widehat{x}(t) - x^*(t)\| \,\mathrm{d}t \\
&\lesssim \int_0^\infty \int_0^t \chi_{E_\epsilon}(s) \|f(\widehat{x},0) - f(x^*,u^*)\| \left\|\Psi^{t\leftarrow s}\right\|_{\mathcal{X}\to\mathcal{X}} \mathrm{d}s \,\mathrm{d}t \\
&= \int_0^\infty \chi_{E_\epsilon}(s) \left(\|f(\widehat{x},0)\| + \|f(x^*,u^*)\|\right) \int_s^\infty \left\|\Psi^{t\leftarrow s}\right\|_{\mathcal{X}\to\mathcal{X}} \mathrm{d}t \,\mathrm{d}s \\
&\leq \left(\epsilon\kappa \sup_s \|f(\widehat{x}(s),0)\| + \kappa\right)\beta.
\end{aligned}$$

with the constant β from (2.18) and "$\lesssim$" as $\epsilon \to 0$. □

3. Singularities and bifurcations

In this chapter, we introduce some types of singularities and bifurcations that are relevant for our study of the optimal control problem. All results in this chapter are standard. A more detailed exposition of bifurcation theory can e.g. be found in [Kie06] or [GG73].

The optimal control problem does not appear in this chapter and we will reuse some variables otherwise associated with it.

Let X and Y be real Banach spaces.

Definition 3.1. *A bounded linear operator $A : X \to Y$ is a* Fredholm operator *of* index $m \in \mathbb{Z}$ *if* $\dim \mathcal{N}(A) < \infty$, $\operatorname{codim} \mathcal{R}(A) < \infty$ *and* $m = \dim \mathcal{N}(A) - \operatorname{codim} \mathcal{R}(A)$.

Lemma 3.2 ([Kat58, Lemma 332])**.** *The range of a Fredholm operator $A : X \to Y$ is a closed subspace of Y.*

Lemma 3.3 ([Kat58, Remark 1])**.** *The index of a Fredholm operator is invariant under bounded perturbations (with respect to the operator norm).*

Let $G : X \to Y$ be continuously (again w.r.t. the operator norm) Fréchet-differentiable with $DG(x)$ a Fredholm operator of index n for all $x \in X$. Because of Lemma 3.3, the latter condition is fulfilled if it holds for any $x \in X$.

Let $\mathcal{S}$ be the set of solutions to $G = 0$, i.e.

$$\mathcal{S} := \{x \in X : G(x) = 0\}.$$

At a generic point x, $DG(x)$ has full rank in the sense that

$$\begin{aligned} \dim \mathcal{N}(DG(x)) &= n, \\ \operatorname{codim} \mathcal{R}(DG(x)) &= 0 \end{aligned} \tag{3.1}$$

and $\mathcal{S}$ is locally an n-dimensional manifold. In the following, we will describe some non-generic situations.

3.1. Fold singularity

First, we study points where $\mathcal{S}$ is still locally a manifold, but where a specific parametrization is not possible.

Let X have a decomposition $x = (\lambda, z) \in \mathbb{R}^n \times Z$ such that $D_z G$ is "square", i.e. has Fredholm index 0. In this situation, λ is often interpreted as a parameter and one wants to find a solution $z(\lambda)$ of $G(\lambda, z) = 0$. Locally z must fulfill $D_z G \, \mathrm{d}z = -D_\lambda G \, \mathrm{d}\lambda$ and we

call points where solutions to this equation may not exist (because $\mathcal{R}(D_\lambda G) \not\subseteq \mathcal{R}(D_z G)$) fold singularities.

Definition 3.4. *Let G as above. A point (λ^*, z^*) with $G(\lambda^*, z^*) = 0$ is a* fold singularity *if there exists $w \in Y$, $w \neq 0$, such that $w \notin \mathcal{R}(D_z G(\lambda^*, z^*))$ and $w \in \mathcal{R}(D_\lambda G(\lambda^*, z^*))$. It is a* simple fold singularity *if* $\operatorname{codim} \mathcal{R}(D_z G(\lambda^*, z^*)) = 1$.

Note that if $D_z G$ is singular but $\mathcal{R}(D_\lambda G) \subseteq \mathcal{R}(D_z G)$, we have a more complicated singularity where multiple solutions may exist. One possible type in this case is the pitchfork bifurcation described in the next section.

At a fold singularity, there exists at least one direction in the parameter space for which locally no solution exists:

Lemma 3.5. *Let $\Pi_\lambda : (\lambda, z) \mapsto \lambda$ be the projection onto the λ-coordinate. If (λ^*, z^*) is a fold singularity, then $\Pi_\lambda T_{\mathcal{S}}(\lambda^*, z^*) \neq \mathbb{R}^n$.*

Proof. Let w as in Definition 3.4. Then there exists a $\lambda \in \mathbb{R}^n$ with $D_\lambda G(\lambda^*, z^*)\lambda = w$ and a functional $v \in Y^*$ with $v \perp \mathcal{R}(D_z G^*(\lambda^*, z^*))$ and $v(w) = 1$. If $\lambda \in \Pi_\lambda T_{\mathcal{S}}(\lambda^*, z^*)$, then there would be $(\lambda, z) \in T_{\mathcal{S}}(\lambda^*, z^*)$ and we would have

$$\begin{aligned} 0 &= D_\lambda G(\lambda^*, z^*)\lambda + D_z G(\lambda^*, z^*)z \\ &= v(D_\lambda G(\lambda^*, z^*)\lambda) + v(D_z G(\lambda^*, z^*)z) \\ &= v(D_\lambda G(\lambda^*, z^*)\lambda) = v(w) = 1. \end{aligned}$$

It follows that $\lambda \notin \Pi_\lambda T_{\mathcal{S}}(\lambda^*, z^*)$. □

Example 3.6. *$G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (\lambda, x) \mapsto \lambda + x^2$ has a fold singularity at $(0, 0)$ with $DG(0, 0) = (0, 0)$ and indeed there are no solutions $x \in \mathbb{R}$ of $G(\lambda, x) = 0$ for $\lambda > 0$, cf. Figure 3.1.*

3.2. Pitchfork singularity

Next, we consider the situation where $DG(x^*)$ is rank-deficient by one. To simplify the presentation, we require that the Fredholm index is one, which is the only case that will occur in the subsequent chapters.

$\mathcal{S}$ would then generically be a curve, but at x^* will be the intersection of two curves forming a *pitchfork bifurcation* as shown in Figure 3.2.

Unlike a fold singularity, a pitchfork bifurcation is not a matter of parameterization but a change in the topological structure of $\mathcal{S}$. Consequently, we will not introduce a special parameter subspace of X. As the precise requirements are not immediately apparent, we will postpone the definition of a pitchfork singularity to the end of this section and begin now the analysis of points with rank-deficiencies.

Let $x^* \in \mathcal{S}$ with

$$\begin{aligned} \dim \mathcal{N}(DG(x^*)) &= 2 \\ \operatorname{codim} \mathcal{R}(DG(x^*)) &= 1. \end{aligned} \tag{3.2}$$

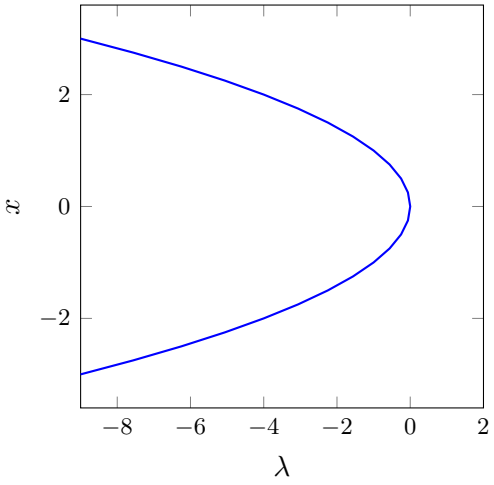

Figure 3.1.: Fold singularity of $G(\lambda, x) = \lambda + x^2$.

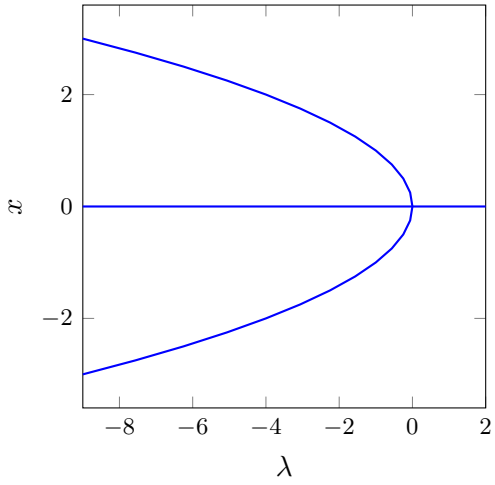

Figure 3.2.: Pitchfork singularity of $G(x) = x_1 x_2 + x_2^3$.

and normalized bases

$$\mathcal{N}(DG(x^*)) = \operatorname{span}\{\Phi_1, \Phi_2\} \subset X$$
$$\mathcal{R}(DG(x^*))^\perp = \operatorname{span}\{\Psi\} \subset Y^*.$$

There exists a complement $\hat{X}$ of $\mathcal{N}(DG(x^*))$ in X, so that we have the decomposition

$$X = \hat{X} \oplus \operatorname{span}\{\Phi_1, \Phi_2\}.$$

3.2.1. Lyapunov-Schmidt reduction

Note that the restriction of $DG(x^*)$ to $\hat{X} \to \mathcal{R}(DG(x^*))$ is (boundedly) invertible by the Open Mapping Theorem. Using the above decompositions we can split $DG(x^*)$ into an invertible part and a finite-dimensional part, which controls the bifurcation. We write $x = x^* + \hat{x} + \alpha_1\Phi_1 + \alpha_2\Phi_2$ with $\hat{x} \in \hat{X}$ and have

$$G(x) = 0 \Leftrightarrow \begin{cases} P_{\mathcal{R}(DG(x^*))} G(x^* + \hat{x} + \alpha_1\Phi_1 + \alpha_2\Phi_2) = 0 \\ \quad \text{and } \Psi\left(G(x^* + \hat{x} + \alpha_1\Phi_1 + \alpha_2\Phi_2)\right) = 0 \end{cases}, \tag{3.3}$$

with $P_{\mathcal{R}(DG(x^*))}$ an arbitrary projection onto $\mathcal{R}(DG(x^*))$.

By the Inverse Function Theorem, the first equation can be solved locally around x^* for $\hat{x} = \hat{x}(\alpha_1, \alpha_2)$.

Note that $\hat{x}(\alpha_1, \alpha_2) = -\left(DG(x^*)|_{\hat{X} \to \mathcal{R}(DG(x^*))}\right)^{-1} DG(x^*)(\alpha_1\Phi_1 + \alpha_2\Phi_2) + \mathcal{O}(\|\alpha\|^2) = 0 + \mathcal{O}(\|\alpha\|^2)$ (for $\alpha \to 0$) since $\Phi_i \in \mathcal{N}(DG(x^*))$.

Inserting into the second equation on the right in (3.3) and Taylor expansion yields

$$\sum_{i,j=1}^{2} \alpha_i \alpha_j \Psi((D^2G)(x^*)(\Phi_i, \Phi_j)) + r(\alpha) = 0 \tag{3.4}$$

with $r(\alpha) = \mathcal{O}(\|\alpha\|^3)$.

The following homotopy argument shows that the quadratic form

$$\sum_{i,j=1}^{2} \alpha_i \alpha_j \Psi((D^2G)(x^*)(\Phi_i, \Phi_j)) =: \alpha^\top A \alpha$$

determines the local structure of the solution set if $A = \left(\Psi((D^2G)(x^*)(\Phi_i, \Phi_j))\right)_{i,j=1}^{2}$ is non-singular: (The result is a variation of the Morse Lemma, cf. [Mor25])

Lemma 3.7. *Let $f : \mathbb{R}^n \to \mathbb{R}, f(x) := x^\top Ax + r(x)$ with $A \in \mathbb{R}^{n\times n}$ non-singular, $r \in \mathcal{C}^1(\mathbb{R}^n)$, $r(0) = 0$ and $\|Dr(x)\| = o(\|x\|)$ as $x \to 0$.*

Let $\mathcal{S}_0 := \{x \in \mathbb{R}^n : x^\top Ax = 0\}$ and $\mathcal{S}_1 := \{x \in \mathbb{R}^n : f(x) = 0\}$.

Then there exist open neighborhoods $U, V \subset \mathbb{R}^n$ of 0 and a diffeomorphism $\varphi : U \to V$ such that

$$x^\top Ax = f(\varphi(x)) \ \forall x \in U, \text{ in particular } \varphi(U \cap \mathcal{S}_0) = V \cap \mathcal{S}_1$$

and

$$\|x - \varphi(x)\| = o(\|x\|) \ (x \to 0).$$

Proof: We use the homotopy $f_\lambda(x) := x^\top Ax + \lambda r(x)$, $\lambda \in [0, 1]$ and construct a flow $F(x, \lambda)$ that leaves f_λ invariant. Overloading notation, $x(\lambda)$ shall now denote trajectories of $\frac{dx}{d\lambda} = F(x, \lambda)$. From the requirement $f_\lambda(x(\lambda)) \overset{!}{\equiv} const.$, we obtain

$$0 = \frac{d}{d\lambda} f_\lambda(x) = \frac{\partial}{\partial \lambda} f_\lambda + \frac{\partial}{\partial x} f_\lambda \cdot \frac{dx}{d\lambda} = r(x) + (2Ax + \lambda Dr(x)) \cdot \frac{dx}{d\lambda}$$

$$\Rightarrow \left\langle 2Ax + \lambda Dr(x), \frac{dx}{d\lambda} \right\rangle = -r(x)$$

Choosing $\left\|\frac{dx}{d\lambda}\right\|$ to be minimal yields

$$F(x, \lambda) := \frac{-r(x)}{\|2Ax + \lambda Dr(x)\|^2}(2Ax + \lambda Dr(x)).$$

As A is non-singular, we have $\|2Ax + \lambda Dr(x)\| = \Omega(\|x\|)$ and hence, for some neighborhood U of 0 and $x \in U\backslash\{0\}$, $F(x, \lambda)$ is well-defined and, as $x \to 0$,

$$\|F(x, \lambda)\| = \frac{o(\|x\|^2)}{\Omega(\|x\|)} = o(\|x\|).$$

It follows that F can be continuously extended to $F(0,\cdot) := 0$ and, by Gronwall's Lemma,

$$\|x(1) - x(0)\| = \left\| \int_0^1 \frac{dx}{d\lambda} d\lambda \right\| = o(\|x(0)\|).$$

The evolution $\varphi : x(0) \mapsto x(1), U \to \varphi(U) =: V$ of $\frac{dx}{d\lambda} = F(x,\lambda)$ is the claimed diffeomorphism. □

3.2.2. Analysis of the quadratic form

We distinguish three cases for the quadratic form $A = \left(\Psi((D^2G)(x^*)(\Phi_i,\Phi_j))\right)_{i,j=1}^2$:

1) **A is definite** ($\det A > 0$, "$x_1^2 + x_2^2 = 0$") x_0 is an isolated solution, contradicting the assumption that it is part of a solution curve.

2) **A is singular** ($\det A = 0$, "$x_1^2 + x_2^3 = 0$") The assumption of Lemma 3.7 is not satisfied. We have a degenerate bifurcation point where the structure of the solution set is determined by higher order terms.

3) **A is indefinite and non-singular** ($\det A < 0$, "$x_1^2 - x_2^2 = 0$") The solution set consists of two curves intersecting at x_0.

Only in case 3 do we have a pitchfork bifurcation. Diagonalization of the symmetric matrix A gives

$$A = T \begin{pmatrix} \mu_1 & 0 \\ 0 & -\mu_2 \end{pmatrix} T^\top,$$

with $\mu_1, \mu_2 > 0$, T orthonormal and $\alpha^\top A \alpha = 0$ has the solutions $\alpha = T \begin{pmatrix} \sqrt{\mu_2} \\ \pm\sqrt{\mu_1} \end{pmatrix}$.

Recalling $x = x^* + \hat{x} + \alpha_1\Phi_1 + \alpha_2\Phi_2$ (and $\hat{x} = \mathcal{O}(\|\alpha\|^2)$) we obtain that the tangents of the solution curves at x^* are

$$\begin{pmatrix} \Phi_1 & \Phi_2 \end{pmatrix} T \begin{pmatrix} \sqrt{\mu_2} \\ \pm\sqrt{\mu_1} \end{pmatrix}. \tag{3.5}$$

The above discussion finally leads to the following definition:

Definition 3.8. *A point x^* with $G(x^*) = 0$ is a* pitchfork singularity *if it fulfills* (3.2) *and* $\det\left(\Psi((D^2G)(x^*)(\Phi_i,\Phi_j))_{i,j=1}\right)^2 < 0$.

Example 3.9. *Let $G : \mathbb{R}^2 \to \mathbb{R}, x \mapsto x_1x_2 + x_2^3$ (cf. Figure 3.2). At $x^* = (0,0)$ we have $DG = (x_2, x_1 + 3x_2^2) = (0,0)$ and $D^2G = \begin{pmatrix} 0 & 1 \\ 1 & 6x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. We can use $\Phi_1 = (1,0)^\top$, $\Phi_2 = (0,1)^\top$ and $\Psi = 1$. Then $A = D^2G$ and $\det A = -1$, i.e. x^* is a pitchfork singularity.*

3.3. Cusp singularity

The most complicated singularity we need to consider is a combination of the previous two: A cusp singularity is a pitchfork bifurcation where a branch of fold singularities with respect to an additional parameter $\lambda \in \mathbb{R}^n$ appears simultaneously with and tangentially to the second solution branch.

Definition 3.10. *Let $(\lambda^*, \mu^*, z^*) \in X = \mathbb{R}^n \times \mathbb{R} \times Z$ such that $D_{(\mu,Z)}G(\lambda^*, \mu^*, z^*)$ has Fredholm index 1, (μ^*, z^*) is a pitchfork singularity of $G(\lambda^*, \cdot, \cdot)$ and one solution branch has the tangent $(0, v)$, $0 \neq v \in Z$. Let*

$$\mathfrak{S} := \{(\lambda, \mu, z) \in X : (\lambda, z) \text{ is a fold singularity of } G(\cdot, \mu, \cdot)\}.$$

Then (λ^*, μ^*, z^*) is a cusp singularity *if $(0, 0, v)$ is a tangent of $\mathfrak{S}$ at (λ^*, μ^*, z^*).*

Example 3.11. *Let $G : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (\lambda, \mu, z) \mapsto 4z^3 + 2\mu z + \lambda$ (cf. Figure 3.3). At $(\lambda^*, \mu^*, z^*) = (0, 0, 0)$ we have $D_{(\mu,z)}G = (2z, 2\mu + 12z^2) = (0, 0)$ and $D^2_{(\mu,z)}G = \begin{pmatrix} 0 & 2 \\ 2 & 24z \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$, so it is a pitchfork singularity w.r.t. (μ, z) with tangents $(1, 0)$ and $(0, 1)$, or $(0, 1, 0)$ resp. $(0, 0, 1)$ if embedded in the full space.*

Furthermore, $(8z^3, -6z^2, z), z \in \mathbb{R}$ is a curve of fold singularities as it fulfills $G = 0$ and $D_z G = 0$. At $(0, 0, 0)$, its tangent is $(0, 0, 1)$, equaling the second tangent of the pitchfork bifurcation. Hence $(0, 0, 0)$ is a cusp singularity.

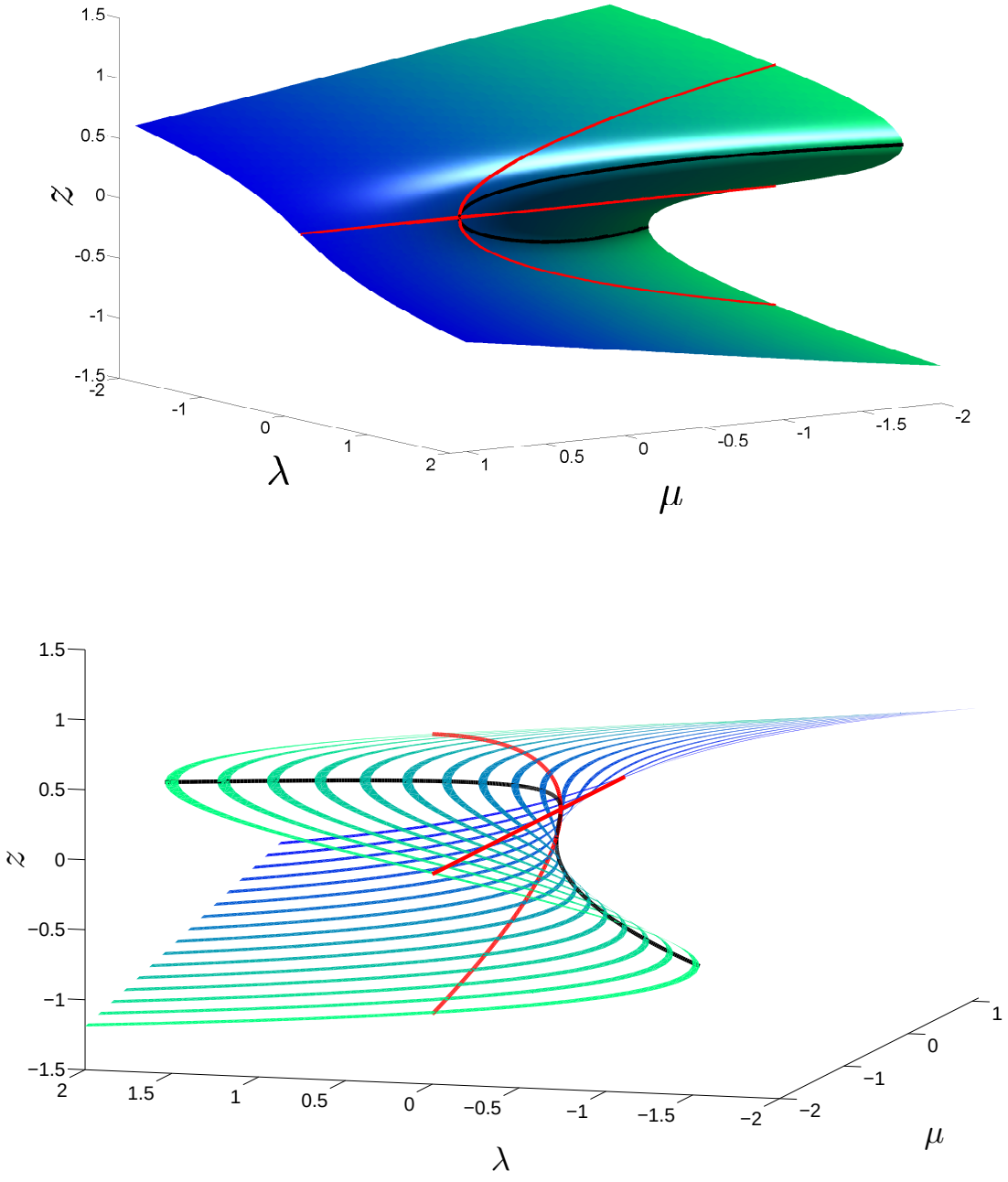

Figure 3.3.: Two views of the cusp singularity of $G(\lambda, \mu, z) = 4z^3 + 2\mu z + \lambda$. The intersection of $\mathcal{S}$ with the (μ, z)-plane, which contains the pitchfork bifurcation, is red. The curve of fold singularities is black.

4. Geometry of optimal control

Having discussed the criteria for identifying a local optimum, we now want to look at the structure of all local optima with the eventual goal of finding the global one among them.

The existence of multiple local optima is closely related to the appearance of nonsmooth regions in the value function. We will explore this connection, at first for finite time horizon problems as in Section 2.1. Recall in particular that the cost functional

$$L(\mathbf{x}, \mathbf{u}) = \int_0^T c(x(t), u(t), t)\,\mathrm{d}t + \varphi(x(T))$$

includes a terminal cost φ.

We are interested in the number of extremal trajectories that are associated with (i.e. start at) an initial condition $x(t_0) = x_0$, but will begin by considering the set of *all* extremals for all initial conditions before returning to the question whether there is a (locally) unique one for a given $t_0 \in [0, T]$ and $x_0 \in \mathcal{X}$.

Restating (2.7), we know that $(\mathbf{x}, \mathbf{u}, \lambda)$ are an extremal trajectory and its adjoint if and only if they fulfill

$$\left.\begin{aligned} \dot{x} &= f, \\ -\dot{\lambda} &= c_x + f_x^\top \lambda, \\ 0 &= c_u + f_u^\top \lambda \end{aligned}\right\} \tag{4.1}$$

and the terminal condition

$$\lambda(T) = \frac{\mathrm{d}}{\mathrm{d}x}\varphi(x(T)). \tag{4.2}$$

We will assume from now on that $c_u = -f_u^\top \lambda$ uniquely determines $u = u^*(x, \lambda)$, which e.g. is the case if c is strictly convex in u and f is finitely generated. With this assumption, (4.1) is an ODE in (x, λ).

In this chapter, we will study its solutions and their variations. We assume throughout that a minimizer exists in $\mathcal{W}^{1,\infty} \times \mathcal{L}^\infty$ and that f, c, φ have a second derivative with uniform bounds on compact sets as in Theorem 2.4, so that the variations exist.

4.1. Non-uniqueness of extremals trajectories

The study of extremals is greatly simplified by the following observation:

Lemma 4.1. *The mapping $\Psi : x_T \mapsto (\mathbf{x}, \lambda)$, where $(\mathbf{x}, \lambda)$ is the solution of (4.1) with terminal condition $x(T) = x_T$ and (4.2), is a continuous bijection between $\mathcal{X}$ and the set*

of extremal trajectories (with the $\mathcal{W}^{1,\infty}$-norm).

Proof. It follows from the above that Ψ produces extremal trajectories and its injectivity is obvious. Under the assumptions of this chapter, solutions to (4.1) vary continuously with the initial conditions and are unique. Hence Ψ is continuous and surjective, as $\Psi(x(T)) = (\mathbf{x}, \lambda)$ for any extremal pair $(\mathbf{x}, \lambda)$. □

In light of this observation, we will study bifurcations in the (possibly multi-valued) mapping $x(t) \mapsto x(T)$. This involves looking at its derivative, so we further introduce $Q(t)$ via the variation $dx(t) := Q(t)dx(T)$, i.e. $Q(t) = \frac{\mathrm{d}x(t)}{\mathrm{d}x(T)}$.

We have the following relationship between λ and the derivative of the value function V:

Lemma 4.2. *If $(\mathbf{x}^*, \mathbf{u}^*, \lambda^*)$ are an extremal starting at $x(t_0) = x_0$ and its adjoint, and there exists a neighborhood $\Omega \subseteq \mathcal{X}$ of x_0 such that $(\mathbf{x}^*, \mathbf{u}^*)$ and its variations with initial conditions $x(t_0) = x \in \Omega$ are global minima then*

$$V_x(x_0, t_0) = \lambda^*(t_0).$$

Proof. Let the extremal $(\mathbf{x}, \mathbf{u})$ be a variation of $(\mathbf{x}^*, \mathbf{u}^*)$. By the Shadow Price Theorem 1.8 we have for $L(\mathbf{x}, \mathbf{u})$ at $(\mathbf{x}, \mathbf{u}) = (\mathbf{x}^*, \mathbf{u}^*)$

$$\frac{\mathrm{d}L}{\mathrm{d}x_0} = \lambda^* \cdot \frac{\mathrm{d}g}{\mathrm{d}x_0} = \lambda^* \cdot \delta_{t_0} = \lambda^*(t_0).$$

By definition of the value function, $L(\mathbf{x}, \mathbf{u}) = V(x(t_0), t_0)$ for $x(t_0) \in \Omega$ and so $V_x(x_0, t_0) = \frac{\mathrm{d}L}{\mathrm{d}x_0}(\mathbf{x}^*, \mathbf{u}^*) = \lambda^*(t_0)$. □

Under the same assumption, $S(t)$ defined by $d\lambda(t) = S(t)dx(t)$ is the second derivative: $S(t) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}V(x(t), t)$.

We will show that there is a relation between S and bifurcation of optima.

Let us consider an infinitesimal variation $(d\mathbf{x}, d\mathbf{u}, d\lambda)$ of an extremal trajectory $(\mathbf{x}, \mathbf{u}, \lambda)$.[1] Such a trajectory is a solution of (1.6), and so its variation solves (1.11). Note that the operators appearing in (1.11) act locally in t. As only the boundary condition is varied, the right hand side of (1.11) is 0 on the interval $(0, T]$. Hence, by solving the third resp. second row, we find that the following holds point-wise for almost all $t \in (0, T)$:

$$\begin{aligned} 0 &= g_x dx + g_u du = f_x dx - d\dot{x} + f_u du \\ &\Rightarrow d\dot{x} = f_x dx + f_u du \\ 0 &= H_{ux} dx + H_{uu} du + g_u^* d\lambda \;\Rightarrow\; du = -H_{uu}^{-1}(H_{ux} dx + f_u^\top d\lambda). \end{aligned}$$

Hence

$$d\dot{x} = (f_x - f_u H_{uu}^{-1} H_{ux})dx - f_u H_{uu}^{-1} f_u^\top d\lambda.$$

[1] The following calculations are standard in optimal control theory, cf. e.g. [BH69, Chapter 6]

Furthermore, from the first row,

$$\begin{aligned} 0 &= H_{xx}dx + H_{xu}du + g_x^* d\lambda = H_{xx}dx + H_{xu}du + d\dot{\lambda} + f_x^\top d\lambda \\ \Rightarrow d\dot{\lambda} &= -H_{xx}dx - H_{xu}du - f_x^\top d\lambda \\ &= (-H_{xx} - H_{xu}H_{uu}^{-1}H_{ux})dx + (-f_x^\top + H_{xu}H_{uu}^{-1}f_u^\top)d\lambda. \end{aligned}$$

So dx and $d\lambda$ fulfill the linear ODE

$$\left.\begin{aligned} d\dot{x} &= A(t)dx - B(t)d\lambda \\ d\dot{\lambda} &= -C(t)dx - A(t)^\top d\lambda \end{aligned}\right\} \tag{4.3}$$

with $A(t) := f_x - f_u H_{uu}^{-1} H_{ux}$, $B(t) := f_u H_{uu}^{-1} f_u^\top$ and $C(t) := H_{xx} + H_{xu}H_{uu}^{-1}H_{ux}$.

From (4.3) we derive differential equations for Q and S:

$$\begin{aligned} d\dot{\lambda} &= \dot{S}dx + Sd\dot{x} \\ d\dot{x} &= Adx - Bd\lambda = Adx - BSdx \\ \dot{Q}(t)dx(T) &= d\dot{x}(t) = (A - BS)dx(t) \\ &= (A - BS)Q(t)dx(T) \qquad \forall dx(T) \\ \Rightarrow \dot{Q} &= (A - BS)Q \\ d\dot{\lambda} &= -Cdx - A^\top d\lambda = -(C + A^\top S)dx = \dot{S}dx + Sd\dot{x} \\ &= (\dot{S} + S(A - BS))dx \\ \dot{S}dx &= (-C - A^\top S - SA + SBS)dx \end{aligned}$$

The last line holds for all $dx(t) = Q(t)dx(T)$ and so S fulfills

$$\dot{S} = SBS - C - A^\top S - SA, \tag{4.4}$$

provided $Q(t)$ is invertible. The equations for Q and S can be integrated backwards form T with the terminal conditions being

$$Q(T) = Id,$$

$$S(T) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} V(x(T), T) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \varphi(x(T)). \tag{4.5}$$

However, the solution can cease to exist at some point, either because Q becomes singular or because the quadratic equation for S undergoes a blow-up. Let us recall what these events mean for the optimal control problem:

- *Q is related to the existence of multiple extrema.*

  $Q(t)$ is the derivative of the mapping $x(T) \mapsto x(t)$. If $Q(T_Q)$ is singular, the Implicit Function Theorem no longer guarantees the existence of a mapping $x(t) \mapsto x(T)$ and, due to Lemma 4.1, there can then be multiple extremal trajectories starting

at $x(T_Q)$.

- *S is related to the smoothness of the value function.*

  We have $S(t) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} V(x(t), t)$, so $\|S\| \to \infty$ as $t \nearrow T_S$ implies that V has no second derivative at $(x(T_S), T_S)$ (if the corresponding extremal is globally optimal).

According to the following Theorem, these events can only occur simultaneously.

Theorem 4.3. *Let $(T_S, T]$, $T_S \in [-\infty, T)$ resp. $(T_Q, T]$, $T_Q \in [-\infty, T)$ be the maximal intervals on which a solution to (4.4) exists, resp. on which $Q(t)$ defined by $x(t) = Q(t)x(T)$ is nonsingular. Then $T_S = T_Q$ and $Q(t)$ exists for all t.*

Proof. Define the solution operator of (4.3) via

$$\begin{pmatrix} x(t) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} D(t) & E(t) \\ F(t) & G(t) \end{pmatrix} \begin{pmatrix} x(T) \\ \lambda(T) \end{pmatrix}.$$

As (4.3) is a linear ODE with Lipschitz coefficients, this solution always exists and in particular D, E, F and G cannot blow up. We have

$$\begin{aligned} \lambda(T) &= S(T)x(T), \\ x(t) &= D(t)x(T) + E(t)\lambda(T) = \underbrace{(D(t) + E(t)S(T))}_{=Q(t)} x(T), \\ \lambda(t) &= F(t)x(T) + G(t)\lambda(T) = \underbrace{(F(t) + G(t)S(T))Q(t)^{-1}}_{=S(t)} x(t). \end{aligned} \tag{4.6}$$

The equation for $x(t)$ provides $Q(t)$ for all t.

We have $\limsup_{t \to T_S} \|S(t)\| = \infty$ because otherwise S would remain in a compact set and so (4.4) would have a Lipschitz-continuous r.h.s. and hence a solution extending beyond T_S. The equation for $\lambda(t)$ shows that S blowing up is equivalent to Q becoming singular and therefore $T_S = T_Q$. □

Hence we have the following Theorem for the existence of V_{xx}:

Theorem 4.4. *If $Q(t)$ is non-singular for all extremals and for all $t \in [t^*, T]$ for some $t^* \in [0, T]$, then $V(x, t)$ is twice differentiable in x for all $x \in \mathcal{X}$ and all $t \in [t^*, T]$.*

Proof. We begin by showing that the local invertibility of the mapping $x_T \mapsto x(t)$ (which follows from Q non-singular) implies its global invertibility. By way of contradiction, assume that for some $\tau \in [t^*, T]$ there exist $x_0 \in \mathcal{X}$ and $x_T(\tau), \widehat{x}_T(\tau) \in \mathcal{X}$ with $x_T(\tau) \neq \widehat{x}_T(\tau)$ such that (x_0, t) is on both the extremals starting at $x_T(t)$ resp. $\widehat{x}_T(t)$ for $t = \tau$. We can extend x_T and $\widehat{x}_T$ to $[\tau, T]$ such that this also holds for $t \in (\tau, T]$ by setting

$$\dot{x}_T(t) := -Q(t)^{-1}\dot{x}(t),$$

where Q and $\dot{x}$ refer to the extremal starting at $x_T(t)$, and similarly for $\widehat{x}_T$. As $x_T \mapsto x(t)$ is locally invertible, we always have a neighborhood of $x_T(t)$ which $\widehat{x}_T(t)$ cannot enter

and so $x_T(t) \neq \widehat{x}_T(t)$ for all $t \in [\tau, T]$. However, $x_0 = x_T(T) = \widehat{x}_T(T)$, completing the contradiction.

By Theorem 4.3, $S(t)$ exists for all extremals and all $t \in [t^*, T]$. Since $x_T \mapsto x(t)$ is globally invertible, there is a unique extremal for each $x(t)$, which must be the global minimum. Hence, by Lemma 4.2, V_{xx} exists and is given by $V_{xx} = S$. □

Remark 4.5. *Times t at which $Q(t)$ is singular are historically known as* conjugate points *and play a role in determining whether an extremal is a local minimum. One can show that $\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{u}^2}L$ is positive definite for $t_0 = T - \tau$ with τ small enough as $c_{uu} > 0$ dominates the other terms. $\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{u}^2}L$ remains positive definite unless one of its eigenvalues reaches 0, i.e. $\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{u}^2}L$ and therefore Q become singular.*

4.2. Global structure of singularities

Let us define a solution operator $\Upsilon^{t,T}$ for (4.1) by $\Upsilon^{t,T}x_T := x(t)$, where $(x(t), \lambda(t))$ solves (4.1) with the terminal conditions $x(T) = x_T$ and $\lambda(T) = \frac{\mathrm{d}}{\mathrm{d}x_T}\varphi(x_T)$. We further define

$$G(x_t, t, x_T) := \Upsilon^{t,T}x_T - x_t.$$

With this definition, $G(x_t, t, x_T)$ is zero if and only if (t, x_t) lies on the extremal emanating from x_T, and hence our goal is to characterize the singularities of the solution set $\{G = 0\}$

Note that in the optimal control problem we are given an initial condition $x(t) = x_t$ and so we regard x_t and t as parameters, whereas x_T is the variable to be solved for.

The matrices $Q(t)$ introduced above depend on the terminal condition x_T, which we will now make explicit by referring to them as $Q(t, x_T)$. In this section, we will consider the set of singularities of Q,

$$\mathfrak{S} := \{(t, x_T) \in [0, T] \times \mathcal{X} : Q(t, x_T) \text{ is singular}\},$$

and also define the set of simple singularities

$$\mathfrak{S}_1 := \{(t, x_T) \in \mathfrak{S} : \dim \mathcal{N}\left(Q(t, x_T)\right) = 1\}.$$

There is a correspondence between the singularities of Q and G, which is given by the following Lemma:

Lemma 4.6. *(x_t, t, x_T) is a fold singularity (with respect to (x_t, x_T)) of the solution set $\{G = 0\}$ if and only if $\Upsilon^{t,T}x_T = x_t$ and $Q(t, x_T)$ is singular.*

Proof. As mentioned above, (x_t, t, x_T) is in the solution set exactly if $\Upsilon^{t,T}x_T = x_t$. As $\frac{\partial}{\partial x_t}G = -Id$, we always have $\mathcal{R}\left(\frac{\partial}{\partial x_t}G\right) = \mathbb{R}^n$ and so (x_t, t, x_T) is a fold singularity exactly if $\frac{\partial}{\partial x_T}G(x_t, t, x_T)$ is singular. We have $\frac{\partial}{\partial x_T}G(x_t, t, x_T) = Q(t, x_T)$ and so the claim holds. □

Furthermore, generic singularities are also pitchfork singularities with respect to (t, x_T):

Theorem 4.7. *Let* $(t^*, x_T^*) \in \mathfrak{S}_1$, $\mathcal{N}(Q(t^*, x_T^*)) = \operatorname{span}\{v\}$, $\mathcal{R}(Q(t^*, x_T^*))^\perp = \operatorname{span}\{w\}$ *and define*

$$\widetilde{G}(t, x_T) := G(\Upsilon^{t,T} x_T^*, t, x_T).$$

If $w^\top \dot{Q}(t^*, x_T^*)v \neq 0$, *then* (t^*, x_T^*) *is a pitchfork singularity (with respect to* (t, x_T)*) of the solution set* $\{\widetilde{G} = 0\}$.

Proof. We have

$$\begin{aligned} D\widetilde{G}(t, x_T) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}(\Upsilon^{t,T} x_T - \Upsilon^{t,T} x_T^*), Q(t, x_T) \right), \\ D\widetilde{G}(t^*, x_T^*) &= (0, Q(t^*, x_T^*)) \end{aligned}$$

and so

$$\begin{aligned} \mathcal{N}(D\widetilde{G}(t^*, x_T^*)) &= \operatorname{span}\{\Phi_1, \Phi_2\}, \\ \mathcal{R}(D\widetilde{G}(t^*, x_T^*))^\perp &= \operatorname{span}\{\Psi\}, \end{aligned}$$

with $\Phi_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\Phi_2 = \begin{pmatrix} 0 \\ v \end{pmatrix}$ and $\Psi = \begin{pmatrix} 0 \\ w \end{pmatrix}$. Further, we compute

$$\begin{aligned} A &:= \left(\Psi((D^2 G)(t^*, x_T^*)(\Phi_i, \Phi_j)) \right)_{i,j=1}^2 \\ &= \begin{pmatrix} 0 & w^\top \dot{Q}(t^*, x_T^*)v \\ w^\top \dot{Q}(t^*, x_T^*)v & w^\top \frac{\mathrm{d}}{\mathrm{d}x_T}(Q(t^*, x_T^*)v)v \end{pmatrix}. \end{aligned}$$

By assumption, $\det A = -(w^\top \dot{Q}(t^*, x_T^*)v)^2 < 0$ and so (t^*, x_T^*) is a pitchfork singularity. □

The quantity $w^\top \dot{Q}(t^*, x_T^*)v$ in the assumptions of Theorem 4.7 is the time derivative of the singular value σ_n of Q at t^* (cf. Section B.8), which is zero at the singularity. We have therefore assumed that the singularity is an isolated point on its extremal.

The tangents of the solution branches at the pitchfork singularity are determined by the solutions of $z^\top A z = 0$ (with A from the proof of Theorem 4.7). $\widetilde{G}(t, x_T) = 0$ always has the solution branch $(t, x_T^*), t \in \mathbb{R}$, which corresponds to the tangent $z_1 = (1, 0)^\top$.

A second branch of solutions exists only at singularities. If $w^\top \frac{\mathrm{d}}{\mathrm{d}x_T}(Q(t^*, x_T^*)v)v = 0$ we have $z_2 = (0, 1)^\top$ and the corresponding second tangent $0 \cdot \Phi_1 + 1 \cdot \Phi_2 = (0, v^\top)^\top$ is orthogonal to the t-axis. In this case, the singularity is a cusp:

Theorem 4.8. *Let* $(t^*, x_T^*) \in \mathfrak{S}_1$, $\mathcal{N}(Q(t^*, x_T^*)) = \operatorname{span}\{v\}$, $\mathcal{R}(Q(t^*, x_T^*))^\perp = \operatorname{span}\{w\}$. *If* $w^\top \dot{Q}(t^*, x_T^*)v \neq 0$ *and* $w^\top \frac{\partial}{\partial x_T}(Q(t^*, x_T^*)v)v = 0$, *then* $(\Upsilon^{t,T} x_T^*, t^*, x_T^*)$ *is a cusp singularity of the solution set* $\{G = 0\}$.

Proof. According to Theorem 4.7, $(\Upsilon^{t,T} x_T^*, t^*, x_T^*)$ is a pitchfork bifurcation with respect to (t, x_T). As calculated above, the second solution branch has the tangent $(0, v^\top)^\top$.

The singular set $\mathfrak{S}$ can also be expressed as $\mathfrak{S} = \{(t, x_T) : \sigma_n(t, x_T) = 0\}$, where $\sigma_n(t, x_T)$ is the smallest singular value of $Q(t, x_T)$. Using (B.16) and the assumption, we have $D\sigma_n(t^*, x_t^*)\begin{pmatrix}0\\ v\end{pmatrix} = w^\top \frac{\mathrm{d}}{\mathrm{d}x_T}(Q(t^*, x_T^*)v)v = 0$. It follows that $(0, v^\top)^\top$ is a tangent of $\mathfrak{S}$ at (t^*, x_T^*) and so (t^*, x_T^*) is a cusp singularity. □

Next, we will show that a singularity which is not a cusp is connected to a cusp by a curve in $\mathfrak{S}$:

We write $z = (t, x_T)$ and, for $z \in \mathfrak{S}$, let $v(z)$ and $w(z)$ be non-zero vectors in $\mathcal{N}(Q(z))$ resp. $\mathcal{R}(Q(z))^\perp$.[2] Now, for any $(t^*, x_T^*) \in \mathfrak{S}$, define a curve $z : [t^*, \tau^*] \to \mathbb{R} \times \mathcal{X}$ by

$$\frac{\mathrm{d}}{\mathrm{d}\tau} z(\tau) = \begin{pmatrix}1\\ 0\end{pmatrix} - \frac{\frac{\partial}{\partial t}\sigma_n(z(\tau))}{\frac{\partial}{\partial x_T}\sigma_n(z(\tau))v(z(\tau))} \begin{pmatrix}0\\ v(z(\tau))\end{pmatrix} \tag{4.7}$$

and $z(t^*) = (t^*, x_T^*)$, with τ^* the maximal time for which the solution exists.

We will now show that z is indeed the connecting curve. Note that the denominator in (4.7) is 0 at the cusp itself, so the main technical difficulty is to show that it goes to 0 slowly enough for the solution to exist.

We have $\frac{\mathrm{d}}{\mathrm{d}\tau} t(\tau) = 1$ and $\frac{\mathrm{d}}{\mathrm{d}\tau}\sigma_n(t(\tau), x_T(\tau)) = 0$. The former implies $t(\tau) = \tau$ and the latter that $z(\cdot)$ remains in $\mathfrak{S}$, and so v and w are well-defined.

Further, it holds that $\tau^* < T$ because otherwise there would exist $z(T) = (T, x_T(T)) \in \mathfrak{S}$, i.e. $Q(T, x_T(T))$ would be singular, but $Q(T, x_T) = Id$ for all x_T.

We abbreviate $g(z) := \frac{\partial}{\partial t}\sigma_n(z(\tau))$ and $h(z) := \frac{\partial}{\partial x_T}\sigma_n(z(\tau))v(z(\tau))$ so that

$$\frac{\mathrm{d}}{\mathrm{d}\tau} x_T = -\frac{g(z)}{h(z)} v(z).$$

The solution of (4.7) can only cease to exist if $h \to 0$, so z will eventually enter some neighborhood of $\{h = 0\}$ as $\tau \to \tau^*$ and clearly $h(z)$ cannot change sign before reaching $\{h = 0\}$. We make the assumptions that g, h and v are generic in the sense that, for a small enough neighborhood, g and $\frac{\partial}{\partial x_T} h \cdot v$ are bounded away from zero. W.l.o.g. we can now assume $g(z) \geq 0$ and $h(z) \geq 0$ (by possibly replacing g, h and v by $-g$, $-h$ and $-v$ as necessary).

Treating g and $\frac{\partial}{\partial x_T} h \cdot v$ as approximately constant we have

$$\begin{aligned}
\frac{\mathrm{d}h}{\mathrm{d}\tau} &\approx -\frac{g}{h}\frac{\partial}{\partial x_T} h \cdot v \\
h\,\mathrm{d}h &\approx -g\frac{\partial}{\partial x_T} h \cdot v\,\mathrm{d}\tau \\
h &\approx \sqrt{2g\frac{\partial}{\partial x_T} h \cdot v \cdot (\tau^* - \tau)}.
\end{aligned}$$

[2] If $z \in \mathfrak{S}\backslash\mathfrak{S}_1$, those spaces have a dimension greater than one, and so defining v and w as their bases as in Theorem 4.7 is not valid.

It follows that $\frac{\mathrm{d}}{\mathrm{d}\tau}x_T$ behaves as $(\tau^* - \tau)^{-1/2}$ and hence z does not go to infinity, $z(\tau^*)$ exists and $\frac{\partial}{\partial x_T}\sigma_n \cdot v = w^\top \frac{\partial}{\partial x_T}(Q(t^*, x_T^*)v)v = 0$ at $z(\tau^*)$, i.e. the curve z leads to a cusp.

SUMMARY

We have now resolved the structure of the singularities in the non-degenerate situation, where the co-dimensions of the sets of folds and cusps in the (t, x_T)-space are the number of scalar conditions, i.e. 1 resp. 2 (cf. Lemma 4.6 resp. Theorem 4.8):

In the $(n+1)$-dimensional (t, x_T)-space[3] there is a $(n-1)$-dimensional set of cusp singularities, of which each generates a curve of fold singularities, which together form the n-dimensional set $\mathfrak{S}$.

The consequences for the value function can be illustrated by a simple one-dimensional example:

Example 4.9. *Let $\mathcal{X} = \mathcal{U} = \mathbb{R}$, $f(x,u) := u$, $c(x,u) := u^2 = \dot{x}^2$ and $\varphi(x) := \mathrm{e}^{-x^2}$. As the problem is autonomous, we can fix $t_0 := 0$ and use T as a parameter instead of t_0, which simplifies some expressions.*

Using (4.1) we find that for all extremal trajectories it holds that $\dot{\lambda} = 0$ and $\dot{x} = u = -\frac{\lambda}{2} = -\frac{\varphi'(x_T)}{2} = x_T \mathrm{e}^{-x_T^2}$. It follows that all extremal trajectories are linear, i.e.

$$x(t) = x_T - (T-t)x_T \mathrm{e}^{-x_T^2} \tag{4.8}$$

and

$$G(x_0, T, x_T) := \Upsilon^{0,T} x_T - x_0 = x_T(1 - T\mathrm{e}^{-x_T^2}) - x_0.$$

For $x_0 = 0$ we see that there is always the solution $x_T = 0$ and that the additional solutions $x_T = \pm\sqrt{\log T}$ appear for $T > 1$ (cf. Figure 4.1).

Taking the derivative of (4.8), we see that

$$Q(t, x_T) = 1 - (T-t)\mathrm{e}^{-x_T^2} + 2(T-t)x_T^2 \mathrm{e}^{-x_T^2}.$$

We further compute

$$\begin{aligned}
\dot{Q}(t, X_T) &= (1 - 2x_T^2)\mathrm{e}^{-x_T^2}, \\
\frac{\partial}{\partial x_T} Q(t, x_T) &= 2x_T(T-t)\mathrm{e}^{-x_T^2} + 4(T-t)x_T \mathrm{e}^{-x_T^2} - 4(T-t)X_T^3 \mathrm{e}^{-x_T^2}, \\
DG(x_0, T, x_T) &= (-1, -x_T \mathrm{e}^{-x_T^2}, Q(t, x_T)).
\end{aligned}$$

As $Q \in \mathbb{R}$, we can always use $v = 1$ and $w = 1$. At $(x_0, T, x_T) = (0, 1, 0)$ we have $G = 0$, $DG = (-1, 0, 0)$ and $w^\top \frac{\partial}{\partial x_T}(Qv)v = 0$, so according to Theorem 4.8 the point $(0, 1, 0)$ is a cusp singularity. Consequently there are two fold singularities in the (x_0, x_T)-plane for $T > 1$, as shown in Figure 4.2 for $T = 1.5$.

We can define a multivalued generalization $\widehat{V}$ of the value function V which assigns to

[3] If we consider instead (x_t, t, x_T), we have n additional dimensions cancelled out by the n additional conditions $\Upsilon^{t,T} x_T - x_t = 0 \in \mathbb{R}^n$.

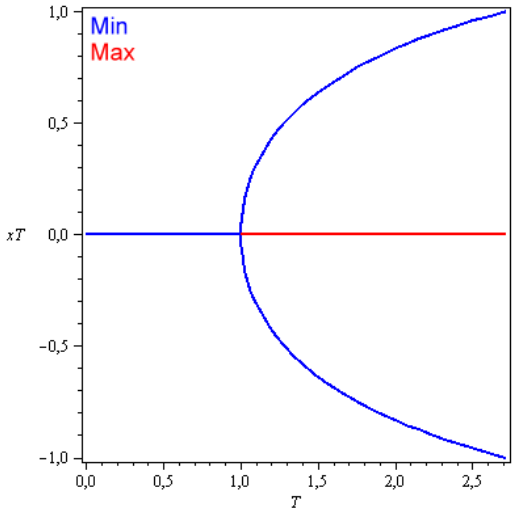

Figure 4.1.: Endpoints x_T of extremal trajectories for varying T and $x_0 = 0$

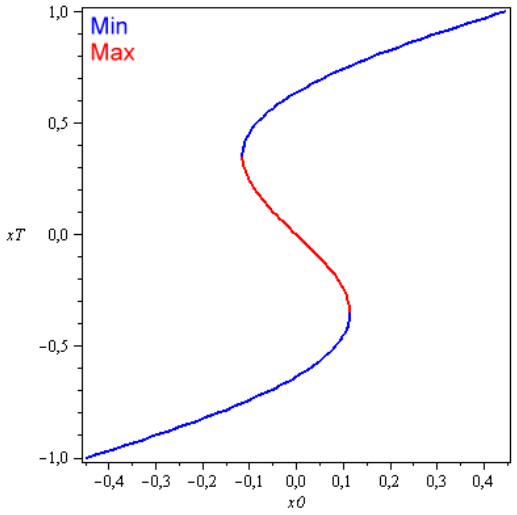

Figure 4.2.: Endpoints x_T of extremal trajectories for varying x_0 and $T = 1.5$

(Minimal trajectories in blue, maximal trajectories in red)

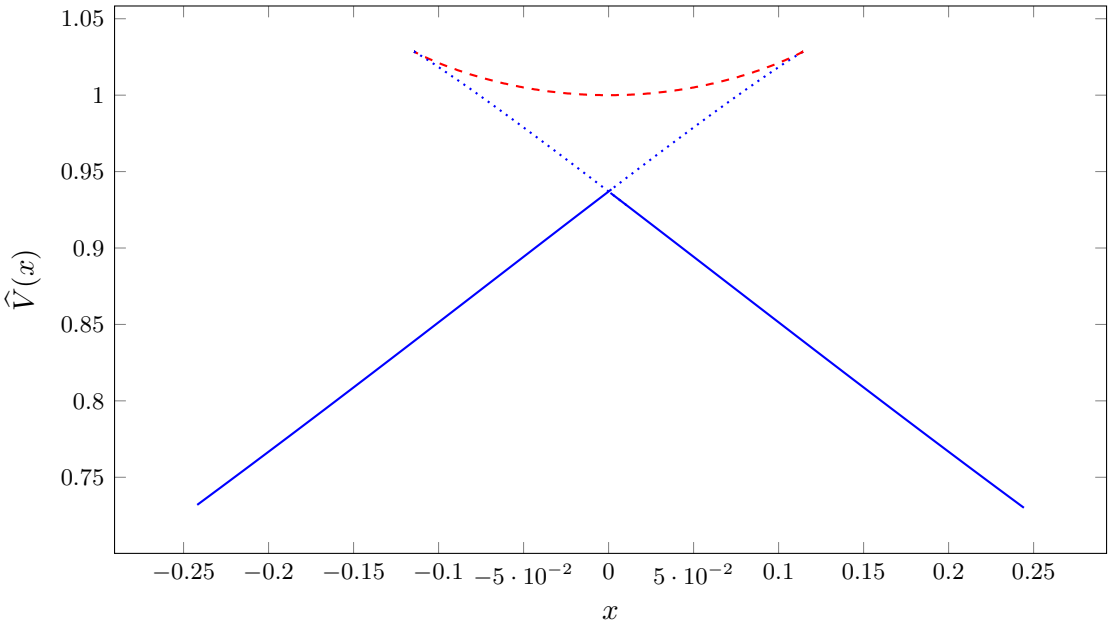

Figure 4.3.: $\widehat{V}$ for $T = 1.5$. The actual value function is solid and values of additional, not globally optimal, extrema are dotted for minima and dashed for maxima.

an initial condition x the values of all associated extremal trajectories, i.e.

$$\widehat{V}(x_0) := \{L(\mathbf{x}, \mathbf{u}) : (\mathbf{x}, \mathbf{u}) \text{ is extremal and } x(0) = x_0\},$$

and from which we recover the value function as

$$V(x) = \min \widehat{V}(x).$$

In Figure 4.3 we see that the cost for a specific branch of local minima is a smooth function of x_0 (until they turn into maxima at a fold), but the overall value function has a discontinuity in the derivative when the global solution switches to a different branch.

In higher dimensions, this structure is extended along the additional dimensions:

Example 4.10. *Let* $\mathbf{X} = \mathbf{U} = \mathbb{R}^2$, $f(x, u, t) := u$, $c(x, u, t) := \sin(x_1) + \sin(x_2) + 0.02(t + 0.95)\,\|u\|^2$, $\varphi(x) := \exp(-x_1^2 - x_2^2)$.

The cost for all extremals at $T = 0.05$ is shown in Figure 4.4. Note the line along the x_1 axis where the branches cross.

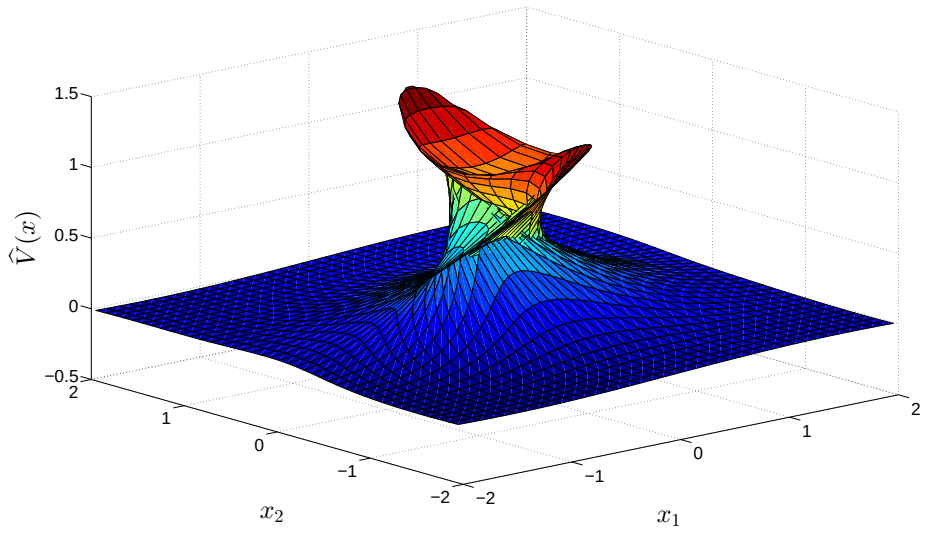

Figure 4.4.: The multi-valued function $\widehat{V}$ giving the cost for extremal trajectories in Example 4.10 for $T = 0.05$. Its graph is a self-intersecting surface. The grid lines show a parametrization by the terminal states $x(T)$ of the associated trajectories. See Figure 5.1 on p. 65 for a wireframe view.

4.3. Infinite horizon

We consider now the infinite horizon problem (2.17) introduced in Section 2.2. We assume that there is a cost-minimal fixed point (x^*, u^*) with $c(x^*, u^*) = \min c = 0$ that can be reached from any initial state and allow only trajectories that converge to x^*.

For this problem, (2.7) still holds (Theorem 2.18) and the terminal condition becomes $\lim_{t\to\infty} \lambda(t) = 0$ (Remark 2.19).

Following Osinga and Hauser [OH06] we lift the problem to coordinates (x, λ) involving both the state and the adjoint. Recall that (4.1) defines a dynamical system. Among its solutions, exactly those with $\lim_{t\to\infty}(x(t), \lambda(t)) = (x^*, 0)$ are extremal trajectories of the optimal control problem. In the language of dynamical systems, such extremal trajectories form $W^{\mathrm{s}}(x^*, 0)$, the stable manifold of $(x^*, 0)$ in (x, λ)-space.

One can define $L(x_0, \lambda_0) := L(\mathbf{x}, \mathbf{u})$, where $\mathbf{x}$ and $\mathbf{u}$ are from the solution of (4.1) with $x(0) = x_0$, $\lambda(0) = \lambda_0$ and recovers the multi-valued value function $\widehat{V}(x)$ as the projection of $V(W^{\mathrm{s}}(0, 0))$ along the λ-coordinate, i.e.

$$\widehat{V}(x) = \{L(x, \lambda) : (x, \lambda) \in W^{\mathrm{s}}(0, 0)\}\,.$$

To apply the results of the previous section, we will decompose the infinite time interval into a finite part, where bifurcations occur, and a well-behaved terminal part.

W.l.o.g. we assume $x^* = 0$. We expect simpler behaviour in a small neighborhood of $(x^*, \lambda^*) = (0, 0)$. The linearization of (4.1) around this point, according to (4.3), is[4]

$$\begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = \underbrace{\begin{pmatrix} A & -B \\ -C & -A^\top \end{pmatrix}}_{=:M} \begin{pmatrix} x \\ \lambda \end{pmatrix}. \tag{4.9}$$

Without a finite endpoint we do not have a direct replacement for the terminal condition $S(T) = \frac{\mathrm{d}}{\mathrm{d}x}\varphi$, so instead we merely guess that

$$\lambda = Sx$$

for some $S \in \mathbb{R}^{n\times n}$ independent of t. With this assumption and (4.9),

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}(Sx - \lambda) = S(Ax - B\lambda) + Cx + A^\top\lambda = SAx - SBSx + Cx + A^\top Sx$$

holds for all x, and so S would be a solution of the *Riccati equation*

$$SA + A^\top S - SBS + C = 0. \tag{4.10}$$

Being a quadratic equation, (4.10) might have spurious solutions. x^* should be a minimum of V, so it is justified to require that $S = \frac{\mathrm{d}^2}{\mathrm{d}x^2}V > 0$, i.e. that S is positive definite.

[4] The matrices A, B and C do not depend on t as we are now linearizing around a fixed point instead of along a trajectory.

Kucera has shown in [Kuc72, Theorem 3] that (4.10) has a unique solution $S > 0$ exactly if the system is stabilizable at x^*, and that $\lim_{t\to\infty} \mathrm{e}^{tM}\begin{pmatrix} x \\ Sx \end{pmatrix} = 0$ for this S and all x. It follows that the linear span of the columns of $\begin{pmatrix} Id \\ S \end{pmatrix}$ is part of the stable eigenspace $E^s(x^*, 0)$ and Proposition 7.2.1 of [vdS96] states that they are in fact equal, i.e.

$$E^s(x^*, 0) = \operatorname{span}\begin{pmatrix} Id \\ S \end{pmatrix},$$

and that $(x^*, 0)$ is a hyperbolic fixed point.

The following Lemma shows that for (x, λ) in a small neighborhood of $(x^*, 0)$, x uniquely determines (x, λ) and therefore the extremal trajectory.

Lemma 4.11. *There exists an open neighborhood $\Omega \subset \mathbb{R}^n \times \mathbb{R}^n$ of $(x^*, 0)$ such that if $(x, \lambda^{(i)}) \in \Omega \cap W^s(x^*, 0)$ for $i = 1, 2$, then $\lambda^{(1)} = \lambda^{(2)}$.*

Proof. By the Stable Manifold Theorem (cf. [Tes12, Theorem 9.4 and 9.5]), there is an open neighborhood $0 \in \Omega \subset \mathbb{R}^n \times \mathbb{R}^n$ and a function $h \in \mathcal{C}^1(E^s(x^*, 0); \mathbb{R}^n \times \mathbb{R}^n)$ such that

$$W^s(x^*, 0) \cap ((x^*, 0) + \Omega) = \{h(a) : a \in E^s(x^*, 0) \cap \Omega\}$$

with $h(x^*, 0) = (x^*, 0)$ and $Dh(x^*, 0) = Id_{2n}$. As a point in $E^s(x^*, 0) = \operatorname{span}\begin{pmatrix} Id \\ S \end{pmatrix}$ is determined by its x-coordinate we can replace $h(x, \lambda)$ by $h(x) := h(x, Sx)$. Let $\Pi_x(x, \lambda) := x$ be the projection onto the x-coordinate. Then $D(\Pi_x h) = Id_n$ and, for a sufficiently small Ω, $\Pi_x h$ is invertible. It follows that $(x, \lambda_1) = h((\Pi_x h)^{-1}(x)) = (x, \lambda_2)$. □

Lemma 4.11 shows that, for any fixed point $(x^*, 0)$, there are no singularities on extremals that lie within $\Omega = \Omega(x^*)$. Recall from the beginning of this section that every locally weakly overtaking minimal trajectory $(\mathbf{x}, \mathbf{u})$ converges to a point (x^*, u^*) with $Dc(x^*, u^*) = 0$ and its adjoint λ to 0. Hence, $(x(t), \lambda(t)) \in \Omega(x^*)$ for all $t \geq T$ and some finite time T.

From this we can finally conclude that all locally minimal extremals can be obtained by extending an extremal in some $\Omega(x^*)$ backwards for a finite time, and so the singularities have the same structure as in the finite time problem.

5. Finding global optima

In this chapter, we will use insights of Chapter 4 to find globally optimal solutions for the optimal control problem (2.17) with infinite horizon, that converge to a fixed point (x^*, u^*). We retain the assumptions of Chapter 4 and in particular Section 4.3.

We begin by discussing two approaches that we will ultimately *not* pursue, as they appear to be impractical.

- Osinga and Hauser ([OH06]) partially computed $W^s(x^*, 0)$ numerically with a continuation algorithm. Once a sufficiently large part of $W^s(x^*, 0)$ is known, one could find extremals (recall that $(x, \lambda) \in W^s(x^*, 0)$ includes λ, which determines u^*), the multi-valued function $\widehat{V}$ and finally the value function V. However, $W^s(x^*, 0)$ has the same dimension as the state space and although it is smooth, it is not a function but a manifold and "[its computation] is a serious challenge" ([OH06, p. 15]). For this reasons, the full computation of $W^s(x^*, 0)$ does not seem to be an efficient way to find the value function or optimal control.

- We have seen in Chapter 4 that singularities in Q and nonsmoothness of the value function are connected. However, the points (x, t) where the value function is smooth are almost disjoint from the points where Q is singular: both kinds of singularity coincide at cusps, but in general, the nonsmoothness in the value function arises from the minimization over different branches and the Q matrices belonging to the corresponding extremals are invertible.

  Essentially, the connection between non-smooth branch switches of the value function and Q-singularities is causal but not local. This is an obstacle to an attempt to find smooth regions of the value function by looking for singularities in Q.

5.1. Homotopies

In fact, the two main lessons we draw from the previous discussion is that multiple solutions of (4.1) (i.e. extremals) exist, and that trying to directly catalogue all of them is difficult. Let us now begin to describe what we intend to do.

As multiple solutions exist, the solution found by a numerical solver depends — often rather unpredictably — on the initial guess. To control which solution we obtain, we will use a homotopy:

We start with a solution $(\mathbf{x}^*, \mathbf{u}^*, \lambda^*)$ for $x(0) = x$ and smoothly change the initial condition to $x(0) = y$, i.e. we use a weakly differentiable function $x_0 : [0, 1] \to \mathcal{X}$ with $x_0(0) = x$ and $x_0(1) = y$.

Recall that extremals are the solutions of $F = 0$ as defined in (1.6) and let us assume that DF is invertible (cf. Section 1.4), otherwise we consider the homotopy failed. Then there exists a parametrized family of extremals $(\mathbf{x}^{(\eta)}, \mathbf{u}^{(\eta)}, \lambda^{(\eta)})$, $\eta \in [0,1]$ for the initial condition $x^{(\eta)}(0) = x_0(\eta)$ with $(\mathbf{x}^{(0)}, \mathbf{u}^{(0)}, \lambda^{(0)}) = (\mathbf{x}^*, \mathbf{u}^*, \lambda^*)$, which one obtains as the solution of

$$\frac{\mathrm{d}}{\mathrm{d}\eta}(\mathbf{x}^{(\eta)}, \mathbf{u}^{(\eta)}, \lambda^{(\eta)}) = -DF^{-1}\left(0, 0, \frac{\mathrm{d}}{\mathrm{d}\eta}x_0(\eta)\delta_0\right). \tag{5.1}$$

(Recall that $\frac{\mathrm{d}}{\mathrm{d}x_0}g = \delta_0$ and that g is the third component of F.)

Note that a homotopy requires a solution $(\mathbf{x}_0, \mathbf{u}_0, \lambda_0)$ to start with. Fortunately we can freely choose $x_0(0)$ and if we start at the target state, i.e. if $x(0) = x^*$, we have the obvious solution $\mathbf{x} \equiv x^*$, $\mathbf{u} \equiv u^*$, $\lambda \equiv 0$.

Path dependence

Definition 5.1. *A* parametrized homoptopy path *between $x \in \mathcal{X}$ and $y \in \mathcal{X}$ is a function $p \in \mathcal{W}^{1,\infty}([0,1]; \mathcal{X})$ that fulfills $p(0) = x \wedge p(1) = y$ or $p(0) = y \wedge p(1) = x$.*

This definition is symmetric in x and y, and w.l.o.g. we will always assume that $p(0) = x$ and $p(1) = y$.

The solution obtained at $\eta = 1$ depends not only on the starting solution at $\eta = 0$, but also on the homotopy path:

Example 5.2. *We continue from example 4.10. If we start with the globally optimal solution at $x_0 = (0, -1)$ and perform a homotopy along a straight line (shown in red in Figure 5.1) to $(0, 0.5)$, we cross a singularity in V_{xx}, where this branch of extremals loses global optimality, and end up with a non-minimal extremal. A different homotopy path (black) with a piecewise linear detour to $(1, 0)$ goes around this singularity and arrives at the globally optimal solution.*

To find globally optimal branches of $\widehat{V}$, we will have to try and compare different paths.

5.2. Description of the Algorithm

To this end, we choose a graph $(\mathcal{V}, \mathcal{E})$ where the vertices are initial states and the edges are homotopy paths connecting them, i.e. we have:

$$\begin{aligned}&\mathcal{V} \subseteq \mathcal{X},\\ &\mathcal{E} \subseteq \{(\{v, w\}, p) : v, w \in \mathcal{V}, v \neq w, p \text{ a homotopy path between } v \text{ and } w\}.\end{aligned}$$

Some ways of choosing such a graph are given in the examples in Section 5.7.

Remark 5.3. *The above definition states that an edge is defined by the pair of vertices it connects and the path between them. Between two vertices v and w there may in principle be several edges with different homotopy paths p. In the examples below, there will only be one edge between two vertices, which is the straight line.*

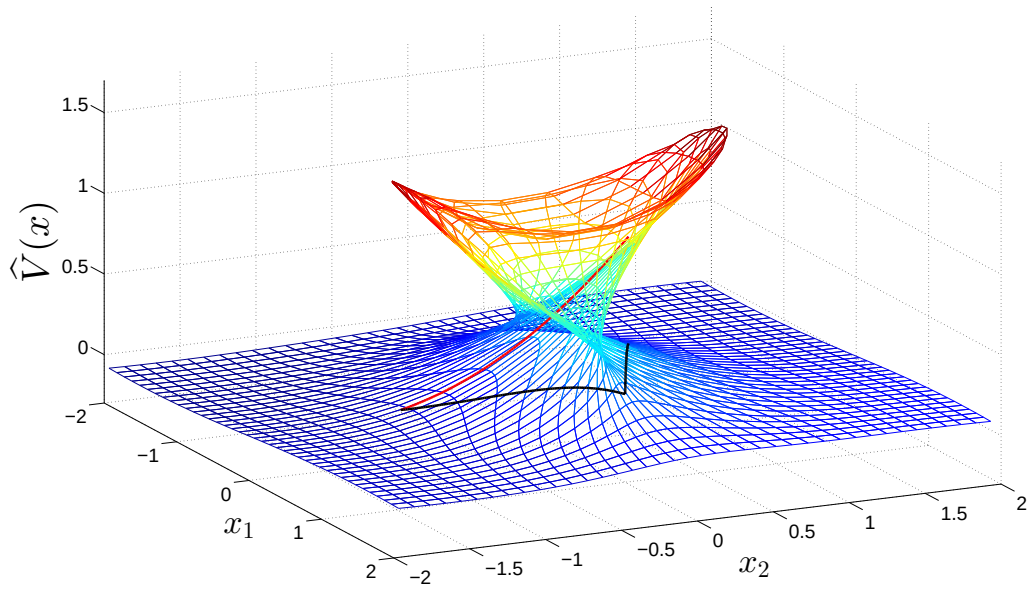

Figure 5.1.: Two homotopy paths resulting in different solutions in Example 5.2.

By a *path in the graph* $(\mathcal{V}, \mathcal{E})$ *between* v *and* w we understand a sequence $(\{v_i, v_{i+1}\}, p_i) \in \mathcal{E}, i = 1, \ldots, n-1$ of edges such that $v = v_1$, $w = v_n$. A path in the graph defines an associated homotopy path through the (re-parametrized) concatenation of the p_i.

Ideally we would like to try all of the paths in the graph. Unfortunately, the number of paths can grow superpolynomially with the number of vertices[1] even if no loops are allowed.

So instead we propose an algorithm which keeps track of only the best known solution for each vertex and iteratively tries to spread good solutions to neighboring vertices.

We associate with each vertex $v \in \mathcal{V}$ the following variables, which may change as the algorithm proceeds:

- The best currently known trajectory $(\mathbf{x}_v, \mathbf{u}_v)$ with $\mathbf{x}_v(0) = v$. This variable may be `null` if no trajectory is known.

- The cost $L_v := L(\mathbf{x}_v, \mathbf{u}_v)$ achieved by the above trajectory (with the convention $L(\texttt{null}) = \infty$).

[1] E.g. in an $n \times n$-grid there are already 2^n paths of length n that start in a the lower left corner and only go up or right.

The algorithm then proceeds as follows:

Algorithm 5.1:

Input: A graph $(\mathcal{V}, \mathcal{E})$ with $x^* \in \mathcal{V}$.
Set $(\mathbf{x}_{x^*}, \mathbf{u}_{x^*}) \equiv (x^*, u^*)$ and $(\mathbf{x}_v, \mathbf{u}_v) \leftarrow$ `null` for all $v \in \mathcal{V} \backslash \{x^*\}$
Create a set `todo` $:= \{x^*\}$
while `todo` *is not empty* **do**
  Pick any $v \in$ `todo`
  Remove v from `todo`
  forall $(\{v, w\}, p) \in \mathcal{E}$ **do**
    Starting from $(\mathbf{x}_v, \mathbf{u}_v)$, compute an extremal trajectory $(\mathbf{x}, \mathbf{u})$ with $\mathbf{x}(0) = w$ by performing the homotopy from v to w along p. $(\mathbf{x}, \mathbf{u}) :=$ `null` if this homotopy fails.
    if $L(\mathbf{x}, \mathbf{u}) < L_w$ **then**
      $(\mathbf{x}_w, \mathbf{u}_w) \leftarrow (\mathbf{x}, \mathbf{u})$
      Set $w \in$ `todo`
    end
  end
end
Output: The best found solution $(\mathbf{x}_v, \mathbf{u}_v)$ for each vertex $v \in \mathcal{V}$

Note that Algorithm 5.1 only computes solutions for initial conditions $v \in \mathcal{V}$. If the objective is to find the solution only for a single x_0, one can of course include x_0 in $\mathcal{V}$. Alternatively, for $x_0 \notin \mathcal{V}$, one can choose some nodes $v \in \mathcal{V}$ close to x_0, perform homotopies from these nodes to x_0 and pick the best solution.

5.3. Conditions for finding the global optimum

If the graph contains a homotopy path that leads to the globally optimal solution for some $v \in \mathcal{V}$ and we would actually try all paths, then we would find this solution. The following Theorem shows that Algorithm 5.1 achieves almost the same despite not trying all paths.

Theorem 5.4. *If the graph $(\mathcal{V}, \mathcal{E})$ contains any path from x^* to $x_0 = v \in \mathcal{V}$ along which all intermediate homotopy solutions are globally optimal (and exist) and Algorithm 5.1 terminates, then the output $(\mathbf{x}_v, \mathbf{u}_v)$ is the globally optimal solution for the initial condition $x(0) = x_0$.*

Proof. Let $x^* = v_1, v_2, \ldots, x_0 = v_n$ be the vertices in the path. Assume, for the sake of contradiction, that the solution at v_n is not globally optimal. As the solution $(\mathbf{x}, \mathbf{u}) \equiv (\mathbf{x}^*, \mathbf{u}^*)$ at v_1 is globally optimal by assumption, there exists a minimal k, $1 < k \leq n$ at which the solution is not globally optimal. Let $v := v_{k-1}$, $w := v_k$ and $(\mathbf{x}_v^*, \mathbf{u}_v^*)$ resp. $(\mathbf{x}_w^*, \mathbf{u}_w^*)$ the associated g. opt. solutions. The assignment $(\mathbf{x}_v, \mathbf{u}_v) \leftarrow (\mathbf{x}_v^*, \mathbf{u}_v^*)$ must have taken place at some point during the execution of the algorithm, followed immediately by the insertion $v \in$ `todo`. Note that $(\mathbf{x}_v, \mathbf{u}_v)$ stays constant after $(\mathbf{x}_v, \mathbf{u}_v) \leftarrow (\mathbf{x}_v^*, \mathbf{u}_v^*)$ as it could only have been overwritten by a better solution. If the algorithm terminates,

v must have been removed from `todo` in a subsequent step of the iteration. In that step the homotopy from $v = v_{k-1}$ to $w = v_k$ would, by assumption of the theorem, have yielded $(\mathbf{x}_w^*, \mathbf{u}_w^*)$ and consequently we would have $(\mathbf{x}_w, \mathbf{u}_w) = (\mathbf{x}_w^*, \mathbf{u}_w^*)$ after the algorithm terminates. This contradiction completes the proof. □

The following theorems provide sufficient conditions to meet the assumptions of Theorem 5.4.

Theorem 5.5. *If the assumptions of Theorem 2.21 are fulfilled and there is no extremal with $x(0) \in \mathcal{V}$ for which $Q(0)$ is singular, then the algorithm terminates.*

Proof. By way of contradiction, assume that the algorithm does not terminate, i.e. `todo` never becomes empty. In each step of the algorithm, one vertex is removed from `todo` and vertices are only added, when their current trajectory is replaced by one with lower cost. Hence there must exist at least one vertex v with an infinite sequence of extremals $(\mathbf{x}_i, \mathbf{u}_i)$, $i \in \mathbb{N}$ of decreasing cost $L(\mathbf{x}_{i+1}, \mathbf{u}_{i+1}) < L(\mathbf{x}_i, \mathbf{u}_i)$.

Using coercivity as in Theorem 2.21, we have a limit trajectory $\mathbf{x}_i \to \mathbf{x}^*$ as $i \to \infty$. Given $\mathbf{x}^*$, we define $\mathbf{u}^*$ and λ^* as the solution of (2.7) and $\lim_{t\to\infty} \lambda(t) = 0$ (cf. Remark 2.19). As $(\mathbf{x}_i, \mathbf{u}_i, \lambda_i)$ fulfills the same ODE, we have $(\mathbf{x}_i, \mathbf{u}_i, \lambda_i) \to (\mathbf{x}^*, \mathbf{u}^*, \lambda^*)$ as $i \to \infty$, which contradicts the local uniqueness implied by a non-singular $Q(0)$. □

The existence of at least one suitable homotopy path is guaranteed, as the optimal trajectory itself is such a path.

Theorem 5.6 (Dynamic Programming Principle)**.** *If $(\mathbf{x}^*, \mathbf{u}^*)$ is the optimal solution of (2.17) with $x_0 = x^*(0)$, then for any $T \in [0, \infty)$, the shifted trajectory $(\mathbf{x}^T, \mathbf{u}^T) := (\mathbf{x}^*(\cdot + T), \mathbf{u}^*(\cdot + T))$ is the optimal solution of (2.17) with $x_0 = x^*(T)$.*

Proof. $(\mathbf{x}^*, \mathbf{u}^*)$ fulfills $\dot{x} = f(x, u)$ and as the dynamics are autonomous, so does $(\mathbf{x}^T, \mathbf{u}^T)$. It remains to show optimality. Assume that $(\mathbf{x}, \mathbf{u})$ is admissible for $x_0 = x^*(T)$ and $L(\mathbf{x}, \mathbf{u}) < L(\mathbf{x}^T, \mathbf{u}^T)$. Define the spliced trajectory $(\widehat{\mathbf{x}}, \widehat{\mathbf{u}})$ as

$$\widehat{x}(t) := \begin{cases} x^*(t) & , t < T \\ x(t-T) & , t \geq T \end{cases}, \qquad \widehat{u}(t) := \begin{cases} u^*(t) & , t < T \\ u(t-T) & , t \geq T \end{cases}.$$

As above, $(\widehat{\mathbf{x}}, \widehat{\mathbf{u}})$ is admissible for $x_0 = x^*(0)$ and we have

$$\begin{aligned} L(\widehat{\mathbf{x}}, \widehat{\mathbf{u}}) &= \int_0^T \mathrm{e}^{-\mu t} c(x^*(t), u^*(t)\,\mathrm{d}t + \int_T^\infty \mathrm{e}^{-\mu t} c(x(t-T), u(t-T))\,\mathrm{d}t \\ &= \int_0^T \mathrm{e}^{-\mu t} c(x^*(t), u^*(t)\,\mathrm{d}t + \mathrm{e}^{-\mu T} L(\mathbf{x}, \mathbf{u}) \\ &< \int_0^T \mathrm{e}^{-\mu t} c(x^*(t), u^*(t)\,\mathrm{d}t + \mathrm{e}^{-\mu T} L(\mathbf{x}^T, \mathbf{u}^T) \\ &= \int_0^T \mathrm{e}^{-\mu t} c(x^*(t), u^*(t)\,\mathrm{d}t + \int_T^\infty \mathrm{e}^{-\mu t} x^*(x^*(t), u^*(t))\,\mathrm{d}t \\ &= L(\mathbf{x}^*, \mathbf{u}^*), \end{aligned}$$

contradicting the optimality of $(\mathbf{x}^*, \mathbf{u}^*)$. □

Corollary 5.7. *Any globally optimal solution $(\mathbf{x}, \mathbf{u})$ is a path from x^* to $x(0)$ fulfilling the condition of Theorem 5.4.*

Remark 5.8. *As global optimality of a branch of $\widehat{V}$ changes only at singularities of V_{xx}, any variation of a path fulfilling the condition is still suitable, as long as no crossing of a singularity is introduced (cf. Figure 5.2).*

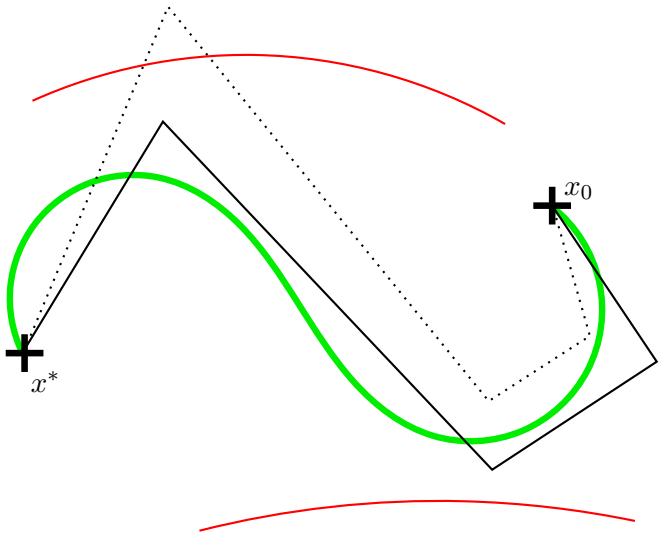

Figure 5.2.: Illustration of Remark 5.8. The bold green trajectory is the optimal solution, the red lines are singularities of V_{xx}. The solid black path fulfills the assumption of Theorem 5.4, the dotted one does not.

5.4. Algorithms for subproblems

For our purposes, we are not interested in the entire family $(\mathbf{x}^{(\eta)}, \mathbf{u}^{(\eta)}, \lambda^{(\eta)})$, but only in $(\mathbf{x}^{(1)}, \mathbf{u}^{(1)}, \lambda^{(1)})$, and so it is unnecessary to actually solve (5.1).

To solve (1.6), we will use Newton's method, which is locally convergent, i.e. if the initial guess is sufficiently close to a solution, then Newton's method will return that solution. In particular, if a solution for $x(0) = x_0 + v$, $\|v\| \ll 1$ is used as an initial guess for the problem with $x(0) = x_0$, the solution found will be similar.

Hence we can employ a *discrete homotopy* where we have $0 = \eta_0 < \eta_1 < \ldots < \eta_k = 1$ and, for $j = 1, \ldots, k$, successively compute $(\mathbf{x}^{(\eta_j)}, \mathbf{u}^{(\eta_j)}, \lambda^{(\eta_j)})$ as the solution for $x(0) = x_0(\eta_j)$ with $(\mathbf{x}^{(\eta_{j-1})}, \mathbf{u}^{(\eta_{j-1})}, \lambda^{(\eta_{j-1})})$ as the initial guess.

The discretization of F and DF (as defined in (1.6) and (1.11)) used in our implementation and further details are described in Appendix A.

5.5. Algorithmic variations

Closed-loop control

So far, we have only computed control trajectories in advance and the question remains what to do if, through some perturbation, the controlled system deviates from the trajectory.

The most straightforward method is to compute a new solution with the current state as the initial state (see above).

A more efficient way for small deviations is to compute S from (4.6) in advance, use the linearization $d\lambda = Sdx$ and solve (2.7b) (or its linearization) for u^*. This approach is known as *perturbation feedback control* ([BH69, Section 6.4]). If it is affordable, a new trajectory can periodically be computed from the current state to get an updated linearization.

Graph refinement

After Algorithm 5.1 finishes, it is possible to refine the graph by adding or removing vertices and edges. In this case, the initialization on the refined graph proceeds as follows:

Algorithm 5.2: Initialization on refined graph

Old vertices v keep $(\mathbf{x}_v, \mathbf{u}_v)$
forall *newly added vertices* v **do**
  $(\mathbf{x}_v, \mathbf{u}_v) \leftarrow$ `null`
end
`todo`$\leftarrow \{v : (\mathbf{x}_v, \mathbf{u}_v) \neq$ `null` and v appears in a new edge$\}$
Continue with **while** loop as in Algorithm 5.1

5.6. Discussion

Although the results of Section 5.3 (cf. in particular Theorem 5.4 and Remark 5.8) imply that the global optimum can be found with a sufficiently fine graph, there is no perfectly reliable way to ascertain whether a given graph is fine enough not to have missed a better solution.

Heuristically, one might try to refine the graph until stagnation is reached (i.e. all nodes shared by the finest and second-finest graph have the same solution in both graphs). A too coarse gird can also sometimes be detected as it may lead to no solution at all being found for some nodes.

We remark that finding global optima can generally not be guaranteed without an exhaustive search ("no free lunch") and limitations like the above are therefore unavoidable.

Computational cost

The proposed algorithm is much more expensive per node than the solution of the HJB equation, as for each node one needs to compute an entire trajectory instead of merely the value V.

On the other hand, the required resolution is also much lower than for the HJB equation. The spatial resolution of the grid is completely independent from the accuracy of the extremals, and matters only for the qualitative question of whether all relevant branches will be found.

In fact, the minimal grid that suffices to find global optima for all initial conditions would have only one node on each branch that contains a globally optimal solution. Of course, such a grid is not known in advance and in practice significantly more than one node will be needed per branch, but nevertheless the computational effort should be roughly proportional to the number of branches. This number can be understood as the inherent complexity of the problem and may scale less than exponentially with the dimension.

5.7. Numerical examples

The relation between neighboring solutions can be deduced from the outcome of the homotopy and the subsequent comparison during Algorithm 5.1. There are three cases, shown in Figure 5.3:

1. The homotopy succeeds and the solution obtained by homotopy is identical to the solution at the target node. In this case, both solutions are on the same branch.

2. The homotopy succeeds and the solution obtained by homotopy is worse than the solution at the target node: The solutions on the nodes are on different branches and along the homotopy path there is a singularity of V_{xx}.

3. The homotopy fails: The solutions on the nodes are on different branches and along the homotopy path there is a singularity of Q.

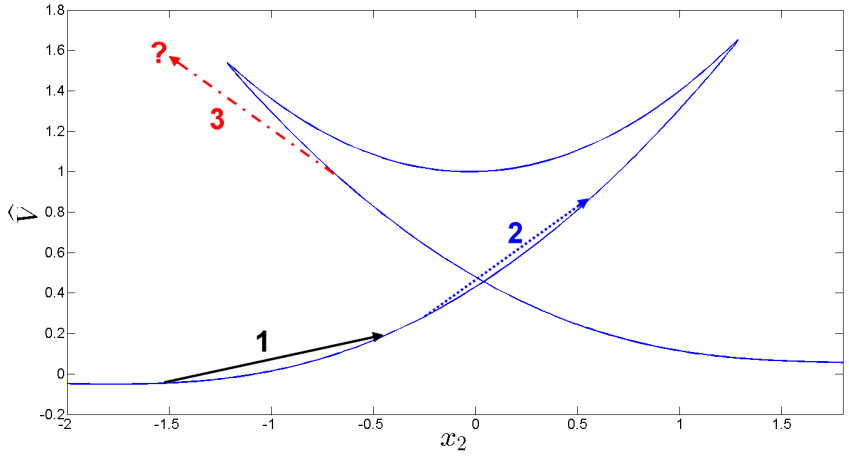

Figure 5.3.: Possible types of homotopy outcomes in Algorithm 5.1, demonstrated on $\widehat{V}(0, x_2)$ from Example 4.10.

Note that if the solutions belonging to two nodes v and w are on different branches, it can occur that the branch v is on extends to w, but not vice versa. In this situation we would have case 2 going from v to w and case 3 in the other direction.

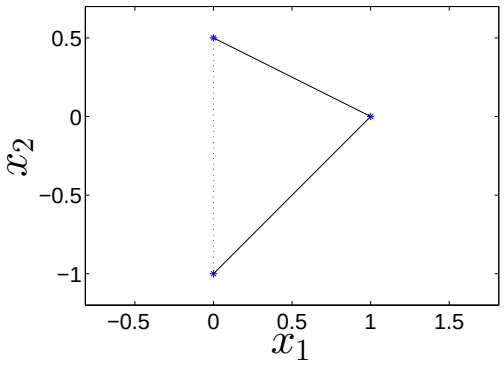

Figure 5.4.: Graph for Example 5.2.

In the following examples, we will draw edges black and solid for case 1, blue and dotted, if case 2 occurs in at least one direction, and red and dash-dotted if case 3 occurs in both directions. For example, the paths from Example 5.2 would be displayed as in Figure 5.4.

The inverted pendulum

We consider a planar inverted pendulum mounted on a cart to which a horizontal force u can be applied (cf. Fig. 5.5). This is a simple example that is frequently studied, e.g. in [OH06].

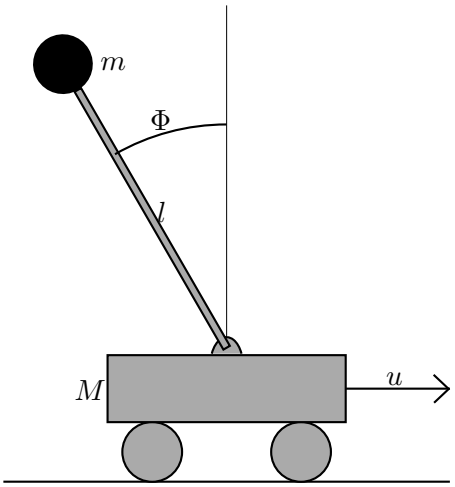

Figure 5.5.: The inverted pendulum (modified from [Kri12]).

We do not keep track of the position of the cart and so the state consists only of the offset angle Φ from the upright position and the corresponding angular velocity Φ_t. The equations of motion are given by

$$\begin{aligned}\dot{\Phi} &= \Phi_t, \\ \dot{\Phi}_t &= \frac{\frac{g}{l}\sin(\Phi) - \frac{1}{2}m_r\Phi_t^2\sin(2\Phi) - \frac{m_r}{ml}\cos(\Phi)u}{\frac{4}{3} - m_r\cos^2(\Phi)},\end{aligned}$$

where g is the gravitational constant, l is the distance between the center of mass and the pivot (half the length of the pendulum for an ideal weightless rod), M is the mass of the cart, m is the mass of the pendulum and m_r is the mass ratio $\frac{m}{m+M}$.

We define the quadratic cost function

$$c(x,u) = c\left((\Phi,\Phi_t),u\right) = q_1\Phi^2 + q_2\Phi_t^2 + ru^2$$

and use the dimensionless parameters $g = 9.8$, $l = 0.5$, $m = 2$, $M = 8$, $q_1 = 0.1$, $q_2 = 0.05$ and $r = 0.01$.

As a state space, we use $\mathbb{R}^2$, which means that we allow the pendulum to be in a downward position (i.e. the cart is e.g. moving on an elevated rail) and keep track of

the number of full rotations the pendulum has performed. The cost-minimal fixed point $(x^*, u^*) = (0, 0)$ represents the pendulum being upright and at rest.

When the pendulum is in a downward position ($\Phi \approx \pm\pi$), there are different strategies for moving it to the upright state ($\Phi = 0$), which we expect to be reflected in the results of the Algorithm: one can either attempt to directly swing the pendulum up in one go or first build up momentum by swinging the pendulum back and forth. In the state space (Φ, Φ_t), the latter is represented by a spiral moving around $(\pm\pi, 0)$ and outward.

Step-by-step demonstration of the algorithm

We begin with a very simple graph. This is not intended to be a serious application of Algorithm 5.1, but a step-by-step demonstration of how it works.

Figure 5.6 shows some intermediate snapshots during its execution. In addition to the graph, we draw the extremal trajectories $\mathbf{x}_v$ starting at the vertices as solid blue lines. Edges along which no homotopy has been performed yet are not drawn.

In the top left figure, the first iteration of the while-loop has been completed. Extremals are known for the neighbors A, B and C of $x^* = (0, 0)$ and since a new (here: the first) extremal has just been found for these vertices, they have been added to `todo`. We decide not to pick A or B for now and perform further iterations until, in the bottom left figure, we reach the vertex E.

The following iteration is the first in which we move to vertices (A and B) for which an extremal is already known. In this case, the extremals obtained by the homotopy from E are different from the old ones obtained by homotopies from x^*. Computation shows that the new extremals achieve a lower cost than the old ones and so they replace them (bottom right). Also, A and B are added to the set `todo`, which has no effect since they are already in it.

In the next iteration, we pick A from `todo` and so have to perform the homotopies to E and x^*. The homotopy back to E gives the solution we already have there. The homotopy from A to x^* yields a trajectory which leaves $(0, 0)$ and goes around in a loop before returning. This is more costly than the old solution, which simply stays at the origin, so the old solution is not replaced. The vertex A gets removed from `todo` and no new vertices are added.

In one further iteration, we pick the only element B from `todo`. The homotopies from B to E resp. x^* are attempted with the same outcome as before from A and the algorithm then terminates with the set `todo` empty. The trajectories shown on the bottom right are the final output.

Random graph with refinement

We will now consider a more elaborate approach which starts with a random graph and subsequently refines it. We want the edges to be short paths that can be combined into larger homotopy paths, so we choose only the vertices randomly and connect vertices which are close together. To be precise, we use the Delaunay triangulation (computed using the Bowyer-Watson algorithm [Bow81, Wat81]) of the vertices as a graph. (The Delaunay triangulation is the dual of the Voronoi diagram: two vertices are connected if their Voronoi cells touch.)

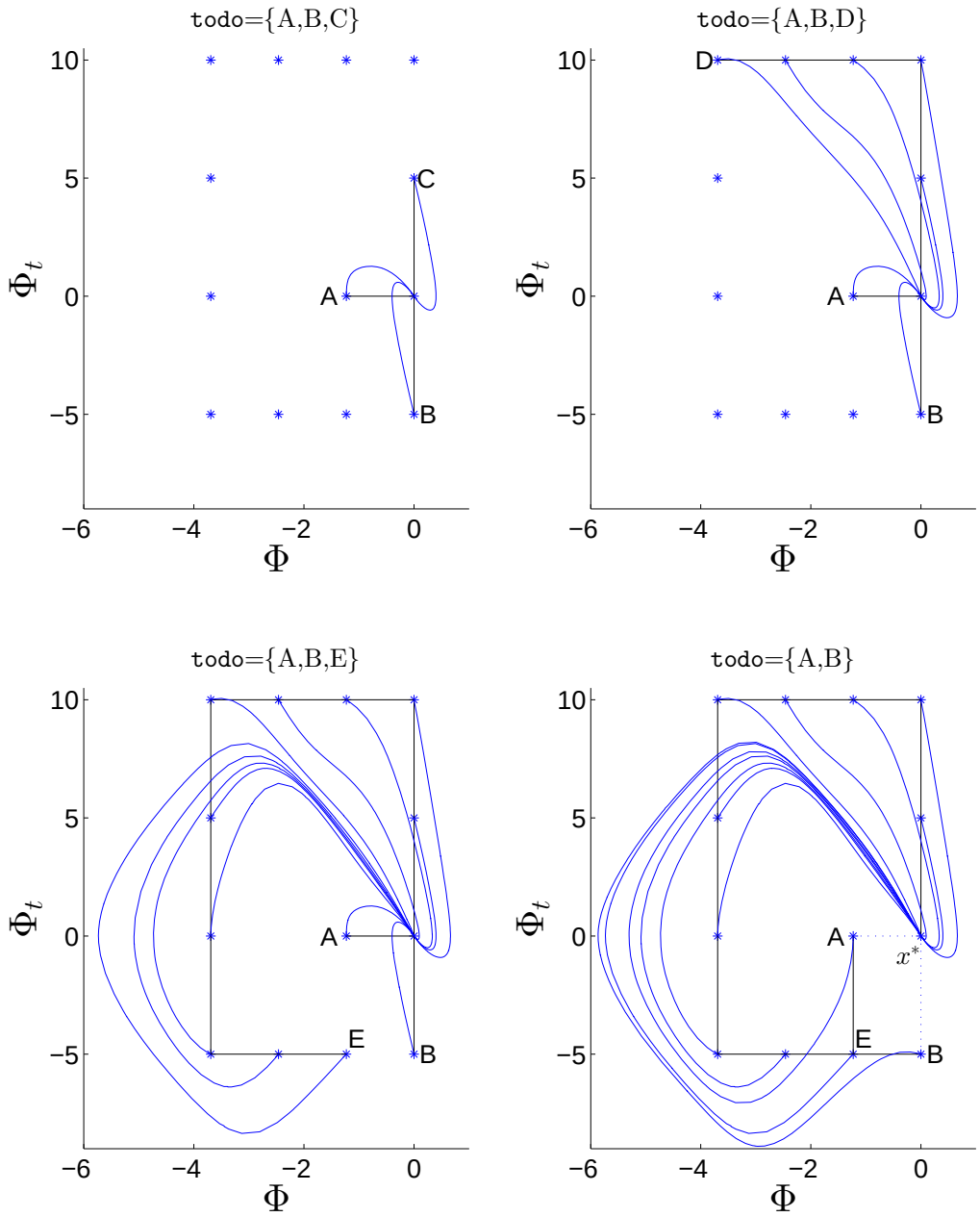

Figure 5.6.: Snapshots (left to right, top to bottom) during the execution of Algorithm 5.1 for the inverted pendulum on a very small graph. 5.1

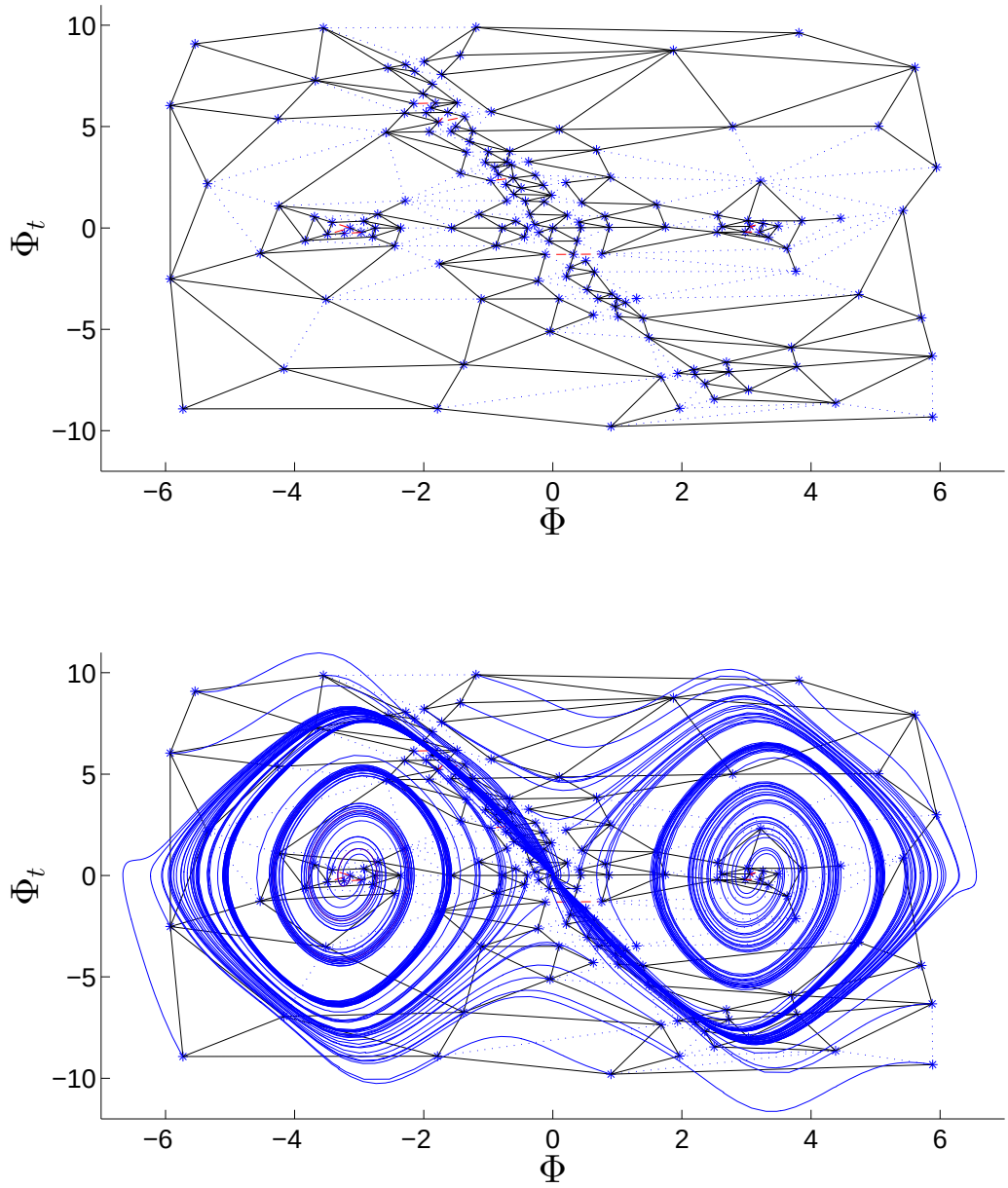

Figure 5.7.: Final graph and trajectories for the Inverted Pendulum with graph refinement

We start with 32 vertices: the fixed point $(x^*, u^*) = (0, 0)$, the downward position $(-\pi, 0)$ for which we want to compute a solution and 30 random points in $[-6, 6] \times [-10, 10]$. We have stated before that the connection between singularities in Q (or equivalently DF) and branch switches is rather loose, but it still proves useful as a guide for refinement: We first run Algorithm 5.1 on the initial 32-vertex-graph. After it finishes, we add new vertices on the midpoints of all edges $(\{v, w\}, p)$ with $\|v - w\| > l_{\min} := 0.05$ if the homotopy along p has failed.[2] We then update the Delaunay triangulation, run Algorithm 5.2 and repeat until there are no edges fulfilling the condition for refinement.

The results are shown in Figure 5.7. The final graph has 162 vertices and 468 edges, 7 refinement iterations have been done and 2335 homotopies have been performed (i.e. on average about 5 per edge of the final graph). We see that the refinement occurs primarily in two areas: near the downward positions $(\pm\pi, 0)$, where extremals with varying numbers of rotations compete for optimality and in the narrow channel (extending diagonally from the origin) where the direct approach is the globally optimal solution. Note that the refinement strategy has produced paths through that channel, which per Remark 5.8 (cf. also the accompanying Figure 5.2) the graph needs to contain to fulfill the conditions of Theorem 5.4.

A PDE example

In this example we consider a high-dimensional problem arising from the discretization of a PDE.

Consider an ensemble of particles on the one-dimensional torus $\Omega = [0, 2\pi)$, where the trajectory $t \mapsto z(t)$ of an individual particle is controlled by a potential $v(z, t)$ and given by the stochastic differential equation

$$\mathrm{d}z = -\frac{\partial}{\partial z} v \,\mathrm{d}t + \sigma \,\mathrm{d}W,$$

with $\sigma > 0$ and W a standard Wiener process (with an independent copy for each particle).

The density $y(z, t)$ of the ensemble evolves according to the *Fokker-Planck equation* (cf. [LM98])

$$y_t = -\operatorname{div}(-v_z \cdot y) + \frac{\sigma^2}{2} \Delta y.$$

As an example of a control problem, we want to move the particles towards $z = \pi$ with a small potential v and choose the cost function[3]

$$c(y(\cdot, t), v(\cdot, t)) = \left\langle \pi^{-1/2} \cos, y(\cdot, t) \right\rangle_{\mathcal{L}^2} + \alpha \|v\|^2_{\mathcal{L}^2}.$$

To avoid any difficulties regarding the existence and regularity of an optimal v, we restrict $v(\cdot, t)$ to be a trigonometric polynomial of at most degree M, thereby turning it into a

[2] p is still always the straight line between v and w.

[3] The choice of the cosine is mostly arbitrary. We only want a minimum at $z = \pi$.

finite-dimensional object.

Using a Fourier-Galerkin discretization, the problem turns into

$$y(z,t) = \sum_{i=0}^{2N} x_i(t)\varphi_i(z),$$
$$v(z,t) = \sum_{i=0}^{2M} u_i(t)\varphi_i(z),$$

with $\varphi_0 \equiv 1/\sqrt{2\pi}$, $\varphi_{2k}(z) = \cos(kz)/\sqrt{\pi}$ and $\varphi_{2k+1} = \sin(kz)/\sqrt{\pi}$, as well as

$$\dot{x}_i = \left\langle \operatorname{div}(v_z \cdot y) + \frac{\sigma^2}{2}\Delta y, \varphi_i \right\rangle.$$

By isometry of the Fourier Transform, the cost function becomes

$$c(x,u) = x_2 + \alpha \|u\|^2.$$

The Fokker-Planck equation conserves the total mass of the particles. Hence x_0 is constant in t and we normalize by setting $x_0 \equiv 1/\sqrt{2\pi}$ (so that $\int_0^{2\pi} y(z,t)\,\mathrm{d}z = 1$). The constant part of u does not influence the dynamics, which depend only on $\frac{\partial}{\partial z}v$. In order to minimize c, we fix $u_0 \equiv 0$. With the 0-th coefficients fixed, the state then consists of $(x_1, \ldots, x_{2N})$ and the control of $(u_1, \ldots, u_{2M})$.

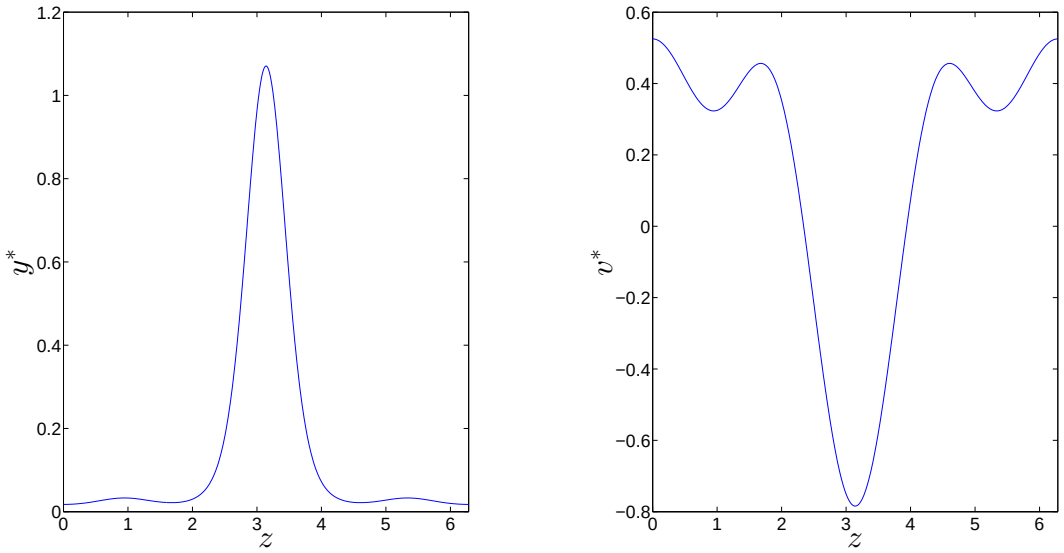

Figure 5.8.: The cost-minimal steady state (y^*, v^*).

The steady state (x^*, u^*) minimizing c can be computed numerically. The corresponding density y^* is concentrated in a peak at π and the potential v^* has a well at π (Figure 5.8).

Now consider an initial state where the density is partially concentrated in a peak but in the wrong place, i.e. not at π. As the domain is a torus, this peak could be moved either left or right and in both cases reach π, and we expect the two directions to correspond to different branches of local optima.

Guided by this, we consider initial states where y^* is shifted and blended with the uniform distribution, i.e.

$$y^{(\phi,r)} := ry^*(\cdot + (\phi - \pi)) + (1 - r)\varphi_0, \ \phi \in [0, 2\pi), r \in [0, 1],$$

where ϕ determines the position of the peak and r the amount of mass in it, and use a grid $\phi_k = 2\pi k/K$, $k = 0, \ldots, K-1$, $r_l = l/L$, $l = 0, \ldots, L$.

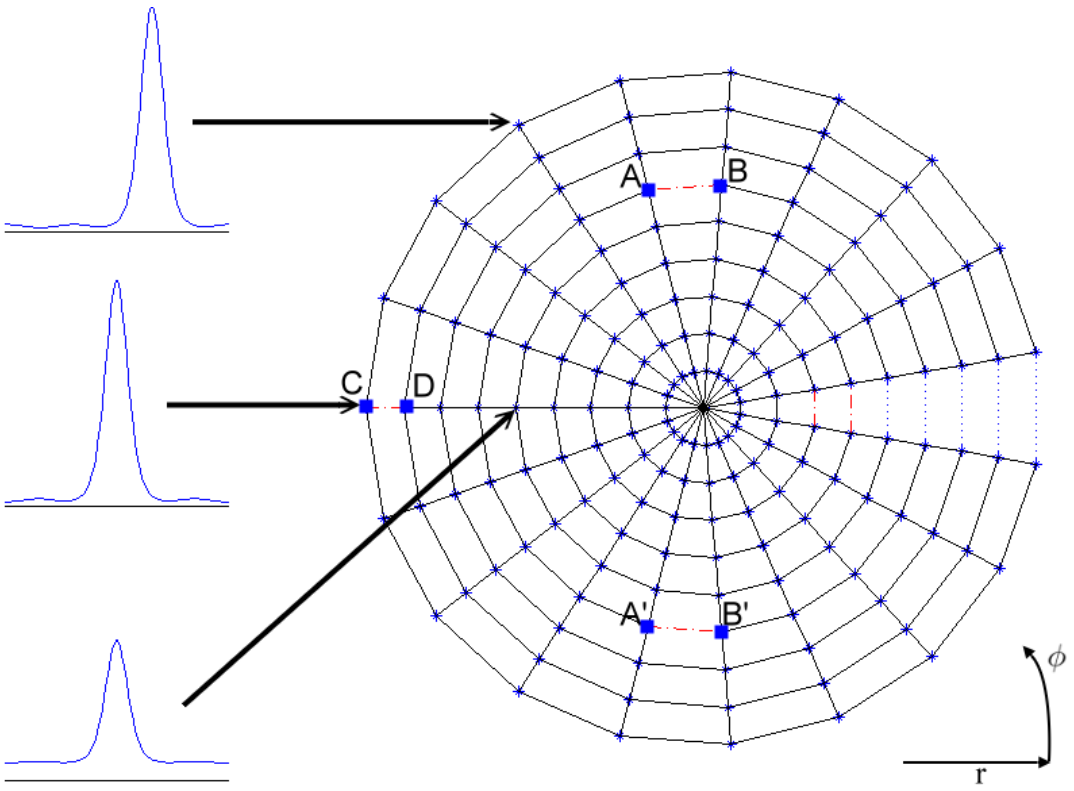

Figure 5.9.: Graph with some initial densities shown on the left.

Figure 5.9 shows the result on this graph for $K = 19$, $L = 9$, $\sigma = 2$ and $\alpha = 0.02$, as well as the initial densities at some nodes.

The graph was designed to capture the switch between going left and right, which is

clearly visible at $\phi = 0$. In addition, there are edges with singularities (implied by failed homotopies), marked A-B, A'-B' and C-D, one might not have predicted a priori.

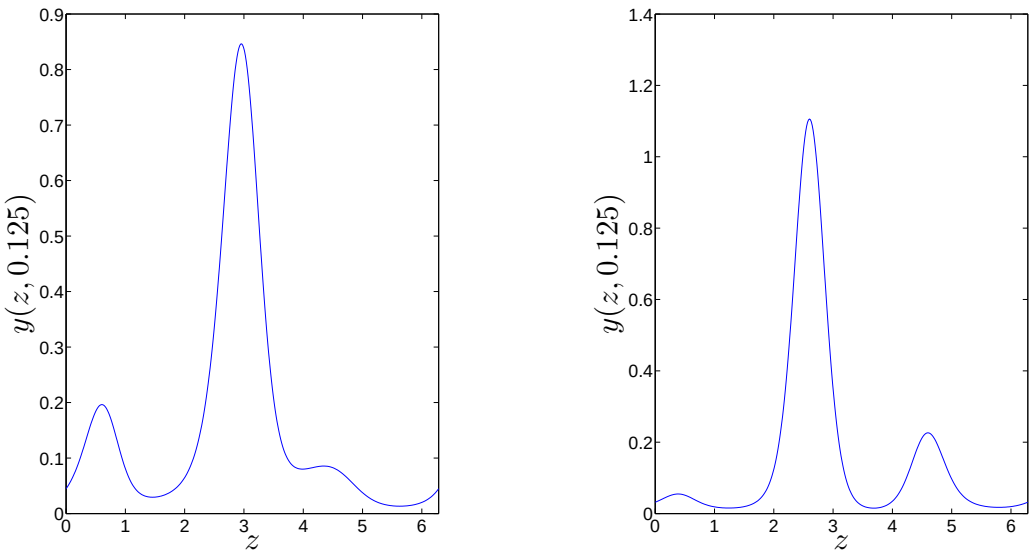

Figure 5.10.: Solutions for node A (left) and B (right) at $t = 0.125$.

Inspection of the solutions at A and B reveals that there is a difference in the strategy for gathering up mass outside the main peak. In both cases, a secondary peak is formed and later merged into the main peak, but the secondary peak is on different sides of the main peak in the globally optimal solutions for A and B (Figure 5.10). The same occurs at A' and B', which are mirror images of A and B.

The density C is the steady state and hence the solution stays constant. At D, there is more mass outside the peak, and the optimal strategy is to gather mass from both sides by moving the potential well back and forth. There are two symmetrical copies of this strategy, distinguished by the direction in which the well moves first, with identical cost. The algorithm will return whichever version is encountered first. The homotopy in a straight line from C to D fails, because the symmetry cannot be broken.

This example shows, that even in a high-dimensional setting, solution branches can be found with a very sparse grid, in this case one of only 172 nodes, and we re-emphasize that the graph $(\mathcal{V}, \mathcal{E})$ matters only for finding different local solutions: Despite the two-dimensional grid, the extremals are computed with a spatial resolution of 40 basis functions (which could be increased further without changing (V, E)) and an adaptive grid in time (which can also become arbitrarily fine if desired, and is chosen indpendently for each solution). For problems without a priori insight, relevant dimensions could be found using methods for model reduction (see e.g. [ASG01] for an overview).

6. The First Order Reliability Method

Stochastic optimization usually targets the expected value of the cost function. However, we want to develop local methods and the unbounded support of Gaussian random variables causes a global dependence of expected values on the entire domain. For this reason, we propose to optimize quantiles instead.

The *p-quantile* $Q_X(p)$ of a random variable X is defined by

$$Q_X(p) := \inf \{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq p\}.$$

If the cumulative distribution function of X is continuous, this is equivalent to

$$Q_X(1-p) := \sup \{x \in \mathbb{R} : \mathbb{P}(X \geq x) \geq p\}.$$

We want to optimize $(1-p)$-quantiles for control problems with stochastic influences, i.e. we consider the problem

$$\min_u Q_{L(u)}(1-p) = \min_u \sup \{l \in \mathbb{R} : \mathbb{P}(L(u) \geq l) \geq p\}$$

and will eventually use the min-sup structure to turn it into a differential game with the randomness as an antagonist.

6.1. Static case without optimization

The method we will develop in the next chapter is based on the *First Order Reliability Method (FORM)*, which was originally used in structural engineering ([HL74], [Rac76]) to determine the probability whether some function (e.g the stress on a structure) will exceed a given value.

In this section, we will review the basic idea of the FORM. For a more extensive treatment see e.g. the textbook [CGC06, chapter 4]. Consider the system

$$L(x) = c(x) \tag{6.1}$$
$$x = x_0 + \Sigma W \tag{6.2}$$
$$W \sim \mathcal{N}(0, I_n) \tag{6.3}$$

with $x, x_0 \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times n}$ a matrix of full rank, $L, c : \mathbb{R}^d \to \mathbb{R}$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Note that x_0 is a given state unperturbed by noise, but $x(\omega)$ is a random variable.

We define $F = F_l := \{w \in \mathbb{R}^n : L(x_0 + \Sigma w) \geq l\}$, the set of outcomes above the quantile threshold. The task is to compute

$$p := \mathbb{P}(W \in F_l)$$

for a given value $l \in \mathbb{R}$.

Generically there is a unique minimum $w^* = \operatorname{argmin}_{w \in \partial F} \|w\|$ on the boundary of F. As the density of W decays rapidly with increasing $\|W\|$, one expects $p = \int_{\{L(x) \geq l\}} e^{-\|w\|^2/2} \, \mathrm{d}w$ to be dominated by contributions close to w^* and linearly approximates F by its tangent cone $\mathcal{T}_{w^*}(F)$ at the *design point* w^*.

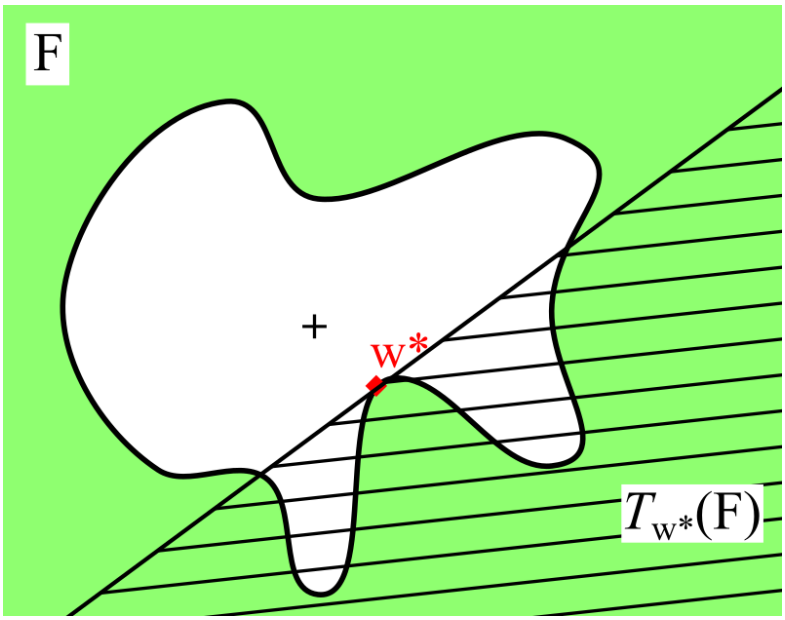

Figure 6.1.: F and its linear approximation at the design point. + is the origin.

Using that F is a superlevel set of L and assuming $\Sigma^\top \nabla L$ not to vanish at $x^* := x_0 + \Sigma w^*$, we have

$$\mathcal{T}_{w^*}(F) = \left\{ w \in \mathbb{R}^n : \left\langle w, \Sigma^\top \nabla L(x^*) \right\rangle \geq \left\langle w^*, \Sigma^\top \nabla L(x^*) \right\rangle \right\}$$

and so p is approximated by

$$\begin{aligned} p = \mathbb{P}(W \in F) &\approx \mathbb{P}\left(\left\langle W, \frac{\Sigma^\top \nabla L(x^*)}{\|\Sigma^\top \nabla L(x^*)\|} \right\rangle \geq \left\langle w^*, \frac{\Sigma^\top \nabla L(x^*)}{\|\Sigma^\top \nabla L(x^*)\|} \right\rangle \right) \\ &= 1 - \Phi\left(\left\langle w^*, \frac{\Sigma^\top \nabla L(x^*)}{\|\Sigma^\top \nabla L(x^*)\|} \right\rangle \right), \end{aligned} \tag{6.4}$$

where $\Phi(q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q} e^{-t^2/2} \, \mathrm{d}t$ is the cumulative distribution function of the standard normal distribution.

6.1.1. Finding the quantile

Determining a quantile changes the problem slightly, as instead of l we are given p and solve for the corresponding $l = Q_{L(x)}(1-p)$.

For $w^* = \operatorname{argmin}_{w \in F} \|w\|$, we have the optimality condition $w^* \perp T_{w^*}(F)$. As F is a superlevel set of $L(x_0 + \Sigma \cdot)$, we conclude that w^* is a multiple of $\Sigma^\top \nabla L(x^*)$, i.e. $\Sigma^\top \nabla L(x^*) = \mu \Sigma^\top c_x(x^*)$, and finally with (6.4) that

$$w^* = \Phi^{-1}(1-p) \frac{\Sigma^\top c_x(x^*)}{\|\Sigma^\top c_x(x^*)\|} = \Phi^{-1}(1-p) \frac{\Sigma^\top c_x(x_0 + \Sigma w^*)}{\|\Sigma^\top c_x(x_0 + \Sigma w^*)\|}.$$

Solving this equation (which is usually done numerically, e.g. by Newton's method) yields w^*. The quantile is then obtained as $Q_{L(x)}(1-p) = c(x_0 + \Sigma w^*)$.

6.1.2. Error estimate

Breitung gives an asymptotic backward error estimate for small p:

Theorem 6.1. *Define $\widetilde{p}$ such that the approximate quantile $\widetilde{l}$ computed by the FORM is the exact quantile for $\widetilde{p}$, i.e. $\widetilde{l} = Q_{L(x)}(1-\widetilde{p})$, and let Σ vary with p such that $\Phi^{-1}(1-p)\Sigma$ is constant.*[1] *Then*

$$\lim_{p \to 0} \frac{p}{\widetilde{p}} = \prod_{j=1}^{n-1} \left(1 - \frac{\kappa_j}{\|\Sigma w^*\|}\right)^{-1/2},$$

where $\kappa_j, j = 1, \ldots, n-1$ are the main curvatures of the surface $\partial F = \{w : c(x_0 + \Sigma w) = \widetilde{l}\}$ at w^*.

Proof. [Bre84, Eqs. 13a and 29] □

6.1.3. Numerical example

As a simple test of the approximation quality we consider the problem $L((x_1, x_2)) = x_1 x_2^2$, $x_0 = (1,1)$, $\Sigma = I_2$, $p = 0.05$. The method predicts a value of $Q_L(1-p) \approx 6.38$. A Monte Carlo simulation with 10^6 samples shows that this value is exceeded with probability $\mathbb{P}(L > 6.38) \approx 0.040$. At $x_0 = (10, 10)$ the function L becomes more linear in the sense that $\left\|D^2 L\right\| / \|DL\|$ decreases and the approximation quality increases with $Q_L(1-p) \approx 1.28 \cdot 10^3$ and $\mathbb{P}(L > 1.28 \cdot 10^3) \approx 0.048$ significantly closer to the target of $p = 0.05$.

[1] Note that this implies that Σw^* is independent of p.

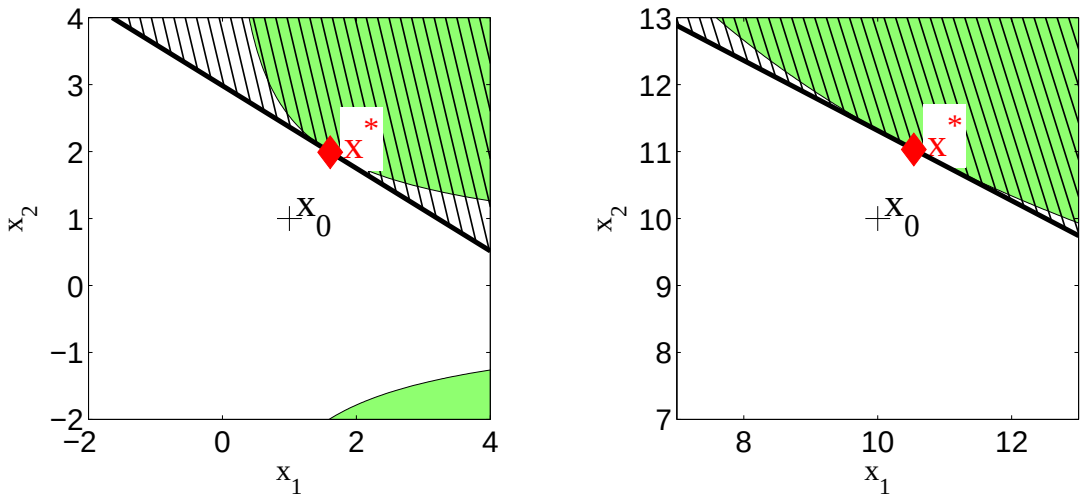

Figure 6.2.: F (in green) and T_{w^*} (hatched) for $x_0 = (1, 1)$ and $x_0 = (10, 10)$. Note that $w = x - x_0$.

6.2. Static case, interpreted as a game

We now introduce a control to optimize and provide an interpretation of the FORM as a game. The problem is now

$$\begin{aligned} L(x, u) &= c(x, u) \\ x &= x_0 + f(x_0, u) + \Sigma W \\ W &\sim \mathcal{N}(0, I_n) \\ Q_L(1-p) &= \min_u ! \end{aligned}$$

with $x, x_0 \in \mathbb{R}^d$, $u \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{d \times n}$, $L, c : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$.

The appearance of the quantile can be turned into a game in which the minimizing player first chooses the control u as well as a forbidden set F under the constraint $\mathbb{P}(W \in F) \leq p$ and then the maximizing player chooses the realization w of W under the constraint $w \notin \text{int}(F)$, i.e. the game reads

$$\begin{aligned} &\min_{(u,F)} \max_{(x,w)} L(x, u) \\ &\quad \text{s.t. } x = x_0 + f(x_0, u) + \Sigma w, \\ &\qquad \mathbb{P}(W \in F) \leq p, w \notin \text{int}(F). \end{aligned}$$

Given a certain u, the obvious choice is for the minimizing player to exclude the values of W with the highest L, i.e. $F = \{w : L \geq Q_L(1-p)\}$ and for the maximizing player

to choose a w realizing $L = Q_L(1-p)$.[2] After this consideration, only the minimization with respect to u is left and as claimed we recover $Q_L(1-p) = \min_u$!.

Proceeding as before, we linearize F at the solution (x^*, u^*) and obtain again that

$$F \approx \left\{ w : \left\langle w, \frac{\Sigma^\top \nabla_x L(x^*, u^*)}{\|\Sigma^\top \nabla_x L(x^*, u^*)\|} \right\rangle \geq \alpha \right\}, \ \alpha = \Phi^{-1}(1-p).$$

To recover the FORM we also need to approximate the behavior of the maximizing player by assuming

1. that he chooses a w maximizing the linearization of L instead of L itself, i.e. a w with $\frac{\Sigma^\top \nabla_x L(x^*,u^*)}{\|\Sigma^\top \nabla_x L(x^*,u^*)\|} \cdot w = \alpha$, and
2. that among the choices allowed by the first condition he selects the one with minimum norm and hence the most likely one, i.e. $w^* = \alpha \frac{\Sigma^\top \nabla_x L}{\|\Sigma^\top \nabla_x L\|}$.

The justification for this approximation is again that $w \notin \text{int}(F)$ with significantly higher $L(w)$ may exist in regions where the linearization is no longer valid, but have very low probability and so cause only a small error.

Setting $w^* = w$ we finally obtain the optimization problem

$$\begin{aligned} &\min_{(x,u)} \ c(x,u) \\ &\text{s.t. } \ x = x_0 + f(x,u) + \alpha \frac{\Sigma\Sigma^\top c_x(x,u)}{\|\Sigma^\top c_x(x,u)\|}. \end{aligned}$$

[2] Note that such a w is on the boundary of F and that the alternative condition $w \notin F$ was deliberately avoided as it would have introduced the technical complication of giving us a maximizing sequence converging to F instead of an actual maximizer.

7. Quantile optimization for dynamic systems

The dynamic problem is given by

$$\begin{aligned} L(\mathbf{x}, \mathbf{u}) &= \int_0^T c(x(t), u(t), t)\,\mathrm{d}t, \\ \mathrm{d}x &= f(x, u, t)\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t, \\ x(0) &= x_0 \\ Q_L(1-p) &= \min_{\mathbf{u}}! \end{aligned}$$

with $x, x_0 \in \mathbb{R}^d =: \mathcal{X}$, $u \in \mathbb{R}^m =: \mathcal{U}$, $\Sigma \in \mathbb{R}^{d\times n}$, $c : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$, $\mathbf{x} \in \mathbf{X} := \mathcal{W}^{1,\infty}([0,T];\mathcal{X})$, $\mathbf{u} \in \mathbf{U} := \mathcal{L}^\infty([0,T];\mathcal{U})$, and W an n-dimensional Wiener process with unit covariance.

We will use results from Section 2.1.2 to prove the existence of solutions, so throughout this chapter, we assume that the conditions of Lemma 2.13 and Theorems 2.14 and 2.16 are fulfilled for some $1 < p < \infty$ (the most prominent of those conditions are that f needs to be finitely generated and c coercive in u). We will also need variations of extremals to exist, and so we assume additionally that f, c and φ have second derivatives with uniform bounds on compact sets as in Theorem 2.4, and that c is strictly convex in u.

7.1. Solution with HJB

Let us first note for comparison how one would proceed using a HJB equation.

The problem can be restated as minimizing the probability to exceed a certain cost target instead of minimizing a quantile (cf. [Kan01]). Let $L^* := Q_L(1-p)$. The control minimizing $Q_L(1-p)$ is also the control minimizing the probability $p = \mathbb{P}(L > L^*)$.[1]

Introducing an additional state variable J with $J(0) = 0$ and $\dot{J} = c(x, u)$ one has $J(T) = L(\mathbf{x}, \mathbf{u})$. Now one can set

$$\varphi\left(x(T), J(T)\right) := \begin{cases} 0 & , J(T) \le q \\ 1 & , J(T) > q \end{cases}$$

with $q \in \mathbb{R}$ and has $\mathbb{E}[\varphi(x(T), J(T))] = \mathbb{P}(L > q)$.

[1] If a different control could achieve a lower probability $\mathbb{P}(L > L^*)$, it would also achieve a lower value $Q_L(1-p)$, contradicting optimality, and similarly for the reverse.

This expected value can be minimized using a HJB equation, but apart from the usual difficulties in high dimensions, one also has to deal with a discontinuous terminal condition. One does not know L^* in advance, so if a specific p is targeted, one has to solve for q, which will require solving the HJB equation several times for different values of q.

7.2. Solution with FORM

Avoiding the use of a global method, we instead want to turn the problem into a game and obtain a method similar to the FORM.

Again, we make the problem tractable through linearization and thereby will eventually obtain a first order method.

We assume that

$$W \text{ is weakly differentiable.} \tag{A1}$$

Note that this assumption is fulfilled with probability zero as it is a crucial feature of Brownian motion that its paths almost surely have unbounded variation and are nowhere differentiable. By assuming (A1) we are treating $\mathrm{d}x = f(x,u,t)\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t$ as an ordinary differential equation instead of a stochastic one and neglecting the second order term in Itô's forumla. In this sense, the assumption is a linearization. Its consequences will be studied in the error analysis below.

To keep notation similar to the deterministic setting, we would like to write $\mathrm{d}x = f\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t$ in standard ODE form, but we will eventually return to non-differentiable W and hence choose the notation

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{x} - \Sigma W) = f.$$

As only the derivative dW enters into the dynamics, we will from now on frequently refer to the noise as $\mathbf{w} \in \mathcal{L}^\infty([0,T])$, defined by $w(t) := dW_t$ almost everywhere, instead of W.

The difference between the two viewpoints discussed in the previous sections is more pronounced for this problem. Using the approach of Section 6.1, one could first find the quantile for any control u and then optimize over u. However, by choosing u first one would end up with an open-loop control, which does not react to the noise.

Of course, the notion of an optimal control is more complicated in the dynamic stochastic case, as instead of a single control $u : t \mapsto u(t)$, we are actually looking for a *non-anticipating control strategy* $U : w|_{[0,t)} \mapsto u(t)$, i.e. a strategy that determines the control $u(t)$ based on everything that has happened up to time t.[2] This is easier to accomplish by viewing the problem as a differential game.

Note that the noise does not react to the control and so w is only a (random) function of t instead of an adaptive strategy.

[2] Note that knowledge of itself and the dynamics can be built into the control strategy, so that w is indeed the only input needed to reconstruct the system's behavior.

Hence, the task is to find the solution $U^*, F^*, \mathbf{w}^*$ of the game

$$\begin{aligned} \min_{U,F} \left(\max_{(\mathbf{x},\mathbf{w})} L(\mathbf{x}, U(\mathbf{w})) \right) \\ \text{s.t. } g(\mathbf{x} - \Sigma W, U(\mathbf{w})) \\ := f(\mathbf{x}, U(\mathbf{w}), \cdot) + \Sigma \mathbf{w} - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} + (x_0 - x(0))\delta_0 \equiv 0, \\ \mathbb{P}(\mathbf{w} \in F) \leq p, \ \mathbf{w} \notin \mathrm{int}(F), \end{aligned} \tag{7.1}$$

where $g : \mathbf{X} \times \mathbf{U} \to \mathcal{Y} := (\mathcal{L}^\infty([0,T];\mathcal{X}) + \mathcal{X}\delta_0) \subset \mathcal{W}^{-1,\infty}([0,T];\mathcal{X})$

We again restrict F to a linearization (thereby changing the game to the control player's disadvantage), in this case to

$$F = \{\langle \mathbf{w}, M(\mathbf{w}) \rangle_{\mathcal{L}^2} \leq \alpha\}, \ \alpha = \Phi^{-1}(1-p),$$

where M is a non-anticipating strategy.

Instead of looking for the full control strategy (U^*, M^*), we start by finding the optimal answer $\mathbf{u}^* := U^*(\mathbf{w}^*), \mu^* := M^*(\mathbf{w}^*)$ to $\mathbf{w}^* := \mathrm{d}W^*$, where U^*, M^*, W^* are the solution of the above game. How to find the optimal answer to an arbitrary noise $\mathbf{w}$ will be considered later.

$(\mathbf{u}^*, \mu^*, \mathbf{w}^*)$ are a Nash equilibrium, i.e. a saddle point of

$$\begin{aligned} \min_{\mathbf{u},\mu} \max_{\mathbf{x},\mathbf{w}} L(\mathbf{x}, \mathbf{u}) \\ \text{s.t. } g(\mathbf{x} - \Sigma W, \mathbf{u}) \equiv 0, \\ g_2(\mathbf{w}, \mu) := \alpha - \langle \mathbf{w}, \mu \rangle_{\mathcal{L}^2} \geq 0, \\ g_3(\mu) := \gamma^2 - \|\mu\|^2_{\mathcal{L}^2} = 0, \end{aligned} \tag{7.2}$$

with $\alpha = \Phi^{-1}(1-p)$ and $\gamma^2 = 1$.

In order to proceed, we need g_2 to be active[3] (i.e. $g_2 = 0$), which we ensure (see Remark 7.1 below) by assuming from now on that

(A2) if $\left\| \Sigma^\top \lambda^{(n)} \right\|_{\mathcal{L}^2} \to 0$ $(n \to \infty)$ for a sequence $\left(\mathbf{x}^{(n)}, \mathbf{u}^{(n)}, \mu^{(n)}, \mathbf{w}^{(n)} \right)$ which is admissible and a local minimum w.r.t. $\left(\mathbf{x}^{(n)}, \mathbf{u}^{(n)}, \mu^{(n)} \right)$ for all n, then this sequence has an accumulation point $(\mathbf{x}, \mathbf{u}, \mu, \mathbf{w})$ which is a local minimum w.r.t. $\mathbf{w}$.

This assumption excludes not only the existence of maxima w.r.t. $\mathbf{w}$ but also stagnation that could lead to a "maximum at ∞", which will be useful later.

[3] Otherwise we would have the KKT conditions $\lambda_2 \geq 0$, $\lambda_2 g_2(\mathbf{w}, \mu) = 0$. With $g_2 \neq 0$, $\lambda_2 = 0$ would imply $\lambda_3 = 0$ and hence no information on $\mathbf{w}$ or μ.

The extremality condition then reads

$$-\nabla_{(\mathbf{x},\mathbf{u},\mathbf{w},\mu)}L = -\begin{pmatrix} c_x \\ c_u \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} g_x^* \\ g_u^* \\ \Sigma^\top \\ 0 \end{pmatrix} \lambda + \lambda_2 \begin{pmatrix} 0 \\ 0 \\ -\mu \\ -\mathbf{w} \end{pmatrix} + \lambda_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ -2\mu \end{pmatrix},$$

with $\lambda \in \mathcal{Y}^*$, $\lambda_2, \lambda_3 \in \mathbb{R}$

It follows that μ, $\mathbf{w}$ and $\Sigma^\top\lambda$ are multiples of each other. From $\|\mu\|_{\mathcal{L}^2} = \gamma = 1$ we see that

$$\mu = \gamma \frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}} = \frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}$$

and from $g_2 = 0$ that

$$\mathbf{w} = \alpha \frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}.$$

The Lagrange multipliers for g_2 and g_3 are

$$\lambda_2 = \frac{\left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2}}{\gamma},$$
$$\lambda_3 = -\frac{\alpha\left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2}}{2\gamma^3}.$$

Remark 7.1. *To understand* (A2)*, note that by the Shadow Price Theorem 1.8*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}L = \lambda_2\frac{\mathrm{d}}{\mathrm{d}\alpha}g_2 = \lambda_2 = \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2},$$

i.e. if $\Sigma^\top\lambda \neq 0$ then increasing α increases L^* and hence a $\mathbf{w}$ with $g_2 > 0$ that does not fully use the allowed α cannot be optimal. (A2) *rules out that $\Sigma^\top\lambda = 0$ for a maximal $\mathbf{w}$. Hence $g_2 = 0$ and the division by $\left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2}$ is allowed.*

Note that there is a second solution in which μ, $\mathbf{w}$, and λ_2 all change their sign. In particular, $\lambda_2 < 0$ for this solution. Arguing as in Remark 7.1, we see that there are $\mathbf{w}$ with $g_2 > 0$ which achieve higher values L. It follows that the second solution is not a maximum with respect to $\mathbf{w}$.

Again, $\mathbf{w}$ can now be eliminated as an independent variable and we obtain an optimal

control problem involving its own adjoint:

$$\min_{\mathbf{u}} \left(\max_{(\mathbf{x},\lambda)} L(\mathbf{x},\mathbf{u}) \right) \tag{7.3}$$
$$\text{s.t. } g(\mathbf{x},\mathbf{u}) := f(x,u,\cdot) + \alpha \frac{\Sigma\Sigma^\top \lambda}{\|\Sigma^\top \lambda\|_{\mathcal{L}^2}} - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} \equiv 0,$$
$$x(0) = x_0,$$
$$-\frac{\mathrm{d}}{\mathrm{d}t}\lambda = c_x + f_x^\top \lambda,$$
$$\lambda(T) = 0.$$

Note that the equations for λ are an extensive form of $c_x = -g_x^* \lambda$.

For a (not necessarily extremal) trajectory in a deterministic optimal control problem, one can start with a given $\mathbf{u}$, integrate forward to obtain $\mathbf{x}$ and finally backward to get λ. In (7.3), λ appears in $\dot{x}$ and so one has to solve for $\mathbf{x}$ and λ simultaneously, and we need to show that a solution exists. The following proofs assume that the reader is familiar with the arguments in Section 2.1.2.

Lemma 7.2. *The maximum over* $(\mathbf{x},\lambda)$ *in* (7.3) *exists for all* $\mathbf{u} \in \mathbf{U}$.

Proof. Consider for $\mathbf{u}$ fixed

$$\max_{(\mathbf{x},\mathbf{w})} L(\mathbf{x},\mathbf{u}) \tag{7.4}$$
$$\text{s.t. } g(\mathbf{x},\mathbf{u}) := f(x,u,\cdot) + \alpha\Sigma\mathbf{w} - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} \equiv 0,$$
$$x(0) = x_0,$$
$$1 - \|\mathbf{w}\|_{\mathcal{L}^2}^2 = 0.$$

Note that the dynamics are finitely generated w.r.t. to w. We do not get a bound on $\|\mathbf{w}\|_{\mathcal{L}^2}$ from coercivity but have one directly in the constraint. So we have a minimizing subsequence $\mathbf{x}^i \to \mathbf{x}^*$, $\mathbf{w}^i \rightharpoonup \mathbf{w}^*$ in $\mathcal{L}^2$ as $i \to \infty$. As L does not directly depend on $\mathbf{w}$, we can immediately conclude from continuity w.r.t. $\mathbf{x}$ that $L(\mathbf{x}^i,\mathbf{u}) \to L(\mathbf{x}^*,\mathbf{u})$. All $\mathbf{w}^i$ fulfill $\langle \mathbf{w}^i, \mathbf{w}^* \rangle_{\mathcal{L}^2} \leq \|\mathbf{w}^i\|_{\mathcal{L}^2} \|\mathbf{w}^*\|_{\mathcal{L}^2} = \|\mathbf{w}^*\|_{\mathcal{L}^2}$ and due to the weak convergence, $\langle \mathbf{w}^*, \mathbf{w}^* \rangle_{\mathcal{L}^2} \leq \|\mathbf{w}^*\|_{\mathcal{L}^2}$, i.e. $\|\mathbf{w}^*\|_{\mathcal{L}^2} \leq 1$. By assumption (A2), $\|\mathbf{w}^*\|_{\mathcal{L}^2} < 1$ would contradict optimality, so $\|\mathbf{w}^*\|_{\mathcal{L}^2} = 1$.

Finally, let $\lambda^* = \lambda(\mathbf{x}^*)$ be defined by $-\frac{\mathrm{d}}{\mathrm{d}t}\lambda = c_x(x^*,u,t) + f_x(x^*,u,t)^\top \lambda$, $\lambda(T) = 0$. The extremality condition

$$-\nabla_{(\mathbf{x},\mathbf{w})} L = -\begin{pmatrix} c_x \\ 0 \end{pmatrix} = \begin{pmatrix} g_x^* \\ \Sigma^\top \end{pmatrix} \lambda + \lambda_2 \begin{pmatrix} 0 \\ -2\mathbf{w} \end{pmatrix}$$

shows that $\Sigma^\top \lambda^*$ is a multiple of $\mathbf{w}^*$ and with $\|\mathbf{w}^*\|_{\mathcal{L}^2} = 1$ we have $\Sigma\mathbf{w}^* = \frac{\Sigma\Sigma^\top\lambda^*}{\|\Sigma^\top\lambda^*\|_{\mathcal{L}^2}}$. It follows that λ^* is admissible in (7.3). □

Remark 7.3. *For $\alpha = 0$, λ does not appear in g and (7.3) coincides with the deterministic problem, for which λ is unique. This remains so unless a bifurcation occurs as $|\alpha|$ grows. In the typical case for $|\alpha|$ small, we expect (7.3) to "maximize" over only a single admissible λ.*

Theorem 7.4. *Problem (7.3) has a solution.*

Proof. Let $\mathbf{x}^i \to \mathbf{x}^*$, $\mathbf{u}^i \rightharpoonup \mathbf{u}^*$ be a minimizing sequence of (7.3) and λ^i the corresponding λ. By the same arguments as in Section 2.1.2, we see that $(\mathbf{x}^*, \mathbf{u}^*) \in \mathcal{W}^{1,\infty} \times \mathcal{L}^\infty$ and that $L(\mathbf{x}^*, \mathbf{u}^*)$ attains the infimum, but we need to show that the corresponding λ solves the maximization.

We define $\lambda^i := \lambda(\mathbf{x}^i)$ with the mapping $\mathbf{x} \mapsto \lambda, \mathcal{L}^p \to \mathcal{W}^{1,p}$ as in Lemma 7.2. This mapping is continuous as it is the solution operator of an ODE with Lipschitz coefficients and hence we have $\lambda^* := \lambda(\mathbf{x}^*) = \lim_{i\to\infty} \lambda^i$.

To show that λ^* solves the maximization, we return to the underlying $\mathbf{w}^*$ as in the proof of Lemma 7.2. Assume the maximization (7.4) with $\mathbf{u} = \mathbf{u}^*$ is solved by $(\widehat{\mathbf{x}}, \widehat{\mathbf{w}}) \neq (\mathbf{x}^*, \mathbf{w}^*)$. Then $L(\widehat{\mathbf{x}}, \mathbf{u}^*) = L(\mathbf{x}^*, \mathbf{u}^*) + \epsilon$ for some $\epsilon > 0$.

As f is finitely generated (i.e. linear in u), $\mathbf{u}_i \rightharpoonup \mathbf{u}^*$ weakly suffices for $\mathbf{x}(\mathbf{u}_i, \widehat{\mathbf{w}}) \to \mathbf{x}(\mathbf{u}^*, \widehat{\mathbf{w}})$ strongly in $\mathcal{L}^p$. Hence $L(\mathbf{x}(\mathbf{u}_i, \widehat{\mathbf{w}}), \mathbf{u}_i) \to L(\mathbf{x}^*, \mathbf{u}^*) + \epsilon$ but $L(\mathbf{x}(\mathbf{u}_i, \mathbf{w}_i), \mathbf{u}_i) = L(\mathbf{x}_i, \mathbf{u}_i) \to L(\mathbf{x}^*, \mathbf{u}^*)$, which for i large enough contradicts the optimality of $\mathbf{w}_i$ in (7.4) at $\mathbf{u} = \mathbf{u}_i$. □

7.3. Time consistency

W, being a random variable, is almost certain to deviate from its realization in the reference solution above, which raises the question how the control should respond to such a deviation. One might, at any time t, use the control solving the $(1-p)$-quantile problem on $[t, T]$ but we need to ask ourselves whether the combination of these controls can in any sense be considered to be a solution for the original problem on $[0, T]$.

The Dynamic Programming Principle (Theorem 5.6) for deterministic problems states that any subtrajectory $(\mathbf{x}^*, \mathbf{u}^*)|_{[t,T]}$ is an optimal solution for the subproblem restricted to the time interval $[t, T]$ with initial condition $x^*(t)$. We will study whether an equivalent statement holds for quantile optimization.

In general, we have $\left\| \mu|_{[t,T]} \right\|_{\mathcal{L}^2} < \|\mu\|_{\mathcal{L}^2}$ and an analogous decrease in $\langle \mathbf{w}, \mu \rangle_{\mathcal{L}^2}$, so we cannot expect $g_2 = 0$ and $g_3 = 0$ to be fulfilled without modifications. Instead, we need to retain more information about that past trajectory than just the current state $x(t)$ and demand that the entire trajectory – past and future – fulfills the constraints. Hence we get for the trajectory on $[t, T]$ the conditions

$$\begin{aligned} \left\langle \mathbf{w}|_{[t,T]} \, , \, \mu|_{[t,T]} \right\rangle_{\mathcal{L}^2} &\leq \alpha - \left\langle \mathbf{w}|_{[0,t]} \, , \, \mu|_{[0,t]} \right\rangle_{\mathcal{L}^2} =: \tilde{\alpha}, \\ \left\| \mu|_{[t,T]} \right\|_{\mathcal{L}^2}^2 &= \gamma^2 - \left\| \mu|_{[0,t]} \right\|_{\mathcal{L}^2}^2 =: \tilde{\gamma}^2, \end{aligned}$$

Subtrajectories of an admissible trajectory $(\mathbf{x}, \mathbf{u}, \mu, \mathbf{w})$ fulfill these conditions for values $\widetilde{\alpha}$ and $\widetilde{\gamma}$ varying according to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\alpha}(t) &= -\langle w(t), \mu(t)\rangle\,,\\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\widetilde{\gamma}(t)\right)^2 &= -\left\|\mu(t)\right\|^2.\end{aligned}$$

To maintain the normalization $\|\mu\| = 1$ (which is required for $\alpha = \Phi^{-1}(1-p)$ to hold), we perform the rescaling $\alpha := \frac{\widetilde{\alpha}}{\widetilde{\gamma}}$, $\gamma := 1$, which does not change the solution (apart from rescaling μ).

$\left(\mathbf{x}|_{[t,T]}\,, \mathbf{u}|_{[t,T]}\,, \mathbf{w}|_{[t,T]}\,, \mu|_{[t,T]} \,/\, \left\|\mu|_{[t,T]}\right\|_{\mathcal{L}^2}\right)$ then fulfill the optimality condition and the constraint g and it remains to determine the values $\widetilde{\alpha}$ and $\widetilde{\gamma}$ for which g_2 and g_3 are fulfilled.

The question of admissibility is also interesting for random outcomes $\mathbf{w} \neq \mathbf{w}^*$, so for $\mu = \mu^* = \frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|}$ but $\mathbf{w}$ arbitrary, we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\alpha}(t) &= -\langle w(t), \mu(t)\rangle = -\frac{w(t)^\top\Sigma^\top\lambda(t)}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}\\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\widetilde{\gamma}(t)\right)^2 &= -\frac{\left\|\Sigma^\top\lambda(t)\right\|^2}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}^2}\\ \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t) &= -\frac{\left\|\Sigma^\top\lambda(t)\right\|^2}{2\widetilde{\gamma}(t)\,\|\Sigma^\top\lambda\|_{\mathcal{L}^2}^2}\\ \frac{\mathrm{d}}{\mathrm{d}t}\alpha(t) &= \left(\frac{\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\alpha}(t)}{\widetilde{\alpha}(t)} - \frac{\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)}{\widetilde{\gamma}(t)}\right)\frac{\widetilde{\alpha}(t)}{\widetilde{\gamma}(t)}\\ &= -\frac{w(t)^\top\Sigma^\top\lambda(t)}{\left\|\Sigma^\top\lambda|_{[t,T]}\right\|_{\mathcal{L}^2}} + \frac{\left\|\Sigma^\top\lambda(t)\right\|^2}{2\left\|\Sigma^\top\lambda|_{[t,T]}\right\|_{\mathcal{L}^2}^2}\alpha(t).\end{aligned} \tag{7.5}$$

To obtain the last line we have used $\widetilde{\gamma}(t) = \left\|\mu|_{[t,T]}\right\|_{\mathcal{L}^2} = \frac{\left\|\Sigma^\top\lambda|_{[t,T]}\right\|}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}$.

For the reference trajectory, the computation is simpler. With $\mathbf{w} = \mathbf{w}^* = \alpha(0)\frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|}$ we directly have

$$\widetilde{\alpha}(t) = \alpha(0)\widetilde{\gamma}(t)^2$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha(t) = -\frac{\left\|\Sigma^\top\lambda(0)\right\|^2}{2\,\|\Sigma^\top\lambda\|_{\mathcal{L}^2}^2}\alpha(0). \tag{7.6}$$

In particular we see that α is decreasing for the reference trajectory $\mathbf{w}^*$, i.e. even if the realization of $\mathbf{w}$ coincides with $\mathbf{w}^*$, the subtrajectories are no longer optimal solutions if

we keep α constant.

Value function

$\widetilde{\alpha}(t)$ defined above keeps track of how much of the allowed value for $\langle \mathbf{w}, \mu \rangle_{\mathcal{L}^2}$ has been used up at time t and in that sense can be understood to be a part of the state.

In fact, if we include its rescaled version $\alpha(t)$, as defined by (7.5), in the state and define $V(x_t, \alpha, t)$ to be the value L^* of (7.3) restricted to $[t, T]$ with the initial condition $x(t) = x_t$, then V is an upper bound on the value function:

Theorem 7.5. *Let* $\mathbf{w} \in \mathcal{L}^\infty([0,T]; \mathbb{R}^n)$, $x(0) \in \mathbb{R}^d$, $\alpha(0) \in \mathbb{R}$. *Let* $x(t), \alpha(t)$, $t \in [0, T^*)$ *be defined by*

$$\begin{aligned}
\dot{x}(t) &= f\left(x(t), u^*(t)\right) + \Sigma w(t), \\
\dot{\alpha}(t) &= -\left\langle w(t), \mu(t) \right\rangle + \frac{\left\|\Sigma^\top \lambda^{(t)}(t)\right\|^2}{2\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}^2} \alpha(t), \\
\mu(t) &:= \begin{cases} \frac{\Sigma^\top \lambda^{(t)}(t)}{\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}}, & \text{for } \left\|\Sigma^\top \lambda^{(t)}\right\| > 0 \\ 0 & \text{for } \left\|\Sigma^\top \lambda^{(t)}\right\| = 0 \end{cases},
\end{aligned}$$

where $u^*(t)$ *and* $\lambda^{(t)}$ *are from the solution of* (7.3) *on* $[t, T]$ *and* $[0, T^*)$ *is the maximal interval on which a solution of the ODE exists. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^t c(x(s), u^*(s), s)\,\mathrm{d}s + V(x(t), \alpha(t), t)\right) \leq 0$$

for $t \in [0, T^*)$ *with equality when* $V(\cdot, \alpha(t), t)$ *is differentiable.*

Proof. If also $\mathbf{w} = \mathbf{w}^*$, then subtrajectories on $[t+\tau, T]$, $\tau \geq 0$ are solutions of (7.2) on that interval, hence $V(x(t+\tau), \alpha(t+\tau), t+\tau) = \int_{t+\tau}^T c(x(s), u^*(s), s)\,\mathrm{d}s$ and taking the derivative at $\tau = 0$ gives

$$\begin{aligned}
\dot{V} :=& \frac{\mathrm{d}}{\mathrm{d}\tau} V\left(x(t) + \tau\left(f\left(x(t), u^*(t)\right) + \alpha(t)\frac{\Sigma\Sigma^\top \lambda(t)}{\left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}}\right), \alpha(t) + \tau \frac{\left\|\Sigma^\top \lambda(t)\right\|^2}{2\left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}^2}\alpha(t), t + \tau\right)\Bigg|_{\tau=0} \\
&+ c(x(t), u^*(t), t) = 0,
\end{aligned}$$

where we have used $w^*(t) = \frac{\left\|\Sigma^\top \lambda(t)\right\|^2}{\left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}^2}\alpha(t)$.

By definition, V equals the value L of the associated solution of (7.2). For arbitrary

$\mathbf{w}$, the Shadow Price Theorem[4] 1.8 tells us that varying $\mathbf{w}$ gives

$$\begin{aligned}\mathrm{d}\dot{L} &= \frac{\mathrm{d}}{\mathrm{d}x_t} L\,\mathrm{d}\dot{x}(t) + \frac{\mathrm{d}}{\mathrm{d}\alpha} L\,\mathrm{d}\dot{\alpha}(t)\\ &= \lambda(t)\cdot\Sigma\,\mathrm{d}w(t) + \lambda_2\left\langle -\mathrm{d}w(t), \mu(t)\right\rangle\\ &= \lambda(t)\cdot\Sigma\,\mathrm{d}w(t) + \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2}\frac{-\Sigma^\top\lambda(t)}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}\cdot\mathrm{d}w(t)\\ &= 0.\end{aligned}$$

In particular this holds for $\mathrm{d}\mathbf{w} = \mathbf{w} - \mathbf{w}^*$ and so, if the varied minimum remains the global minimum, then the claim holds with equality. If the global minimum varies discontinuously, then $V(\cdot, \alpha(t), t)$ is not differentiable and $\dot{V} \leq \dot{L}$. □

Note that $\left\|\Sigma^\top\lambda\big|_{[t,T]}\right\|_{\mathcal{L}^2} \to 0$ in the denominator of (7.5) as $t \to T$, making a blow-up a common occurrence for arbitrary $\mathbf{w}$. For $\langle\mathbf{w}, \mu\rangle = \alpha$ it can be compensated by the numerator also going to zero (cf. (7.6), the behavior of the reference trajectory), otherwise the sign of the blow-up tells us whether $\mathbf{w}$ is admissible:

Lemma 7.6. *Under the assumptions and definitions of Theorem 7.5, if* $\langle\mathbf{w}, \mu\rangle_{\mathcal{L}^2} \neq \alpha(0)$, *then*

$$\begin{aligned}\alpha(t) &\xrightarrow[t\to T^*]{} +\infty \Leftrightarrow \langle\mathbf{w}, \mu\rangle_{\mathcal{L}^2} < \alpha(0),\\ \alpha(t) &\xrightarrow[t\to T^*]{} -\infty \Leftrightarrow \langle\mathbf{w}, \mu\rangle_{\mathcal{L}^2} > \alpha(0).\end{aligned}$$

Proof. We have $\alpha(t) = \frac{\tilde{\alpha}(t)}{\tilde{\gamma}(t)}$, so $\alpha(t)$ stops being differentiable at T^* only if $\tilde{\gamma}(T^*) = 0$. Hence $\tilde{\gamma}(t) = 0$ and $\mu(t) = 0$ for all $t \geq T^*$ giving us

$$\tilde{\alpha}(T^*) = \tilde{\alpha}(0) - \int_0^{T^*} w(t)\cdot\mu(t)\,\mathrm{d}t = \alpha(0) - \langle\mathbf{w}, \mu\rangle_{\mathcal{L}^2}\,.$$

The claim follows from $\alpha(t) = \frac{\tilde{\alpha}(t)}{\tilde{\gamma}(t)}$ and $\tilde{\gamma}$ being non-negative. □

Furthermore, if a blow-up occurs at $T^* < T$, then it has become impossible to stay below the target cost regardless of $\mathbf{w}$:

Lemma 7.7. *If* $\alpha(t) \xrightarrow[t\to T^*]{} -\infty$ *for some* $\mathbf{w}|_{[0,T^*]}$ *with* $\mathbf{x}$, α *as in Theorem 7.5 on* $[0, T^*]$, *then*

$$\int_{T^*}^{T} c(x(t), u(t), t)\,\mathrm{d}t \geq \lim_{t\to T^*} V(x(t), \alpha(t), t)$$

for all $(\mathbf{u}, \mathbf{w})|_{(T^*,T]}$ *and* $\dot{x}(t) = f(x(t), u(t), t) + \Sigma w(t)$ *on* $(T^*, T]$.

[4] Note that the Shadow Price Theorem merely requires z^* to be a critical point, so it does not matter that we have a saddle point instead of a minimum.

Proof. Let $\widetilde{L}(\mathbf{w}) := \int_{T^*}^{T} c(x(t), u^*(t), t)\,\mathrm{d}t$ (note that $\mathbf{x}$ and $\mathbf{u}^*$ depend on $\mathbf{w}$), $\widetilde{V} := \lim_{t\to T^*} V(x(t), \alpha(t), t)$, $\widetilde{L}^* := \inf_{\mathbf{w}} \widetilde{L}(\mathbf{w})$ and assume by way of contradiction that $\widetilde{L}^* < \widetilde{V}$. Then there exists $\widetilde{\mathbf{w}}$ such that $\widetilde{L}(\widetilde{\mathbf{w}}) \leq (\widetilde{L}^* + \widetilde{V})/2 < \widetilde{V}$. We consider now

$$\widetilde{L}_\eta := \widetilde{L}\left((1-\eta)\widetilde{\mathbf{w}} + \widetilde{\alpha}(\eta)\frac{\Sigma^\top\lambda}{\|\Sigma^\top\lambda\|_{\mathcal{L}^2}}\right), \eta \in [0,1]$$

and want to find $\widetilde{\alpha} : [0,1] \to \mathbb{R}$, $\widetilde{\alpha}(0) = 0$ such that $\widetilde{L}_\eta$ is constant w.r.t. η. This implies $\frac{\mathrm{d}}{\mathrm{d}\eta}\widetilde{L}_\mu = -\langle \Sigma\widetilde{\mathbf{w}}, \lambda\rangle + \frac{\mathrm{d}\widetilde{\alpha}}{\mathrm{d}\eta}\left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2} = 0$, which is fulfilled, e.g., if

$$\frac{\mathrm{d}\widetilde{\alpha}}{\mathrm{d}\eta} = \begin{cases} \left\langle \widetilde{\mathbf{w}}, \Sigma^\top\lambda\right\rangle_{\mathcal{L}^2} / \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2}, & \text{for } \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2} > 0 \\ 0, & \text{for } \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2} = 0 \end{cases}.$$

The solution of this ODE exists because $\left|\frac{\mathrm{d}\widetilde{\alpha}}{\mathrm{d}\eta}\right| \leq \|\widetilde{\mathbf{w}}\|_{\mathcal{L}^2}$. By definition of V it follows that

$$V(x(T^*), \widetilde{\alpha}(1), T^*) \leq \widetilde{L}(\widetilde{\mathbf{w}}) < \widetilde{V} = \lim_{t\to T^*} V(x(t), \alpha(t), t).$$

As $\alpha(t) \xrightarrow[t\to T^*]{} -\infty$ and $\frac{\mathrm{d}}{\mathrm{d}\alpha}V = \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2} \geq 0$, we have $\lim_{t\to T^*} V(x(t), \alpha(t), t) \leq V(x(T^*), \bar{\alpha}, T^*)$ for any finite $\bar{\alpha}$, so this is a contradiction. □

This unachievability is reflected in the fact that the FORM approach does not provide a control strategy after T^*.

Remark 7.8. *Note that $\alpha \to +\infty$ at a $T^* < T$ would violate* (A2)*. Of course, this could still occur if the method is applied to a problem not fulfilling* (A2)*. Even then one could not get a statement analogous to lemma 7.7 with "$\leq$" because the proof used the infimum in the definition of V.*

Finally, we need to relate the cost actually incurred to the value function and consider V at $T^* = T$. For $\alpha = 0$, the value function $V(\cdot, \alpha, t) = \int_t^T c\,\mathrm{d}s$ goes to 0 for $t \to T$ as the corresponding trajectory $(\mathbf{x}, \mathbf{u})$ remains bounded. Combining Theorem 7.5 with $\frac{\mathrm{d}}{\mathrm{d}\alpha}V = \left\|\Sigma^\top\lambda\right\|_{\mathcal{L}^2} \geq 0$ and Lemma 7.6, we find (with $\mathbf{u}^*$ as in Theorem 7.5)

Theorem 7.9. *$L(\mathbf{x}, \mathbf{u}^*(\mathbf{w}), \mathbf{w}) = \int_0^T c(x, u^*, t)\,\mathrm{d}t \leq V(x(0), \alpha(0), 0)$ for all $\mathbf{w} \in \mathcal{L}^\infty([0,T])$ with $\alpha(t) \xrightarrow[t\to T^*]{} +\infty$, i.e. for all admissible outcomes $\mathbf{w}$ up to a zero set.*

7.4. Alternative choices for α

Restatement as probability minimization

Let $Q_L^{[a,b]}$ be the quantile function of the problem restricted to the time interval $[a,b]$ with the cost function $L^{[a,b]} = \int_a^b c(x,u)\,\mathrm{d}t$.

After L^* has been determined, the minimization of $\mathbb{P}(L > L^*)$ is a goal that can be pursued consistently in time (cf. Section 7.1) and can be translated back to a quantile optimization by finding a $p(t)$ such that

$$Q_L^{[t,T]}(1 - p(t)) = L^* - \int_0^t c(x, u)\,\mathrm{d}t. \tag{7.7}$$

Remark 7.10. *If* (A1) *holds, then* $V = L^* - \int_0^t c(x,u)\,\mathrm{d}t$ *for all* t*, unless the "*$<$*" case of Theorem 7.5 occurs on a set of nonzero measure, and we obtain that* $p(t)$ *is the probability corresponding to* $\alpha(t)$ *as determined by* (7.5)*. If* W *is a Browninan motion, this no longer holds (cf. section 7.6).*

For practical applications this restatement has the advantage that one only needs to measure the cost c.

α locked to reference trajectory

A pure focus on minimizing only a certain quantile might not be desirable in practice. In particular, we have seen that this would result in giving up completely if the target value has become unattainable. As an alternative, we suggest to act as if $\mathbf{w}$ were to follow the reference trajectory and vary α according to (7.6). This would still give approximately the same control if $\mathbf{w}$ is close to the most critical outcome (i.e the reference trajectory), but a generally more well-behaved behavior because $|\alpha(t)|$ is non-increasing.

7.5. Behavior near a target state

Cost functionals are often designed with a target state (x^*, u^*) with $f(x^*, u^*) = 0$ such that $c(x^*, u^*)$ is minimal, in which case the optimal trajectory will typically converge to the target state on larger time intervals. (This was assumed in Chapter 5.)

This situation requires some attention, as our approach relies on the problem being essentially linear and treats only the component of the noise pointing in the direction of the gradient (i.e. the adjoint λ) as increasing the cost (see also the error analysis in the next section). At the target state x^*, this assumption breaks down as the quadratic term begins to dominate in c and any noise will increase the cost, regardless of direction.

It follows that the control derived by our quantile optimization procedure should be used to guide the state towards its target but not to keep it there. We thus require a method for the later task and a criterion when to switch.

Close to the target x^* the problem is approximately linear-quadratic. As the optimal control for the deterministic LQ problem also optimizes the expected value for the stochastic version, we use the optimal control for the problem without noise.

The switching strategy could be based on the distance to x^* with the LQ approximation around x^* being used to compute how far the state is expected to stray from the target due to noise (the controlled dynamics in the LQ case is an Ornstein-Uhlenbeck process).

7.6. Error analysis in V

We can estimate the error in V by using an Itô version of Theorem 7.5 that does not require W to be weakly differentiable (Assumption (A1)):

Writing x and α as random variables, we have the SDEs

$$\begin{aligned}
\mathrm{d}x_t &= f(x, u^*, t) + \Sigma \,\mathrm{d}W_t, \\
\mathrm{d}\alpha_t &= \frac{\left\|\Sigma^\top \lambda^{(t)}(t)\right\|^2}{2\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}} \alpha_t \,\mathrm{d}t - \frac{\Sigma^\top \lambda^{(t)}(t)}{\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}} \cdot \mathrm{d}W_t.
\end{aligned}$$

By Itô's lemma,

$$\begin{aligned}
&V\left((x,\alpha)_t, t\right) := V(x_t, \alpha_t, t) \\
=&V\left((x,\alpha)_0, 0\right) + \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} V \,\mathrm{d}s + \int_0^t \frac{\mathrm{d}}{\mathrm{d}(x,\alpha)} V \,\mathrm{d}(x,\alpha)_s \\
&+ \frac{1}{2}\int_0^t \frac{\mathrm{d}^2}{\mathrm{d}(x,\alpha)^2} V : \mathrm{d}\left\langle (x,\alpha)\right\rangle_s \\
=&V\left((x,\alpha)_0, 0\right) - \int_0^t c(x,u,s)\,\mathrm{d}s + \frac{1}{2}\int_0^t \frac{\mathrm{d}^2}{\mathrm{d}(x,\alpha)^2} V : \mathrm{d}\left\langle (x,\alpha)\right\rangle_s .
\end{aligned}$$

Similarly to Theorem 7.9, we hence have

Theorem 7.11.

$$\int_0^T c(x,u^*,t)\,\mathrm{d}t \le V(x(0),\alpha(0),0) + \frac{1}{2}\int_0^t \frac{\mathrm{d}^2}{\mathrm{d}(x,\alpha)^2} V : \mathrm{d}\left\langle (x,\alpha)\right\rangle_s$$

for all realizations of W with $\alpha(t) \xrightarrow[t\to T^*]{} +\infty$, i.e. for all admissible outcomes up to a zero set.

This estimate shows that, as in the static case, the error (i.e. the last integral in Theorem 7.11) depends on the magnitude of the quadratic part, but unfortunately we cannot give any further bounds.

Note that the last term is still defined and finite even when the second derivative of V is measure valued, cf. [Eis01]. This means that the Theorem holds even if V consists of multiple branches (cf. Chapter 4).

For reference we note that for the quadratic variations in the error term we have

$$\begin{aligned}
\mathrm{d}\left\langle x\right\rangle_t &= \Sigma\Sigma^\top, \\
\mathrm{d}\left\langle x,\alpha\right\rangle_t &= -\frac{\Sigma\Sigma^\top \lambda^{(t)}(t)}{\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}}, \\
\mathrm{d}\left\langle \alpha\right\rangle_t &= \frac{\lambda^{(t)\top}(t)\Sigma\Sigma^\top \lambda^{(t)}(t)}{\left\|\Sigma^\top \lambda^{(t)}\right\|_{\mathcal{L}^2}}.
\end{aligned}$$

7.7. Error analysis in p

In this section, we will show that our method can be interpreted as using the "wrong" type of stochastic integral for $\langle \mathbf{w}, \mu \rangle_{\mathcal{L}^2} = \int_0^T w\mu \, \mathrm{d}t$. Since $p = \mathbb{P}\left(\langle \mathbf{w}, \mu \rangle_{\mathcal{L}^2} > \alpha\right)$, we can regard the computed quantile as the correct value for a *different* probability.

This result is based on a Theorem in [IW89] which states that there are smooth approximations of Brownian motions such that integrals of the approximations converge to their Stratonovich counterparts for the approximated Brownian motion.

Recall that the *Itô* integral is defined as a limit of random variables by

$$\int_0^T f(t) \, \mathrm{d}W_t := \lim_{N\to\infty} \frac{T}{N} \sum_{i=0}^{N-1} f\left(\frac{i}{N}\right) \left(W_{(i+1)/N} - W_{i/N}\right),$$

and the *Stratonovich* integral by

$$\int_0^T f(t) \circ \mathrm{d}W_t := \lim_{N\to\infty} \frac{T}{N} \sum_{i=0}^{N-1} \frac{1}{2}\left(f\left(\frac{i}{N}\right) + f\left(\frac{i+1}{N}\right)\right) \left(W_{(i+1)/N} - W_{i/N}\right).$$

Note also that we have seen above that the extended state $(x(t), \alpha(t))$ contains sufficient information to determine the value of the game and consequently we restrict control strategies to be functions of (x, α, t). The game (7.8) defined below differs from (7.1) in that it uses this notion of control strategies, allows W to be in a larger set and uses the Stratonovich integral in g_2^S as g_2 is undefined for non-differentiable W.

Theorem 7.12. *Let V be the value of the differential game (7.1). There exists a set $\mathcal{W}$ of Brownian paths such that $\mathbb{P}(W \in \mathcal{W} = 1)$ and that the value $\tilde{V}$ of the differential game*

$$\begin{aligned} \min_{U,M} \left(\max_{W\in\mathcal{W},\mathbf{x}} L(\mathbf{x}, \mathbf{u}) \right) \qquad & (7.8) \\ s.t.\ \ & g(\mathbf{x} - \Sigma W, \mathbf{u}) \\ & := \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{x} - \Sigma W) - f(x, u, \cdot) + (x_0 - x(0))\delta_0 \equiv 0, \\ & g_2^S(W, \mu) := \alpha(0) - \int_0^T \mu(t) \circ \mathrm{d}W_t \geq 0, \\ & g_3(\mu) := 1 - \|\mu\|_{\mathcal{L}^2}^2 = 0 \\ & u(t) = U(x(t), \alpha(t), t), \ \ \mu = M(x(t), \alpha(t), t) \ \forall t \in [0, T], \end{aligned}$$

where the minimum is over functions $U \in \mathcal{C}^1(\mathcal{X} \times \mathbb{R} \times [0,T]; \mathcal{U})$ and $M \in \mathcal{C}^2(\mathcal{X} \times \mathbb{R} \times [0,T]; \mathbb{R}^n)$, and α is defined on $(0,T]$ (analogously to $\widetilde{\alpha}/\widetilde{\gamma}$) by

$$\alpha(t) := \frac{\alpha(0) - \int_0^T \mu(t) \circ \mathrm{d}W_t}{\sqrt{1 - \int_0^T \|\mu(t)\|^2 \, \mathrm{d}t}},$$

fulfills $V = \tilde{V}$.

Proof. Let U^*, M^* be the solution of (7.1). By the Corollary to Theorem 7.3 in chapter VI of [IW89], there exist almost surely $W^i, \mathbf{u}^i, \mathbf{x}^i, \mu^i, i \in \mathbb{N}$ with $g(\mathbf{x}^i, \mathbf{u}^i, W^i) \equiv 0$, $W^i, \mu^i \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ and $W^i \to W$, $\mathbf{x}^i \to \mathbf{x}$, $\mathbf{u}^i \to U^*(W)$, $\mu^i \to M^*(W)$ uniformly, as well as $\int_0^T \mu^i \cdot \dot{W}^i \,\mathrm{d}t \to \int_0^T M^*(W) \circ \mathrm{d}W_t$, $L(\mathbf{x}^i, \mathbf{u}^i) \to L(\mathbf{x}(W), U(W))$ $(i \to \infty)$. Let $\mathcal{W}$, $\mathbb{P}(W \in \mathcal{W} = 1)$ be a set for which the above holds and which includes all weakly differentiable W. The latter is possible because for these W the regular integral exists and coincides with the Stratonovich integral.

Clearly $\tilde{V} \geq V$ as in (7.8) the maximum is over a larger and the minimum over a smaller set than in (7.1)

For the converse inequality, note that u^* and μ^* computed in Section 7.2 vary smoothly with (x, α, t) and so (U^*, M^*) is admissible for (7.8). If $\widehat{V} > V$, there is an $W \in \mathcal{W}$ such that $L(\mathbf{x}(U^*(W), W), U^*(W)) > V$ and W is admissible in (7.8). Its approximations W^i from above would be admissible in (7.1), leading to a contradiction. □

The game does not provide a p-quantile because generally

$$\mathbb{P}\left(\int_0^T \mu(t) \circ \mathrm{d}W_t \geq \alpha\right) \neq p$$

and the desired equality

$$\mathbb{P}\left(\int_0^T \mu(t) \,\mathrm{d}W_t \geq \alpha\right) = p$$

holds only for the Itô integral.

So V is actually a $(1 - \tilde{p})$-quantile, where we can estimate

$$\tilde{p} \leq \underbrace{\mathbb{P}\left(\int \mu(t) \,\mathrm{d}W_t \geq \alpha\right)}_{=p} + \mathbb{P}\left(\int \mu(t) \,\mathrm{d}W_t < \alpha \wedge \int \mu(t) \circ \mathrm{d}W_t \geq \alpha \wedge L > V\right)$$

and see that the difference is caused by outcomes which are admissible for the Itô, but not the Stratonovich version of the constraint.

For the difference between the Stratonovich and Itô integral we have

$$\int_0^T \mu(t) \,\mathrm{d}W_t = \int_0^T \mu(t) \circ \mathrm{d}W_t - \frac{1}{2} \int_0^T \frac{\mathrm{d}}{\mathrm{d}W_t} \mu(t) : \mathrm{d}\left\langle W_t \right\rangle .$$

Further statements are hard to make in general, but the difference can be estimated e.g. by using the values of $\frac{\mathrm{d}}{\mathrm{d}W_t}\mu$ along the reference path.

7.8. Implementation

The term $\left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}$ appearing as a denominator in g causes some inconvenience when solving (7.3). Firstly, it depends on the entire adjoint λ and therefore prevents the

derivative of the constraint g, that would otherwise be local in time, from being sparse. Secondly, it happens to be zero for the stationary solution $(\mathbf{x}, \mathbf{u}) \equiv (x^*, u^*)$, which we often like to use as a starting point for homotopies. To overcome these problems, we introduce the additional variable $\beta :\approx \left\|\Sigma^\top \lambda\right\|$ and a regularization parameter β_{reg} and solve a discretization (cf. Section A.1.6) of

$$F(\mathbf{x}, \mathbf{u}, \lambda, \beta) := \begin{pmatrix} c_x(\mathbf{x}, \mathbf{u}) - \lambda \cdot g_x(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ c_u(\mathbf{x}, \mathbf{u}) - \lambda \cdot g_u(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ g(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ \beta^2 - \left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}^2 - \beta_{\text{reg}} \end{pmatrix} \overset{!}{\equiv} 0 \tag{7.9}$$

with

$$g(\mathbf{x}, \mathbf{u}, \lambda, \beta) := f(x, u, \cdot) + \alpha \frac{\Sigma\Sigma^\top \lambda}{\beta} - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} + (x_0 - x(0))\delta_0.$$

In general, λ does not vanish and we can set $\beta_{\text{reg}} = 0$ to recover $\beta = \left\|\Sigma^\top \lambda\right\|_{\mathcal{L}^2}$.

The algorithm

At an abstract level, we have the following algorithm:

1. Determine the reference solution for a given initial condition $x(0) = x_0$ and with initial $\alpha(0) = \Phi^{-1}(1-p)$
2. for each t_i
   - 2.1. Compute the solution on $[t_i, T]$ starting at current position $x(t_i)$
   - 2.2. Use $u|_{[t_i, t_{i+1})}$ as control
   - 2.3. Update α

Note that in step 2.1. one can reduce the computational effort by computing the solution approximately by using a linearization around a previous solution and computing a full solution only every couple of time steps.

Step 2.3. varies according to which of the approaches for time consistency is used. Note that using the noiseless control near the target is equivalent to setting $\alpha = 0$ and so the switching strategy is included in this step. The possible choices are:

(i) *Probability maximization for a cost target*: The derivative $\frac{\mathrm{d}L}{\mathrm{d}\alpha} = \lambda_2$ is available as a Lagrange multiplier and so an α fulfilling (7.7) can be computed by Newton's method. We set $\alpha = 0$ if the target has become unachievable.

(ii) α *from reference trajectory*: α is varied according to (7.6). As α goes to zero with this approach, no further switching strategy is needed.

(iii) *Constant* α: α remains unchanged until a neighborhood of the target is reached. α is then set to zero.

7.9. Numerical example

Discretization of a viscous Burgers equation with boundary control

As an example, we use a boundary control problem for the viscous Burgers equation taken from [KVX04] to which we add a noise term which also acts on the boundary:

$$\begin{aligned} y_t - \nu y_{zz} + y y_z &= 0 && \text{in } [0,1] \times (0,T) \\ \nu y_z(\cdot, 0) &= u + \sigma dW && \text{in } (0,T) \\ \nu y_z(\cdot, 1) &= 0 && \text{in } (0,T) \\ y(0,\cdot) &= y_0 && \text{in } [0,1] \end{aligned}$$

$$L(\mathbf{y}, \mathbf{u}) = \frac{1}{2} \int_0^T \int_{[0,1]} |y(t,z)|^2 \,\mathrm{d}z + \eta \,|u(t)|^2 \,\mathrm{d}t$$

We approximate $y(\cdot, t) \approx \sum_{j=1}^{d} x_j(t) \Phi_j(\cdot)$ with Φ_j the normalized Legendre polynomials. Galerkin projection gives the weak equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} x_j &= \left\langle \frac{\mathrm{d}}{\mathrm{d}t} y, \Phi_j \right\rangle \\ &= -\left(\int_0^1 \nu \frac{\mathrm{d}}{\mathrm{d}z} y(z) \frac{\mathrm{d}}{\mathrm{d}z} \Phi_j(z) + y(z) \frac{\mathrm{d}}{\mathrm{d}z} y(z) \frac{\mathrm{d}}{\mathrm{d}z} \Phi_j(z) \,\mathrm{d}z + (u(t) + \sigma dW_t)\, \Phi_j(0) \right). \end{aligned}$$

In particular $\Sigma \in \mathbb{R}^{d \times 1}$, $\Sigma_{j,1} = \sigma \Phi_j(0)$.

As the Φ_j are orthonormal, we obtain the cost function

$$L(\mathbf{x}, \mathbf{u}) = \frac{1}{2} \int_0^T \|x(t)\|^2 + \eta \,|u(t)|^2 \,\mathrm{d}t.$$

For the experiment we always use $\nu = 0.05$, $\mu = 0.1$, $T = 1$, $y_0(x) = (1-x)\sin(3\pi(x - 1/2))$ and $d = 20$, whereas the noise intensity σ will vary.

The quantile functions are approximated by a Monte Carlo method. We draw a number of samples $\mathbf{w}_i$ from the distribution of W and compute $L(\mathbf{u}(\mathbf{w}_i), \mathbf{w}_i)$ for different control strategies. (Note that the same $\mathbf{w}_i$ are used for all control strategies to improve comparability.) When studying the following figures one should keep in mind that $Q_L(p) \to \infty$ as $p \to 1$ and hence the rightmost part of the quantile functions is subject to a large sampling error.

Figure 7.1 shows the difference in outcomes compared to the deterministic control for $\sigma = .05$. The time-consistent strategies perform best, in accordance with theoretical predictions. Note also that for $\alpha = 1$ we optimize the quantile for $p \approx 0.84$ and that the naive strategy with constant α performs worse than the deterministic control for this p.

However, we obtain a different picture if we increase the noise to $\sigma = .1$ (Figure 7.2). The time-consistent strategies are still an improvement over the deterministic control but they are clearly not optimal at $p \approx .84$, which suggests that the linearity assumption underlying the model no longer valid. Remarkably, the naive strategy is more robust and

performs well in the high-noise case.

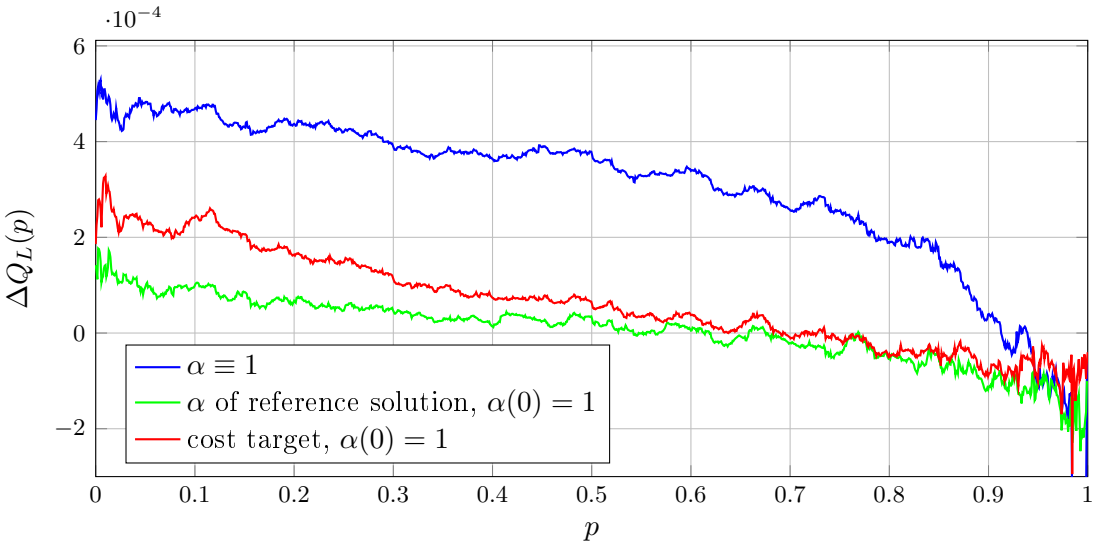

Figure 7.1.: Difference of quantile functions for $\sigma = .05$ compared to deterministic control for different control strategies (Monte Carlo simulation with 1216 samples)

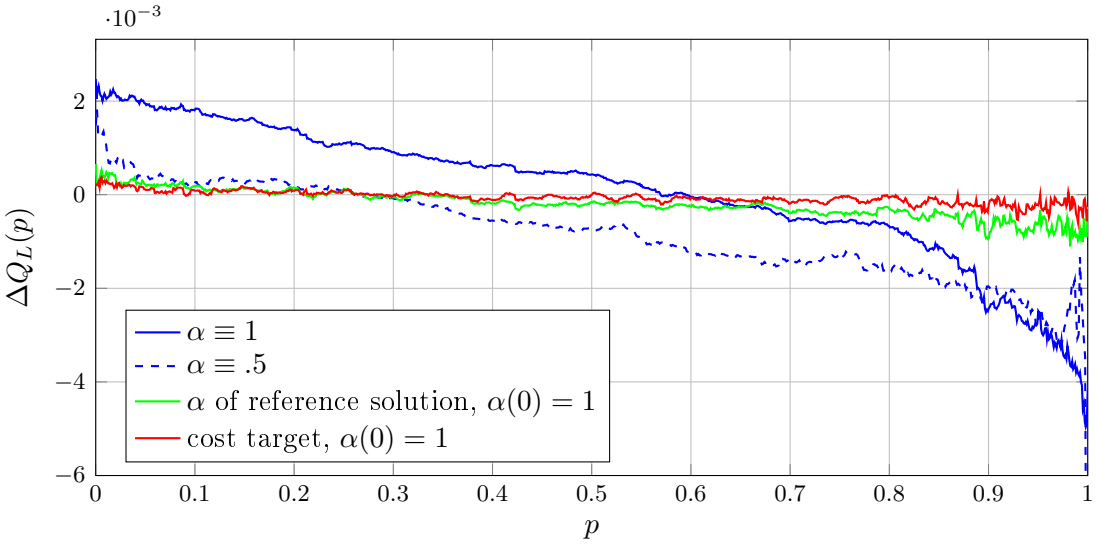

Figure 7.2.: Difference of quantile functions for $\sigma = .1$ compared to deterministic control (Monte Carlo simulation with 1024 samples)

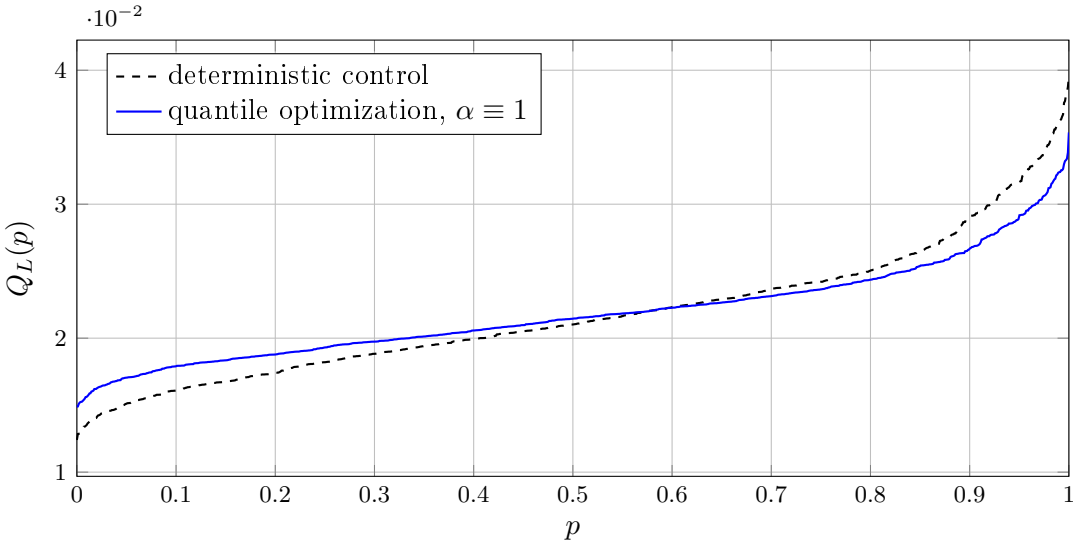

Figure 7.3.: Quantile functions for different control strategies with $\sigma = .1$ (Monte Carlo simulation with 1024 samples)

A. Solution of local subproblems

The subproblems (1.6) arising during the discrete homotopies in Algorithm 5.1 need to be discretized in order to be solved numerically.

As we are dealing with extremals which converge to (x^*, u^*) we can reduce the problem from an infinite to a finite time horizon $[0, T]$ through a cut-off. Due to the exponential decay of the solution near (x^*, u^*) (recall that extremals are in the stable manifold), an adaptive grid allows a high enough T for the solution to become essentially constant at low computational cost.

A.1. Discretization of the optimality conditions

We will describe the structure of the discrete optimality condition and then determine the coefficients appearing in it. W.l.o.g. we assume again $t_0 = 0$.

A.1.1. Structure of the discrete equations

We choose to always represent functions by nodal values instead of coefficients in some basis and hence choose a time grid $t_i \in [0, T]$, $i = 1, \ldots, N$ and have as unknowns the values x_i, u_i and λ_i associated to each t_i. Note that x_0 is not an unknown but the initial state.

The integral in the cost function L is approximated by some quadrature rule

$$L \approx \sum_{i=1}^{N} L_i^{\mathrm{c}} c(x_i, u_i)$$

specified by the coefficients $L_i^{\mathrm{c}} \in \mathbb{R}$, $i = 1, \ldots, N$.

For this chapter, we define $g \equiv f(\mathbf{x}, \mathbf{u}) - \dot{\mathbf{x}} \in \mathcal{W}^{-1,\infty}([0, T]; \mathcal{X})$ and treat the initial condition $x(0) = x_0$ separately. Those constraints are translated into a set of $\mathcal{X}$-valued constraints g_j, $j = 1, \ldots N$ of the form

$$g_j(x_0, \ldots, x_N, u_1, \ldots, u_N) = \sum_{i=1}^{N} G_{ji}^{\mathrm{x}} x_i + \sum_{i=1}^{N} G_{ji}^{\mathrm{f}} f(x_i, u_i) + G_j^{\mathrm{ic}} x_0,$$

where $G^{\mathrm{x}} \in \mathbb{R}^{N \times N}$, $G^{\mathrm{f}} \in \mathbb{R}^{N \times N}$ and $G^{\mathrm{ic}} \in \mathbb{R}^{N}$ depend on the choice of integrator used for $\dot{x} = f(x, u)$.

A.1.2. Discretization by the Discontinuous Galerkin method

We now derive values for the coefficients L^{c}, G^{x}, G^{f} and G^{ic} using a Discontinuous Galerkin (DG) method for the dynamical constraint.

The interval $[0,T]$ is divided in $\widehat{N}$ subintervals $[\hat{t}_{i-1},\hat{t}_i]$, $i=1,\ldots,\widehat{N}$, $0=\hat{t}_0<\hat{t}_1<\ldots<\hat{t}_{\widehat{N}}=T$. The approximation is sought in the space of discontinuous piecewise polynomials of degree k,

$$\Pi := \mathbb{P}_k((\hat{t}_0,\hat{t}_1))\times\ldots\times\mathbb{P}_k((\hat{t}_{\widehat{N}-1},\hat{t}_{\widehat{N}})).$$

Hence the DG method allows u to be discontinuous, which can occur if c is discontinuous on t. This will not happen in the infinite horizon problem (2.17), but is possible for the quantile optimization problem.

Treatment of discontinuities

The discontinuities leaves undefined the values at subinterval boundaries $\hat{t}_i$, which are critical for the evaluation of the integrals that will appear in the weak form of the ODE. We now define those to be some combination of the left and right limits, i.e.

$$p(\hat{t}_i) := \alpha p(\hat{t}_i^-) + (1-\alpha)p(\hat{t}_i^+). \tag{A.1}$$

This definition is equivalent to defining the function values as the limit of suitable mollifications of the function. Let $\eta\in\mathcal{C}_0^\infty(\mathbb{R})$ with $\eta\geq 0$, $\int_{-\infty}^{\infty}\eta(t)\,\mathrm{d}t=1$ and — in addition to the usual requirements — $\int_{-\infty}^{0}\eta(t)\,\mathrm{d}t=\alpha\in[0,1]$ be a mollifier and define $\eta_\epsilon(t)=\frac{1}{\epsilon}\eta\left(\frac{1}{\epsilon}t\right)$. For $p\in\Pi$ we have $\lim_{\epsilon\to0}(p*\eta_\epsilon)(t)=\alpha p(t^-)+(1-\alpha)p(t^+)$,[1] i.e. the value of $p(t)$ as defined in (A.1). Furthermore, for any $\varphi\in\Pi$,[2] at any t where $p\in\Pi$ is discontinuous:

$$\begin{aligned}
&\lim_{h\to0}\lim_{\epsilon\to0}\int_t^{t+h}(p*\eta_\epsilon)'(s)\varphi(s)\,\mathrm{d}s\\
&=\lim_{h\to0}\left(\varphi(t^+)\lim_{\epsilon\to0}\int_t^{t+h}(p*\eta_\epsilon)'(s)\,\mathrm{d}s\right)\\
&=\lim_{h\to0}\Big(\varphi(t^+)\lim_{\epsilon\to0}(p*\eta_\epsilon)(t+h)-(p*\eta_\epsilon)(t)\Big)\\
&=\varphi(t^+)\left(p(t^+)-\left(\alpha p(t^-)+(1-\alpha)p(t^+)\right)\right)\\
&=\alpha(p(t^+)-p(t^-))\varphi(t^+)
\end{aligned}\tag{A.2a}$$

and similarly

$$\lim_{h\to0}\lim_{\epsilon\to0}\int_{t-h}^{t}(p*\eta_\epsilon)'(s)\varphi(s)\,\mathrm{d}s=(1-\alpha)(p(t^+)-p(t^-))\varphi(t^-). \tag{A.2b}$$

[1] For the purpose of convolutions, we assume functions to be zero outside their domains.

[2] More generally, for any φ which is continuous on $(t,t+\delta]$ resp. $[t-\delta,t)$ for some $\delta>0$ and has a limit for $s\searrow t$ resp. $s\nearrow t$.

Essentially we are assigning a fraction α of the jump $p(t^+) - p(t^-)$ at t to t^+ and the remaining fraction $1-\alpha$ to t^-.

Weak formulation

With (A.2) we can now derive a weak formulation of the ODE. It is convenient to choose first a basis $\left\{\phi^{(0)}, \ldots, \phi^{(k)}\right\}$ of $\mathbb{P}_k((0,1))$ and then use $(\varphi^{(i,j)})_{i=1,\ldots,\widehat{N},j=0,\ldots,k}$,

$$\varphi^{(i,j)}(t) = \begin{cases} \phi^{(j)}\left(\frac{t-\hat{t}_{i-1}}{\hat{t}_i - \hat{t}_{i-1}}\right) & t \in (\hat{t}_{i-1}, \hat{t}_i) \\ 0 & \text{else} \end{cases}$$

as a basis for Π.

Projection of the dynamical constraint $f(\mathbf{x}, \mathbf{u}) \equiv 0 - \dot{\mathbf{x}}$ then yields conditions of the form

$$0 \stackrel{!}{=} \langle f_l(\mathbf{x}, \mathbf{u}) - \dot{\mathbf{x}}_l, \varphi \rangle = \int_a^b \left(f(x(t), u(t)) - \dot{x}(t)\right) \varphi(t) \, \mathrm{d}t$$

with $a = \hat{t}_{i-1}$, $b = \hat{t}_i$, $\varphi = \varphi^{(i,j)}$.

The critical term is $\int_a^b \dot{x}(t)\varphi(t)\,\mathrm{d}t$, which does not exist in the classical sense due to the possible discontinuities of x at a and b, so we define instead the new dual product

$$\langle \mathbf{p}, \varphi \rangle_\Pi := \lim_{\epsilon \to 0} \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}(\eta_\epsilon * p)(t)\varphi(t)\,\mathrm{d}t$$

and using (A.2) obtain

$$\begin{aligned}
\langle \dot{\mathbf{x}}, \varphi \rangle_\Pi &= \lim_{h \to 0} \int_a^{a+h} \dot{x}(t)\varphi(t)\,\mathrm{d}t + \int_{a+h}^{b-h} \dot{x}(t)\varphi(t)\,\mathrm{d}t + \int_{b-h}^{b} \dot{x}(t)\varphi(t)\,\mathrm{d}t \\
&= \alpha(x(a^+) - x(a^-))\varphi(a^+) - \int_a^b x(t)\dot{\varphi}(t)\,\mathrm{d}t + x\varphi|_{a^+}^{b^-} + (1-\alpha)(x(b^+) - x(b^-))\varphi(b^-) \\
&= -\left(\alpha x(a^-) + (1-\alpha)x(a^+)\right)\varphi(a^+) + \left(\alpha x(b^-) + (1-\alpha)x(b^+)\right)\varphi(b^-) - \int_a^b x(t)\dot{\varphi}(t)\,\mathrm{d}t \\
&= -x(a)\varphi(a^+) + x(b)\varphi(b^-) - \int_a^b x(t)\dot{\varphi}(t)\,\mathrm{d}t, \qquad \text{(A.3)}
\end{aligned}$$

where we have used (A.1) to arrive at the last line. Note that if φ is continuous in a and b, this is equal to the result one would get by formally applying partial integration.

Finally , we remark that the dual product $\langle \cdot, \cdot \rangle_\Pi$ depends on α but not otherwise on η. For the forward integration of an ODE it is advantageous to choose $p(\hat{t}_i) := p(\hat{t}_i^-)$, i.e. $\alpha = 1$ so that values in $[\hat{t}_i, \hat{t}_{i+1}]$ depend only on the previous subinterval through $p(\hat{t}_i)^-$ but not on the next one, i.e. the future. This allows one to solve the ODE advancing from one subinterval to the next instead of simultaneously on the entire interval.

Error estimates

A priori The DG method has been shown to be of order k in Theorem 3.1 of [Est95]. Restated for our setting this theorem becomes the following

Theorem A.1. *Let* $\mathbf{u}$ *be given,* $\widetilde{\mathbf{x}}$ *be the true solution of* $\dot{x}(t) = f(x(t), u(t))$, $t \in [0, T]$ *and* $\mathbf{x}$ *its DG approximant. Let* $\Omega \subseteq \mathcal{X}$ *be a compact set such that* $\widetilde{\mathbf{x}}$ *and* $\mathbf{x}$ *stay in* Ω *and* $(x, t) \mapsto f(x, u(t))$ *has continuous partial derivatives of order* k *on* Ω*. Then there is a constant* C *such that for all* $i = 1, \ldots, \widehat{N}$*,*

$$\sup_{t \in [0, \hat{t}_i]} \|\widehat{x}(t) - x(t)\| \leq C \max_{j \leq i} (\hat{t}_i - \hat{t}_{i-1})^{k+1} \sup_{t \in [\hat{t}_{i-1}, \hat{t}_i]} \left\| \widehat{x}^{(k+1)}(t) \right\|.$$

A posteriori The approximation space Π always includes the step functions $\chi_{(\hat{t}_{i-1}, \hat{t}_i)}$, $i = 1, \ldots, \widehat{N}$, so the discrete solution $\mathbf{x}$ fulfills

$$\begin{aligned} 0 &= \left\langle f(\mathbf{x}, \mathbf{u}) - \dot{\mathbf{x}}, \chi_{(\hat{t}_{i-1}, \hat{t}_i)} \right\rangle_{\Pi} \\ \Rightarrow \underbrace{\int_{\hat{t}_{i-1}}^{\hat{t}_i} f(x(t), u(t)) - \dot{x}(t) \, \mathrm{d}t}_{:= Res_i} &= \alpha \left(x(\hat{t}_{i-1}^+) - x(\hat{t}_{i-1}^-) \right) - (1 - \alpha) \left(x(\hat{t}_i^+) - x(\hat{t}_i^-) \right). \end{aligned}$$

The local residual Res_i on the ith subinterval can therefore be directly computed from the jumps at $\hat{t}_{i-1}$ and $\hat{t}_i$. (Note that when fully discretizing the problem, one also needs to account for the quadrature error.)

A.1.3. Nodes, quadrature and computation of coefficients

Nodes

To obtain the nodes t_i we choose nodes θ_l, $l = 0, \ldots, k$ on the reference interval $[0, 1]$, which are then transformed to each subinterval:

$$t_j := \hat{t}_i + \theta_l \text{ for } j = i(k+1) + l + 1, \; j = 1, \ldots, N, \; N := (k+1)\widehat{N}.$$

In our implementation we choose Chebyshev nodes

$$\theta_l := \left(1 + \cos\left((k - l) \frac{\pi}{k+1} \right) \right) / 2.$$

Note that for those nodes $\theta_0 = 0$, $\theta_k = 1$ and so we have duplications $t_{i(k+1)+k+1} = t_{(i+1)(k+1)+1} = \hat{t}_{i+1}$. As $t_{i(k+1)+l+1}$ is supposed to belong to the ith subinterval, this is to be interpreted as $t_{i(k+1)+k+1} = \hat{t}_{i+1}^-$, $t_{(i+1)(k+1)+1} = \hat{t}_{i+1}^+$.

Quadrature

In order to discretize the integral in the cost functional a quadrature rule

$$\int_0^1 f(t)\,\mathrm{d}t \approx \sum_{l=0}^{k} w_l f(\theta_l) \tag{A.4}$$

on $[0,1]$ is required. The composite quadrature rule on $[0,T]$ then reads

$$\int_0^T e^{-\mu t} c(x(t),u(t))\,\mathrm{d}t \approx \sum_{j=0}^{N} L_j^{\mathrm{c}} e^{-\mu t_j} c(x_j,u_j)$$

with

$$L_j^{\mathrm{c}} := w_j(\hat{t}_{i+1}-\hat{t}_i) \text{ for } j = i(k+1)+l+1,\ j=1,\dots,N.$$

On Chebyshev nodes the optimal order of $k+1$ is achieved by Clenshaw-Curtis quadrature. Although this is lower than the order $2k+2$ that could be achieved with Gauss quadrature on a different set of nodes, Clenshaw-Curtis quadrature often performs almost equally well ([Tre08]). The weights are given by

$$w_l = \frac{c_l}{2k}\left(1 - \sum_{j=1}^{\lfloor k/2\rfloor} \frac{b_j}{4j^2-1}\cos\left(2j(k-l)\frac{\pi}{k+1}\right)\right),\ l=0,\dots,k$$

with

$$b_j = \begin{cases} 1, & j = k/2 \\ 2, & j < k/2 \end{cases}, \qquad c_l = \begin{cases} 1, & l = 0 \text{ or } l = k \\ 2, & \text{otherwise} \end{cases}.$$

An efficient computation of the weights is possible by using the DFT, see e.g. [Wal06].

Polynomial basis

While the DG method works for any basis, our choice to use the values $x_j = x(t_j)$, $u_j = u(t_j)$ as the unknown variables implies the use of the Lagrange basis

$$\phi^{(j)}(\theta) = \prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{\theta-\theta_i}{\theta_j-\theta_i}$$

and we have

$$\mathbf{x} = \sum_{i=1}^{N} x_i\varphi_i,\ \mathbf{u} = \sum_{i=1}^{N} u_i\varphi_i.$$

Computation of coefficients

We consider only the case $\alpha = 1$, i.e. $x(t) = x(t^-)$. Before quadrature, the jth constraint ($j = i(k+1) + l + 1$ as above) reads

$$\begin{aligned} g_j &= \langle f(\mathbf{x}, \mathbf{u}) - \dot{\mathbf{x}}, \varphi_j \rangle \\ &\overset{(A.3)}{=} \int_{\hat{t}_i}^{\hat{t}_{i+1}} f(x(t), u(t))\varphi(t)\,\mathrm{d}t + \int_{\hat{t}_i}^{\hat{t}_{i+1}} x(t)\dot{\varphi}(t)\,\mathrm{d}t + x(\hat{t}_i)\varphi(\hat{t}_i^+) - x(\hat{t}_{i+1})\varphi(\hat{t}_{i+1}^-) \\ &\overset{!}{=} 0. \end{aligned}$$

Note that the second integral already is a linear function $\sum_i G^{\mathrm{x}}_{ji} x_i$ of the x_i, whereas our choice to approximate the first integral using only the values $f_i = f(x_i, u_i)$ by $\sum_i G^{\mathrm{f}}_{ji} f_i$ is a significant restriction.

From the above equation it is obvious that g_j depends only on values for $t \in [\hat{t}_i^-, \hat{t}_{i+1}^-]$, corresponding to indices $i(k+1), \ldots, (i+1)(k+1)$, which leads to sparse matrices G^{x}, G^{f}. Furthermore, we apply the quadrature rule (A.4) to the integral $\int f\varphi$ and, due to the Lagrange basis, find that G^{f} becomes diagonal, because

$$\int_{\hat{t}_i}^{\hat{t}_{i+1}} f(x(t), u(t))\varphi_j(t)\,\mathrm{d}t \approx w_l(\hat{t}_{i+1} - \hat{t}_i) f_j$$

involves only f_j.

To summarize, we have, for all $j = 1, \ldots, N$

$$\begin{aligned} G^{\mathrm{f}}_{j,j} &= w_l(\hat{t}_{i+1} - \hat{t}_i) \\ G^{\mathrm{x}}_{j,i(k+1)+r+1} &= \textstyle\int_0^1 \phi_r(t)\dot{\phi}_l(t)\,\mathrm{d}t \quad r = 0, \ldots, k \\ G^{\mathrm{x}}_{j,(i-1)(k+1)+k+1} &= \phi_l(1) = \delta_{lk} \qquad \text{if } i > 1, \end{aligned}$$

with all other entries being 0. If $i = 1$, then the index $(i-1)(k+1) + k + 1 = 0$ in the last line corresponds to the initial condition and hence

$$G^{\mathrm{ic}}_j = \delta_{lk}.$$

A.1.4. Optimization

Discretize-then-optimize

A discrete version of the optimality conditions can now be obtained by solving the discrete problem $\tilde{L} = \max!$ s.t. $g_j = 0\ \forall j$, an approach known as "discretize-then-optimize".

Using Lagrange multipliers and $\frac{\partial g_j}{x_i} = G^{\mathrm{x}}_{ji} + G^{\mathrm{f}}_{ji}\frac{\partial f}{\partial x}(x_i, u_i)$, $\frac{\partial g_j}{u_i} = G^{\mathrm{f}}_{ji}\frac{\partial f}{\partial u}(x_i, u_i)$, the optimality conditions then turn out to be

$$L^{\mathrm{c}}_i \frac{\partial c}{\partial x}(x_i, u_i) + \sum_{j=1}^{N} \lambda_j \cdot \left(G^{\mathrm{x}}_{ji} + G^{\mathrm{f}}_{ji}\frac{\partial f}{\partial x}(x_i, u_i) \right) = 0, \tag{A.5a}$$

$$L_i^{\mathrm{c}} \frac{\partial c}{\partial u}(x_i, u_i) + \sum_{j=1}^{N} \lambda_j \cdot G_{ji}^{\mathrm{f}} \frac{\partial f}{\partial u}(x_i, u_i) = 0 \ \forall i \tag{A.5b}$$

Optimize-then-discretize

Alternatively, one could obtain discrete optimality conditions by discretizing the optimality conditions of the original, continuous problem ("optimize-then-discretize").

We will show that this approach yields the same conditions, if the ODE $-\dot{\lambda} = e^{-\mu t} c_x + f_x^\top \lambda$ is integrated with the DG method with $\alpha = 0$. This corresponds to the fact that λ is a function in the dual space and hence is treated by the dual product $\langle \cdot, \cdot \rangle_\Pi$ as if it is left-continuous. It also reflects that the adjoint equation is an ODE that is integrated backwards from a terminal condition.

The dependence of the dual product on α will now be made explicit by writing $\langle \cdot, \cdot \rangle_{\Pi,\alpha}$. From $\alpha = \int_{-\infty}^{0} \eta(t)\,\mathrm{d}t$ one sees that replacing α by $(1-\alpha)$ corresponds to replacing η by $\eta(-\cdot)$, which in turn corresponds to transposing the convolution operator. Hence

$$\langle \psi, \varphi \rangle_{\Pi,\alpha} = \lim_{\epsilon \to 0} \int (\eta_\epsilon * \psi)\varphi = \lim_{\epsilon \to 0} \int \psi (\eta_\epsilon * \varphi(-\cdot)) = \langle \varphi, \psi \rangle_{\Pi,1-\alpha}\,.$$

$$\left\langle e^{-\mu \cdot} c_x + f_x^\top \lambda + \dot{\lambda}, \varphi_i \right\rangle_{\Pi,0} = 0 \ \forall i$$

$$\left\langle e^{-\mu \cdot} c_x, \varphi_i \right\rangle_{\Pi,0} + \langle \lambda, f_x \varphi_i \rangle_{\Pi,0} - \langle \lambda, \dot{\varphi}_i \rangle_{\Pi,0} = 0 \ \forall i$$

$$\left\langle \varphi_i, e^{-\mu \cdot} c_x \right\rangle_{\Pi,1} + \langle f_x \varphi_i, \lambda \rangle_{\Pi,1} - \langle \dot{\varphi}_i, \lambda \rangle_{\Pi,1} = 0 \ \forall i$$

With $\lambda = \sum_j \lambda_j \varphi_j$ this becomes

$$\left\langle \varphi_i, e^{-\mu \cdot} c_x \right\rangle_{\Pi,1} + \sum_j \lambda_j \left\langle f_x \varphi_i, \varphi_j \right\rangle_{\Pi,1} + \lambda_j \sum_j \left\langle \varphi_i, \dot{\varphi}_j \right\rangle_{\Pi,1} = 0 \ \forall i$$

and turns out to be (A.5a), if the same quadrature rules are used.

The second discrete optimality condition, (A.5b), is just the pointwise condition $e^{-\mu t} c_u(t_j) + f_u(t_j)^\top \lambda(t_j) = 0$ for all $j = 1, \ldots, N$ (recall that G^{f} is diagonal).

A.1.5. Assembly of F and DF

The derivatives of the discrete cost and constraint are

$$\begin{aligned} \nabla_x L := \nabla_{x_1,\ldots,x_N} L &= \begin{pmatrix} L^{\mathrm{c}} e^{-\mu t_1} c_x^\top(x_1, u_1) \\ \vdots \\ L^{\mathrm{c}} e^{-\mu t_N} c_x^\top(x_N, u_N) \end{pmatrix} \\ &= L^{\mathrm{c}} \circ e^{-\mu \vec{t}} \circ (c_x^\top(x_i, u_i))_{i=1,\ldots,N} \\ &=: L^{\mathrm{c}} \circ e^{-\mu \vec{t}} \circ c_x^\top, \end{aligned}$$

where $\vec{t} = (t_1, \dots, t_N)^\top$ and $\circ$ denotes, depending on context, both the regular elementwise (Hadamard) product and the "elementwise" product in which the elements of the second factor are the blocks of a block matrix.

Similarly, with $g := (g_1^\top, \dots, g_N^\top)^\top$,

$$\begin{aligned}
\nabla_u L &= L^{\mathrm{c}} \circ e^{-\mu \vec{t}} \circ c_u^\top, \\
\nabla_x g &= G^{\mathrm{x}} \otimes I_d + G^{\mathrm{f}} \circ f_x^\top, \\
\nabla_u g &= G^{\mathrm{f}} \circ f_u^\top,
\end{aligned}$$

where $\otimes$ denotes the Kronecker product.

The discrete version of (1.6) is

$$F := \begin{pmatrix} \nabla_x L + \nabla_x g \lambda \\ \nabla_u L + \nabla_u g \lambda \\ g \end{pmatrix} \stackrel{!}{=} 0$$

and for its derivative we have

$$D_{(x,u,\lambda)} F = \begin{pmatrix} A_{xx} + B_{xx} & A_{xu} & \nabla_x g \\ A_{ux} + B_{ux} & A_{uu} & \nabla_u g \\ D_x g & D_u g & 0 \end{pmatrix}, \tag{A.6}$$

with the block matrices $(i, j = 1, \dots, N)$

$$A_{xx} := \begin{pmatrix} \ddots & & \\ & L_j^{\mathrm{c}} e^{-\mu t_i} \left(\frac{\partial^2}{\partial x^2} c(x_j, u_j) \right)^\top & \\ & & \ddots \end{pmatrix},$$

$$B_{xx} := \begin{pmatrix} & \vdots & \\ \cdots & G_{i,j}^{\mathrm{f}} \left(\frac{\partial^2}{\partial x^2} \lambda_j^\top f(x_j, u_j) \right)^\top & \cdots \\ & \vdots & \end{pmatrix}$$

and A_{ux}, B_{ux} etc. similarly.

A.1.6. Changes for quantile optimization

With some modifications, this discretization can also be used for the Quantile Optimization problem. We describe here the necessary changes for the formulation

$$F(\mathbf{x}, \mathbf{u}, \lambda, \beta) := \begin{pmatrix} c_x(\mathbf{x}, \mathbf{u}) - \lambda \cdot g_x(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ c_u(\mathbf{x}, \mathbf{u}) - \lambda \cdot g_u(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ g(\mathbf{x}, \mathbf{u}, \lambda, \beta) \\ \beta^2 - \left\| \Sigma^\top \lambda \right\|^2 - \beta_{\mathrm{reg}} \end{pmatrix} \stackrel{!}{\equiv} 0 \tag{7.9}$$

with

$$g(\mathbf{x}, \mathbf{u}, \lambda, \beta) := f(x, u, \cdot) + \alpha \frac{\Sigma\Sigma^\top \lambda}{\beta} - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}$$

from section 7.8.

The discretizations for the first two components carry over unchanged. In the third component, the new term $\alpha\frac{\Sigma\Sigma^\top\lambda}{\beta}$ can be treated as an addition to f, giving

$$g_j = \sum_{i=1}^{N} G^{\mathrm{x}}_{ji} x_i + \sum_{i=1}^{N} G^{\mathrm{f}}_{ji} \left(f(x_i, u_i) + \alpha \frac{\Sigma\Sigma^\top \lambda_i}{\beta} \right) + G^{\mathrm{ic}}_j x_0.$$

Applying the quadrature rule to the fourth component, we obtain

$$\beta^2 - \sum_{j=1}^{N} w_l (\hat{t}_{i+1} - \hat{t}_i) \left\| \Sigma^\top \lambda_j \right\|^2 - \beta_{\mathrm{reg}} \tag{A.7}$$

with $j = i(k+1) + l + 1$.

For the derivative, we now have

$$D_{(x,u,\lambda,\beta)} F = \begin{pmatrix} A_{xx} + B_{xx} & A_{xu} & \nabla_x g & 0 \\ A_{ux} + B_{ux} & A_{uu} & \nabla_u g & 0 \\ D_x g & D_u g & D_\lambda g & D_\beta g \\ 0 & 0 & D_\lambda \left(- \left\| \Sigma^\top \lambda \right\|^2 \right) & 2\beta \end{pmatrix}.$$

Terms already appearing in (A.6) remain unchanged. From (A.7) we see that

$$\frac{\partial}{\partial \lambda_i} g_j = G^{\mathrm{f}}_{ji} \alpha \frac{\Sigma\Sigma^\top}{\beta},$$
$$\frac{\partial}{\partial \beta} g_j = -\alpha \frac{\Sigma\Sigma^\top}{\beta^2} \sum_{i=1}^{N} G^{\mathrm{f}}_{ji} \lambda_i,$$

and finally compute

$$\frac{\partial}{\partial \lambda_j} \left(- \left\| \Sigma^\top \lambda \right\|^2 \right) = -2 w_l (\hat{t}_{i+1} - \hat{t}_i) \left\| \Sigma^\top \lambda_j \right\| \Sigma \lambda_j^\top.$$

A.2. Discrete homotopy

We perform discrete homotopies using Newton's method as described in Section 5.4 and have attempted two different strategies to choose the step size, i.e. the sequence η_j.

The first strategy is to attempt the homotopy in a single step (i.e. $\eta_1 = 0$, $\eta_2 = 1$) and thereby push any difficulties (in particular, this includes determining whether the homotopy can succeed at all) to the solver, which means that we use the globalized

version

$$x^{(k+1)} := x^{(k)} - \nu^{(k)} DF\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right).$$

of Newton's method, which includes a step size $0 < \nu^{(k)} \leq 1$, whereas the standard local version always uses $\nu^{(k)} = 1$. Appendix B describes how to choose the $\nu^{(k)}$ such that it can be easily determined whether a solution exists.

The second strategy is cruder but proved to be more efficient in practice. We attempt a step for some initial step size $\Delta\eta_j = \eta_{j+1} - \eta_j$ and try to find the associated extremal using the local version of Newton's method. If the Newton iterations fail the convergence test $\left\|\Delta x^{(k+1)}\right\| < \rho \left\|\Delta x^{(k)}\right\|$, with $\rho \in (0,1)$ a parameter and $\Delta x^{(k)} := DF\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right)$, we abort the Newton iteration and retry with a smaller $\delta\eta_j$. This is repeated until we succeed with $\eta_j = 1$ or fail with $\Delta\eta_j$ below a given threshold.

In both strategies, we need to solve the linear equation

$$\Delta x^{(k)} := DF\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right)$$

(where F and DF refer to the discrete versions). There are specialized algorithms, e.g. [BP84] to exploit the structure of DF. However, multi-purpose solvers, like MATLAB's *backslash*-operator, already perform sufficiently well that the time required to solve the linear system is negligible next to the time required to assemble the matrix.

B. Step size control for Newton's method in the presence of singularities

Nonlinear equations of the form $F(x) = 0$, $F : \mathbb{R}^d \to \mathbb{R}^d$ are often solved numerically using Newton's method

$$x^{(k+1)} := x^{(k)} - DF\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right).$$

Although Newton's method converges locally to roots of F, it may fail to do so — and even show chaotic behavior — globally. This may be overcome by the use of the so-called globalized or damped Newton method, which uses a step size[1] $0 < \lambda^{(k)} \leq 1$ and reads

$$x^{(k+1)} := x^{(k)} - \lambda^{(k)} DF\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right).$$

It can be understood as an integrator tracing the Newton path

$$\dot{x} = -DF^{-1}F \tag{B.1}$$

with a suitable step size being necessary to ensure stability. As (B.1) implies $\dot{F} = DF \cdot \dot{x} = -F$, the globalized Newton method can be expected to converge to a root of F for small enough step sizes. The choice of suitable and efficient step sizes has been studied extensively, see e.g. [Deu04] for an overview.

However, the above reasoning breaks down and even the globalized Newton method may fail to find a root of F, if the Jacobian matrix DF becomes singular. The simplest example for this situation is $F : \mathbb{R}, \mathbb{R}, x \mapsto x^2 + 1$, where at $x = 0$, $F(x) = 1$ the function "turns around" with $DF(x) = 0$ and hence it is impossible to get any closer to $F = 0$.

In practice, this may result in a Newton algorithm becoming stuck or eventually aborting due to the step size becoming too small. We are therefore interested in methods which can efficiently determine whether the Newton path terminates at a singular Jacobian.

This paper shows that the affine covariant stepsize control introduced by Deuflhard ([Deu74]) and a modification based on Projected Natural Level Functions proposed by Steinhoff ([Ste11]) both possess this ability. To this end we will present a singularity indicator based on [GR84] (cf. Appendix B.7 for the connection) and derive from it stepsize controls that turn out to coincide with those mentioned above. These schemes are then analyzed and we conclude with some numerical experiments.

[1] The letter λ is reused here. The step size has no relation to the adjoint.

B.1. The singularity indicator

B.1.1. Definition

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be sufficiently smooth. We define for $v \neq 0$

$$g(x,v) := \begin{cases} \left\| DF(x)^{-1} \frac{F(x)}{\|F(x)\|} \right\|^{-1} & \text{if } DF(x) \text{ is invertible and } F(x) \neq 0, \\ 0 & \text{if } F(x) \notin \mathcal{R}(DF(x)), \\ \lim\limits_{\epsilon \searrow 0} g(x + \epsilon \frac{v}{\|v\|}) & \text{otherwise.} \end{cases} \tag{B.2}$$

Note that g is not properly defined everywhere as the limit may fail to exist. We will see that this does not matter for our purposes. In most cases g will not depend on v and we will speak only of $g(x)$.

The motivation of this definition is that, informally, by division by zero, $\left\| DF(x)^{-1} \frac{F(x)}{\|F(x)\|} \right\|$ should be infinite and g zero if $DF(x)$ is singular.

Note. *The indicator g is a special case of the indicator proposed by Griewank and Reddien in [GR84]. However, their indicator is only defined if the rank-deficiency is at most one. Details can be found in Section B.7.*

B.2. Exact stepsize control

We begin by deriving a stepsize control from the first case of (B.2) and deal with the other cases later.

For the following calculations it is convenient to introduce some abbreviations. Let $J(x) := DF(x)$, $H(x) := D^2F(x)$, $R(x) := \frac{F(x)}{\|F(x)\|}$ and $T(x) := \frac{DF(x)^{-1}F(x)}{\|DF(x)^{-1}F(x)\|}$.

Using $g = f \circ h$ with $f(x) := \|x\|^{-1}$ and $h(x) := J(x)^{-1}R(x)$ we then compute

$$\begin{aligned} Df(x)(v) &= -\|x\|^{-2} \left\langle \frac{x}{\|x\|}, v \right\rangle = -\|x\|^{-3} \langle x, v \rangle \\ D^2 f(x)(v,u) &= 3\|x\|^{-5} \langle x, v \rangle \langle x, u \rangle - \|x\|^{-3} \langle u, v \rangle \end{aligned}$$

and, suppressing x,

$$\begin{aligned} Dh(v) &= -J^{-1}H(v,\cdot)J^{-1}R + J^{-1}DR(v) \\ D^2h(v,u) &= J^{-1}H(u,\cdot)J^{-1}H(v,\cdot)J^{-1}R - J^{-1}DH(v,u,\cdot)J^{-1}R \\ &\quad + J^{-1}H(v,\cdot)J^{-1}H(u,\cdot)J^{-1}R - J^{-1}H(u,\cdot)J^{-1}DR(v) \\ &\quad - J^{-1}H(v,\cdot)J^{-1}DR(u) + J^{-1}D^2R(v,u) \qquad \text{(B.3)} \\ Dg(v) &= \left\|J^{-1}R\right\|^{-2} \left\langle T, J^{-1}H(v,\cdot)J^{-1}R - J^{-1}DR(v) \right\rangle \qquad \text{(B.4)} \end{aligned}$$

Along the Newton direction $\Delta x := -J^{-1}F$ we have $J(\Delta x) = -F$ and hence $DR(\Delta x) =$

0. (B.4) then yields

$$\begin{aligned} Dg(\Delta x) &= \left\| \frac{\Delta x}{\|F\|} \right\|^{-2} \left\langle -\frac{\Delta x}{\|\Delta x\|}, J^{-1}H(\Delta x, \cdot)\frac{\Delta x}{\|F\|} \right\rangle \\ &= \frac{\|F\|}{\|\Delta x\|} \left\langle -\frac{\Delta x}{\|\Delta x\|}, J^{-1}H\left(\Delta x, \frac{\Delta x}{\|\Delta x\|}\right) \right\rangle. \end{aligned}$$

As we do not want to go past points with singular DF, this leads to the stepsize restriction

$$\begin{aligned} \lambda^{(k)} &\overset{!}{\leq} \frac{g(x^{(k)})}{Dg(x^{(k)})(\Delta x^{(k)})} \\ &= \left\langle \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}, DF(x^{(k)})^{-1}D^2F(x^{(k)}) \left(\Delta x^{(k)}, \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}\right) \right\rangle^{-1}, \end{aligned} \tag{ES}$$

whenever $\left\langle \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}, DF(x^{(k)})^{-1}D^2F(x^{(k)}) \left(\Delta x^{(k)}, \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}\right) \right\rangle > 0$, for the damped Newton step

$$x^{(k+1)} := x^{(k)} + \lambda^{(k)}\Delta x^{(k)}.$$

Note. *In his thesis [Ste11], Steinhoff proposed the stepsize control*

$$\lambda^{(k)} = \max\left\{ \left| \left\langle \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}, DF(x^{(k)})^{-1}D^2F(x^{(k)}) \left(\Delta x^{(k)}, \frac{\Delta x^{(k)}}{\|\Delta x^{(k)}\|}\right) \right\rangle \right|^{-1}, 1 \right\},$$

which obviously fulfills (ES).

For the purposes of stepsize control it is sufficient to consider the behavior of g along lines, i.e. of $g(x + \epsilon v, v)$ as ϵ varies. As mentioned in Section B.1.1, g is undefined in the interior of a line segment on which points with $F = 0$ or singular DF are dense. However, the Newton method should terminate upon encountering such a segment and hence there is no need to move along it. Double roots, i.e. x where $F(x) = 0$ and $DF(x)$ is singular, are beyond the scope of this paper.

B.3. Case-by-case analysis

The remaining cases will be considered after introducing some necessary perturbation results.

Perturbation Lemmas

Lemma B.1. *Let $L, M \in \mathcal{L}(X, Y)$ be bounded linear operators, such that L is invertible and $\|M - L\| < \frac{1}{\|L^{-1}\|}$.*

Then M is invertible and

$$\left\|M^{-1}\right\| \leq \frac{\left\|L^{-1}\right\|}{1 - \left\|L^{-1}\right\| \|L - M\|}.$$

Furthermore,

$$\left\|L^{-1} - M^{-1}\right\| \leq \frac{\left\|L^{-1}\right\|^2 \|L - M\|}{1 - \left\|L^{-1}\right\| \|L - M\|}.$$

Proof. [AH09] □

Lemma B.2. *Let X, Y be Banach spaces and A, B: $X \to Y$ bounded linear operators. Assume that*

i) $\dim \mathcal{N}(A) < \infty$

ii) $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$,

iii) $Y = B\mathcal{N}(A) \oplus \mathcal{R}(A)$ *with corresponding projectors P, $Id - P$.*

Let $\widehat{X} \subseteq X$ be a complement of $\mathcal{N}(A)$ with projectors $Id - Q$, Q and $A^* : \mathcal{R}(A) \to \widehat{X}$, $B^* : B\mathcal{N}(A) \to \mathcal{N}(A)$ the inverses of the respective restrictions of A and B. Then

- *$(A + \epsilon B)$ is invertible for sufficiently small $\epsilon > 0$*
- *$\forall b \in Y : \exists u, w \in X : (A + \epsilon B)(u + \frac{1}{\epsilon} w) \to b$ as $\epsilon \to 0$*
- *$\left\|(A^*(Id - P) + \frac{1}{\epsilon} B^* P) - (A + \epsilon B)^{-1}\right\| = O(1) \quad (\epsilon \to 0)$*

Proof. All limits are with respect to $\epsilon \to 0$. For the case of infinite dimensions, note that $B\mathcal{N}(A)$ has finite dimension. Hence B^* is bounded, the complement $\mathcal{R}(A)$ is closed and so A^* is also bounded.

Let $u := A^*(Id - P)b$ and $w := B^* P b \in \mathcal{N}(A)$. Then $(A + \epsilon B)(u + \frac{1}{\epsilon} w) = (Id - P)b + \epsilon B A^*(Id - P)b + Pb = b + O(\epsilon) \to b$.

A norm on X is given by $\|x\|_\epsilon := \|(Id - Q)x + \epsilon Q x\|$ for $x \in X$. We also use $\|\cdot\|_\epsilon$ to denote operator norms which are induced by $(X, \|\cdot\|_\epsilon)$. We have

$$\left(A^*(Id - P) + \frac{1}{\epsilon} B^* P\right)^{-1} = A(Id - Q) + \epsilon B Q = A + \epsilon B - \epsilon(Id - Q)B$$

and $\left\|A^*(Id - P) + \frac{1}{\epsilon} B^* P\right\|_\epsilon = O(1)$ (recall that $\mathcal{R}(B^* P) = \mathcal{N}(A)$), $\|A(Id - Q) + \epsilon B Q\|_\epsilon = O(1)$ and $\|\epsilon(Id - Q)B\|_\epsilon = O(\epsilon)$. With Lemma B.1 it follows that $A + \epsilon B$ is invertible for sufficiently small $\epsilon > 0$ and that $\left\|(A + \epsilon B)^{-1} - \left(A^*(Id - P) + \frac{1}{\epsilon} B^* P\right)\right\|_\epsilon = O(\epsilon)$. As the ordinary and $\epsilon-$norms are equivalent with constants of order ϵ resp. $\frac{1}{\epsilon}$, the final claim follows. □

Remark B.3. *The assumptions of Lemma B.2 imply that A is a Fredholm operator of index 0.*

Remark B.4. *If $\mathcal{N}(A) \cap \mathcal{N}(B) \neq \{0\}$, $A + \epsilon B$ is not invertible for any ϵ. In this case one can first restrict A and B to a complement of $\mathcal{N}(A) \cap \mathcal{N}(B)$ and then apply Lemma B.2.*

Remark B.5. *If $B\mathcal{N}(A) \oplus \mathcal{R}(A) \neq Y$, then $A + \epsilon B$ has a singular value $O(\epsilon^2)$ (cf. Appendix B.8) and hence its inverse (if it exists) grows at least with order ϵ^{-2}.*

Remark B.6. *The term $A^*(Id - P)$ in the approximate inverse of $A + \epsilon B$ matters only in the ϵ-norm and we also have*

$$\left\| \frac{1}{\epsilon} B^* P - (A + \epsilon B)^{-1} \right\| = O(1) \quad (\epsilon \to 0).$$

Lemma B.7. *If $\widehat{X} = B^{-1}\mathcal{R}(A) := \{x \in X : Bx \in \mathcal{R}(A)\}$ in the setting of Lemma B.2, then, for all $y \in \mathcal{R}(A)$,*

$$\left\| \left(A^* - (A + \epsilon B)^{-1}\right) y \right\| = O(\epsilon) \quad (\epsilon \to 0).$$

Proof. Note that $\mathcal{R}\left((A + \epsilon B)A^*\right) \subseteq \mathcal{R}(A)$. Hence for the iteration $x_0 := A^* y$, $r_i := (A + \epsilon B)x_i - y$, $x_{i+1} := x_i + A^* r_i$ $(i = 0, 1, \ldots)$ it holds that $r_i \in \mathcal{R}(A)$, $r_i = O(\epsilon^{i+1})$ and, for sufficiently small ϵ, $x_i \xrightarrow[i\to\infty]{} x^* =: (A + \epsilon B)^{-1} y$ with $\|x_0 - x^*\| = O(\epsilon)$. □

We consider now separately the three cases in the definition of g.

DF singular, $F \notin \mathcal{R}(DF)$

First, we briefly note that g is continuous:

Lemma B.8. *Let $DF(x_0)$ be singular and $F(x_0) \notin \mathcal{R}\left(DF(x_0)\right)$. Then g is continuous at x_0.*

Proof. As $F(x_0) \notin \mathcal{R}\left(DF(x_0)\right)$, there exists $u \in \mathcal{R}(DF(x_0))^\perp$ such that $\langle u, R(x)\rangle > c > 0$ in a neighborhood of x_0. If $v(x) := DF(x)^{-1}R(x)$ exists, then

$$c < \langle u, DF(x)v(x)\rangle = \langle u, (DF(x) - DF(x_0))v(x)\rangle = v(x) \cdot O(\|x - x_0\|).$$

Otherwise $g(x) = 0$ and so in either case $g(x) = O(\|x - x_0\|)$ and hence $g(x) \to 0 = g(x_0)$ as $x \to x_0$. □

The interesting result for convergence is the following:

Theorem B.9. *Let $J(x_0)$ be singular, $\mathcal{N}(J(x_0)) \cap \mathcal{N}(H(x_0)(v, \cdot)) = \{0\}$ and $F(x_0) \notin \mathcal{R}\left(J(x_0)\right)$. Then for any $v \neq 0$, $\epsilon \mapsto g(x_0 + \epsilon v, v)$ has a directional derivative at $\epsilon = 0$ which is locally Lipschitz-continuous.*

Proof. During this proof, evaluations of any function at x_0 are denoted by the subscript 0, otherwise functions are evaluated at $x_0 + \epsilon v$, which is suppressed, and we redefine $H := H(v, \cdot)$. All limits are with respect to $\epsilon \searrow 0$.

Appendix B. Step size control for Newton's method in the presence of singularities

$\mathcal{N}(J_0) \cap \mathcal{N}(H_0) = \{0\}$ implies that J is nonsingular for small $\epsilon \neq 0$ (cf. Section B.8) and hence $\frac{d}{d\epsilon} g$ is given by (B.4). It remains to show that the limit exists as $\epsilon \searrow 0$ and that $\frac{d}{d\epsilon^2} g$ is bounded.

Note that $R \to R_0$ with DR bounded in a neighborhood of x_0 since $F_0 \neq 0$. Applying Lemma B.1 and Remark B.6 with $A = J_0$, $B = H_0$ we have

$$\begin{aligned} J^{-1}R &= (J_0 + \epsilon H_0 + O(\epsilon^2))^{-1} R = (J_0 + \epsilon H_0)^{-1} R + O(1) \\ &= \frac{1}{\epsilon} H_0^* P R + O(1) = \frac{1}{\epsilon} H_0^* P R_0 + O(1). \end{aligned}$$

Similarly $J^{-1} H J^{-1} R = \frac{1}{\epsilon^2} H_0^* P H_0 H_0^* P R + O\left(\frac{1}{\epsilon}\right) = \frac{1}{\epsilon^2} H_0^* P R + O\left(\frac{1}{\epsilon}\right)$ and $J^{-1} DR = O\left(\frac{1}{\epsilon}\right)$.

$H_0^* P R_0 \neq 0$ because $F_0 \notin \mathcal{R}(J_0)$ and hence $T \to \frac{H_0^* P R_0}{\left\| H_0^* P R_0 \right\|} =: T_0$. Inserting into (B.4) we obtain

$$\begin{aligned} \frac{d}{d\epsilon} g &= \left\| J^{-1} R \right\|^{-2} \left\langle T, J^{-1} H(v, \cdot) J^{-1} R - J^{-1} DR(v) \right\rangle \\ &= \epsilon^2 \left(\left\| H_0^* P R_0 \right\|^{-2} + O(\epsilon) \right) \left\langle T, \frac{1}{\epsilon^2} H_0^* P R_0 + O\left(\frac{1}{\epsilon}\right) \right\rangle \\ &\to \frac{\langle T_0, H_0^* P R_0 \rangle}{\left\| H_0^* P R_0 \right\|^2}. \end{aligned}$$

For $h = J^{-1} R$ we have, using (B.3),

$$\begin{aligned} \frac{d}{d\epsilon^2} h &= 2 J^{-1} H J^{-1} H J^{-1} R + O\left(\frac{1}{\epsilon^2}\right) = \frac{2}{\epsilon^3} H_0^* P H_0 H_0^* P H_0 H_0^* P R + O\left(\frac{1}{\epsilon^2}\right) \\ &= \frac{2}{\epsilon^3} H_0^* P R + O\left(\frac{1}{\epsilon^2}\right) \text{ and} \\ \frac{d}{d\epsilon^2} g &= 3 \left\| J^{-1} R \right\|^{-5} \left\langle J^{-1} R, J^{-1} H J^{-1} R + O\left(\frac{1}{\epsilon}\right) \right\rangle^2 \\ &\quad - \left\| J^{-1} R \right\|^{-3} \left\| J^{-1} H J^{-1} R + \left(\frac{1}{\epsilon}\right) \right\|^2 \\ &\quad - \left\| J^{-1} R \right\|^{-3} \left\langle J^{-1} R, 2 J^{-1} H J^{-1} H J^{-1} R + O\left(\frac{1}{\epsilon^2}\right) \right\rangle \\ &= 3 \left\| \frac{1}{\epsilon} H_0^* P R + O(1) \right\|^{-5} \left\langle \frac{1}{\epsilon} H_0^* P R + O(1), \frac{1}{\epsilon^2} H_0^* P R + O\left(\frac{1}{\epsilon}\right) \right\rangle^2 \\ &\quad - \left\| \frac{1}{\epsilon} H_0^* P R + O(1) \right\|^{-3} \left\| \frac{1}{\epsilon^2} H_0^* P R + O\left(\frac{1}{\epsilon}\right) \right\|^2 \\ &\quad - 2 \left\| \frac{1}{\epsilon} H_0^* P R + O(1) \right\|^{-3} \left\langle \frac{1}{\epsilon} H_0^* P R + O(1), \frac{1}{\epsilon^3} H_0^* P R + O\left(\frac{2}{\epsilon^2}\right) \right\rangle \\ &= O(1) \end{aligned}$$

□

In general, full differentiability is not possible unless the gradient at a singularity of J is 0.

Example Let $F(x) := \begin{pmatrix} x_1 & 0 \\ 0 & x_2 \end{pmatrix} x$. Then $g = 0$ on both $\{x : x_1 = 0\}$ and $\{x : x_2 = 0\}$.

If rank-deficiency of J is 1, g only fails to be differentiable because the sign does not change when crossing the singular manifold. This could be remedied by using $\det(J) \cdot g$ instead of g and differentiability could then be deduced from [GR84], cf. also Section B.7.

DF singular, $0 \neq F \in \mathcal{R}(DF)$

Again the subscript 0 denotes values at the singular point x_0. We have $F = F_0 + \epsilon J_0 v + O(\epsilon^2)$ and $R = R_0 + O(\epsilon)$. With Lemma B.7, $\left\|J^{-1}R\right\| = \|J_0^* R_0 + O(\epsilon)\| < C$ and $\left\|D(J^{-1}R)\right\| = \|-J_0^* H_0 J_0^* R_0 + O(1)\| = O(1)$, as $H_0 J_0^* R_0 \in \mathcal{R}(J_0)$. Hence $|g/(Dg \cdot v)| = 1/C/O(1)$, which means that the stepsize control does not detect such singularities. Since, in this particular case, they are no obstruction to solving $F = 0$, one can justify not regarding this as a defect.

Note that J_0^* and hence $g(x, v)$ depend on v through the decomposition

$$X = H_0(v, \cdot)^{-1}\mathcal{R}(J_0) \oplus \mathcal{N}(J_0).$$

DF regular, $F = 0$

Near a regular solution x_0, i.e. for $F(x_0) = 0$, $J(x_0)$ invertible, we have $F = \epsilon J_0 v + O(\epsilon^2)$, $R = \frac{J_0 v}{\|J_0 v\|} + O(\epsilon)$, $J^{-1}R = \frac{v}{\|J_0 v\|} + O(\epsilon)$, $C > \left\|J^{-1}R\right\|$ and $\left\|D(J^{-1}R)\right\| = \left\|-J^{-1}HJ^{-1}R + DR\right\| = O(1)$ as $x \to x^*$, uniformly for all v and in some neighborhood of x^*. This implies $g > C^{-1}$, $Dg = O(\|x - x^*\|)$ and so the indicator will not erroneously predict a nearby singularity although g is in general not continuous at x_0 and $g(x_0, v)$ depends on v.

B.4. Convergence to singularity

By the preceding remarks, the constraint is only active in the neighborhood of points with $F \notin \mathcal{R}(DF)$. If the direction $\Delta x / \|\Delta x\|$ is assumed to be fixed, Theorem B.9 ensures local quadratic convergence of g to 0.[2] In well-behaved cases, in particular if $\mathcal{N}(DF)$ is one-dimensional and Δx is therefore always approximately aligned to the direction uniquely given by $\mathcal{N}(DF)$, there is a common region of quadratic convergence for all $\Delta x^{(k)}$ which arise during the iteration.

[2] Assuming, of course, that equality is chosen in (ES) and $g/(Dg \cdot \Delta x) < 0$, i.e. the Newton direction actually points towards the singular manifold.

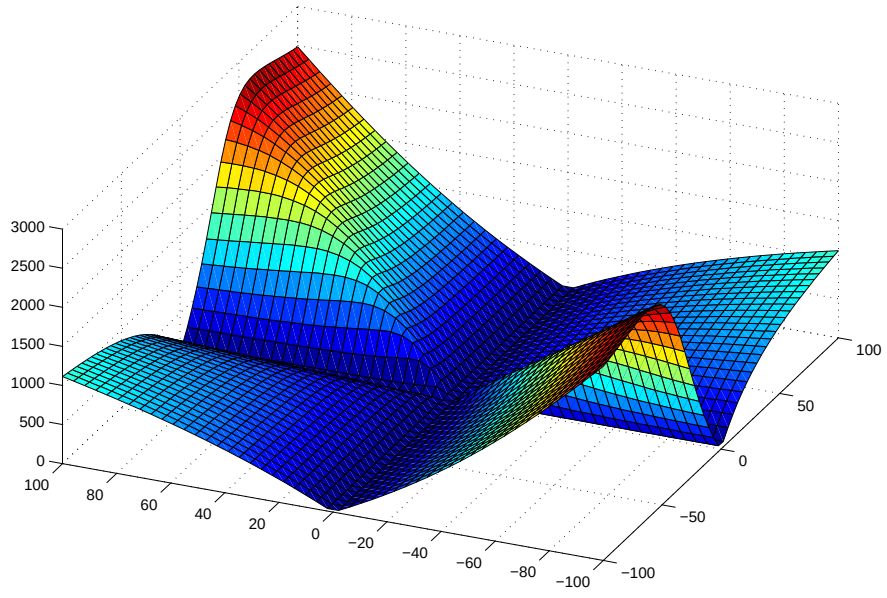

Figure B.1.: g at the crossing of two singular manifolds

In general this is hard to guarantee and problems arise when singular manifolds intersect. Figure B.1 shows g for the function

$$F : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto -\left(x_1 \begin{pmatrix} 5 & 10 \\ 2 & 4 \end{pmatrix} + x_2 \begin{pmatrix} 4 & 2 \\ 6 & 3 \end{pmatrix}\right) x - 10^6 \begin{pmatrix} 1.1 \\ 1 \end{pmatrix},$$

which has a singular Jacobian along both axes and a rank-deficiency of two at the origin. E.g. at $(x_1, 0)$, $|x_1| \ll 1$, $x_1 \neq 0$, the Jacobian has a singular value which is small but not zero. Consequently the pseudoinverse J^* has a large norm and dominates $\frac{1}{\epsilon}H^*$ even for quite small ϵ, which limits the neighborhood where the approximation of Theorem B.9 is applicable.

In contrast, the Jacobian of,

$$F : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto -x_1 \begin{pmatrix} 1 & 7 \\ 8 & 3 \end{pmatrix} x - 10^6 \begin{pmatrix} 1.1 \\ 1 \end{pmatrix}$$

has rank-deficiency two along the whole x_2-axis, but no actual crossing of singular manifolds (Figure B.2).

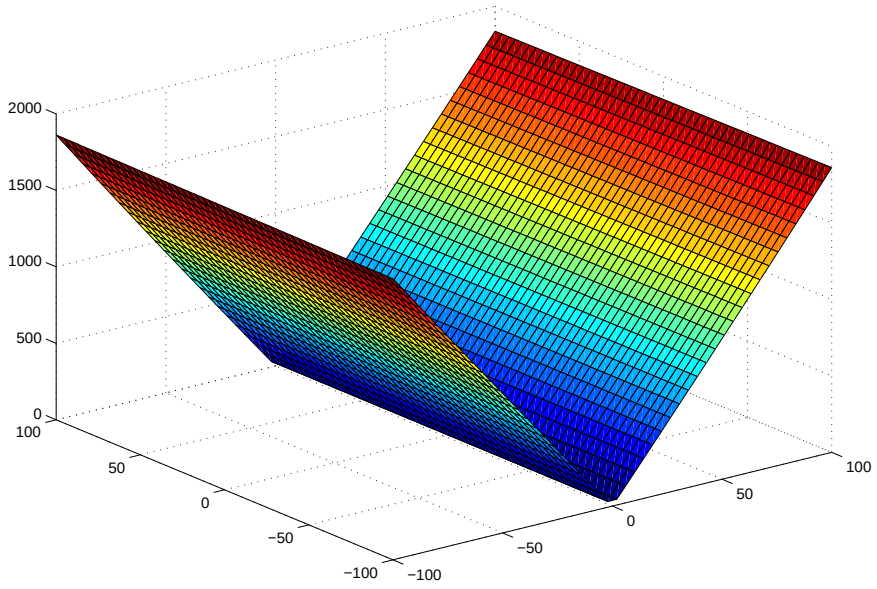

Figure B.2.: g near coinciding singular manifolds

B.5. Approximate stepsize control

We observe that (ES) is similar to the stepsize control in Deuflhard's affine covariant globalization, which is

$$\lambda^{(k)} \overset{!}{\leq} \left\| DF\left(x^{(k)}\right)^{-1} D^2F\left(x^{(k)}\right)\left(\Delta x^{(k)}, \frac{\Delta x^{(k)}}{\left\|\Delta x^{(k)}\right\|}\right)\right\|^{-1}. \tag{AS}$$

Near singular DF we find for any v that $DF^{-1}v$, and hence both sides of the scalar product in (ES), will approximately be multiples of the left singular vector belonging to the smallest singular vector of DF, which suggests that (ES) and (AS) coincide in the limit.

Indeed, we can show that this choice of stepsize works well if the smallest singular value of DF is isolated and some genericity conditions hold:

Let $DF(x) = U(x)\Sigma(x)V(x)^T$ be a smooth SVD of $DF(x)$ as defined in Section B.8, with $U = (u_1, \ldots, u_n)$, $V = (v_1, \ldots, v_n)$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Then

Appendix B. Step size control for Newton's method in the presence of singularities

$$
\begin{aligned}
DF^{-1}F &= \sum_{k=1}^{n} \frac{1}{\sigma_k} \langle u_k, F\rangle v_k \\
&= \frac{1}{\sigma_n}\left[\langle u_n, F\rangle v_n + O\left(\frac{\sigma_n}{\sigma_{n-1}}\|F\|\right)\right].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
-\Delta x &= DF^{-1}D^2F(\cdot,-\Delta x) \\
&= \frac{1}{\sigma_n}\left[\left\langle u_n, D^2F(\cdot,-\Delta x)\right\rangle v_n + O\left(\frac{\sigma_n}{\sigma_{n-1}}\left\|D^2F(\cdot,-\Delta x)\right\|\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
&DF^{-1}D^2F(-\Delta x,-\Delta x) = \\
&= \tfrac{1}{\sigma_n}\left[v_n u_n^T D^2F\left(\tfrac{1}{\sigma_n}\left[\langle u_n, F\rangle v_n + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|F\|\right)\right], -\Delta x\right)\right. \\
&\quad \left. + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\left\|D^2F(-\Delta x,-\Delta x)\right\|\right)\right] \\
&= \tfrac{1}{\sigma_n^2}\left[\langle u_n, F\rangle v_n u_n^T D^2F(v_n,-\Delta x) + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|F\|\left\|D^2F(\cdot,-\Delta x)\right\|\right)\right] \\
&\quad + \tfrac{1}{\sigma_n} O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\left\|D^2F(-\Delta x,-\Delta x)\right\|\right) \\
&= \tfrac{1}{\sigma_n^2}\left[\langle u_n, F\rangle \langle \nabla\sigma_n, -\Delta x\rangle v_n + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|F\|\left\|D^2F(\cdot,-\Delta x)\right\|\right)\right] \\
&\quad + \tfrac{1}{\sigma_n} O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\left\|D^2F(-\Delta x,-\Delta x)\right\|\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\tfrac{\|\Delta x\|}{\|DF^{-1}D^2F(-\Delta x,-\Delta x)\|} = \\
&= \left|\frac{\tfrac{1}{\sigma_n}\left[\langle u_n,F\rangle + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|F\|\right)\right]}{\tfrac{1}{\sigma_n^2}\left[\langle u_n,F\rangle\langle\nabla\sigma_n,-\Delta x\rangle + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|F\|\|D^2F(\cdot,-\Delta x)\|\right)\right] + \tfrac{1}{\sigma_n}O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|D^2F(-\Delta x,-\Delta x)\|\right)}\right| \\
&= \left|\frac{1 + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\tfrac{\|F\|}{\langle u_n,F\rangle}\right)}{\tfrac{1}{\sigma_n}\left[\langle\nabla\sigma_n,-\Delta x\rangle + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\tfrac{\|F\|}{\langle u_n,F\rangle}\|D^2F(\cdot,-\Delta x)\|\right)\right] + O\left(\tfrac{\sigma_n}{\sigma_{n-1}}\|D^2F(-\Delta x,-\Delta x)\|\right)}\right|.
\end{aligned}
$$

As

$$
\left\|D^2F(-\Delta x,-\Delta x)\right\| = O\left(\frac{1}{\sigma_n}\|F\|\left\|D^2F(\cdot,-\Delta x)\right\|\right)
$$

and

$$
\langle\nabla\sigma_n,-\Delta x\rangle = \left\langle u_n, D^2F(v_n,-\Delta x)\right\rangle,
$$

it follows that

$$
\begin{aligned}
&\frac{\|\Delta x\|}{\|DF^{-1}D^2F(-\Delta x,-\Delta x)\|} = \\
&\left|\frac{\sigma_n}{\langle\nabla\sigma_n,-\Delta x\rangle}\right|\left(1 + O\left(\frac{\sigma_n}{\sigma_{n-1}}\frac{\|F\|}{\langle u_n,F\rangle}\frac{\left\|D^2F(\cdot,-\Delta x)\right\|}{\langle u_n, D^2F(v_n,-\Delta x)\rangle}\right)\right),
\end{aligned}
\tag{B.5}
$$

and hence for the stepsize

$$\lambda^{(k)} \overset{!}{\leq} \left| \frac{\sigma_n}{\langle \nabla \sigma_n, -\Delta x \rangle} \right| \approx \left\| DF\left(x^{(k)}\right)^{-1} D^2F\left(x^{(k)}\right) \left(\Delta x^{(k)}, \frac{\Delta x^{(k)}}{\left\|\Delta x^{(k)}\right\|} \right) \right\|^{-1}.$$

If $\frac{1}{\sigma_{n-1}} \frac{\|F\|}{\langle u_n, F \rangle} \frac{\left\|D^2F(\cdot, -\Delta x)\right\|}{\langle u_n, D^2F(v_n, -\Delta x)\rangle}$ is bounded in some suitable neighborhood, the error term in (B.5) reduces to $O(\sigma_n)$ and σ_n converges quadratically to 0 for the iteration with stepsize (AS).[3]

B.6. Numerical experiments

We apply both stepsize controls to the test problem *Expsin* from [NW92], which also appears as Example 3.2 in [Deu04]. The task is to find solutions of

$$\begin{aligned} \exp(x^2 + y^2) - 3 &= 0, \\ x + y - \sin(3(x + y)) &= 0. \end{aligned} \tag{B.6}$$

The Jacobian is singular along the lines $y = x$ and

$$y = -x \pm \frac{1}{3} \arccos\left(\frac{1}{3}\right) \pm \frac{2}{3}\pi.$$

Figure B.3 shows the path of Newton iterates with stepsize control (ES) started at points in the square $[-1.5, 1.5]^2$. Note that (B.6) has solutions only in the six middle regions and observe that iterates to initial points in regions without a solution eventually converge to a singular manifold while oscillating around it. Some of the paths do, however, cross several singular manifolds and end up at a "wrong" one. The approximate stepsize control appears more robust as the norm in (AS) does not ignore the components perpendicular to T. Indeed, all iterations converged to the correct solution or singular manifold when applying the approximate stepsize control to the same example, but the convergence to singular manifolds was slightly slower as exemplified in Figure B.4.

As mentioned in Section B.5, (ES) and (AS) are in general approximately equal close to singular manifolds, which suggests to compute both and use (ES) only when they differ by less than some small factor. Due to shared terms this involves almost no additional effort. In our experiments this heuristic combined the faster convergence of (ES) with the robustness of (AS).

B.7. Connection to Griewank and Reddien

The indicator g is inspired by the following Lemma of Griewank and Reddien, which gives an indicator for the rank-deficiency of rectangular matrices.

[3] As before assuming equality in (AS) and $\sigma_n / \langle \nabla \sigma_n, -\Delta x \rangle < 0$.

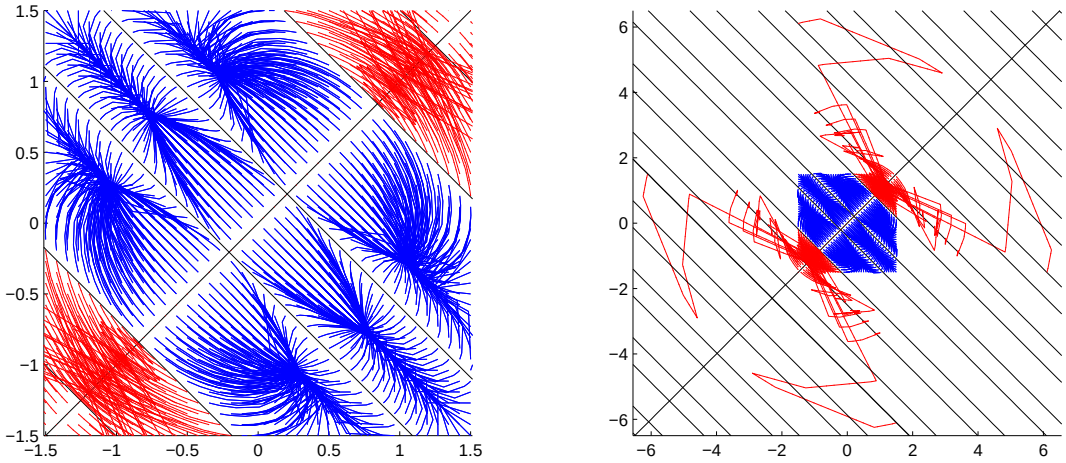

Figure B.3.: Iterates with stepsize control (ES) for several initial values at two zoom levels. Blue paths converge to solutions, red paths converge to singular manifolds, black lines are singular manifolds

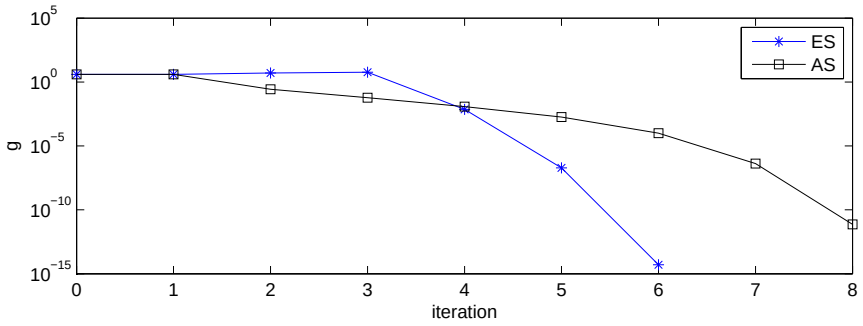

Figure B.4.: Convergence of g for $(x_0, y_0) = (-0.5, -1.5)$

Lemma B.10. *([GR84], Lemma 2.1) Let $D \subset \mathbb{R}^d$, $m \leq n$ and define $p := n+1-m$. Let $A : D \to \mathbb{R}^{m\times n}$, $T : D \to \mathbb{R}^{n\times p}$ and $R : D \to \mathbb{R}^m$ be continuously differentiable functions so that*

$$\begin{pmatrix} A & -R \\ -T^T & 0 \end{pmatrix}$$

is nonsingular for all $x \in D$. Then there are unique functions $u : D \to \mathbb{R}^m$, $V : D \to \mathbb{R}^{n\times p}$ and $g : D \to \mathbb{R}^p$ such that

$$\begin{aligned} AV &= Rg^T, & T^TV &= I, \\ u^TA &= g^TT^T, & u^TR &= 1, \end{aligned}$$

and

$$\operatorname{rank}(A(x)) = m - 1 \text{ if and only if } g(x) = 0.$$

Lemma B.11. *For square matrices the indicator g of Lemma B.10 coincides with* (B.2).

Proof. As $n = m$, we have $p = 1$, g is a scalar and T, R and V are vectors. From $T^T V = I$ we obtain the decomposition $V = T + V^{(\perp)}$, where $V^{(\perp)} \in T^\perp := \operatorname{span}\{T\}^\perp$. Projection of $gR = AV = AV^{(\perp)} + AT$ onto $(AT^\perp)^\perp$ yields $gP_{(AT^\perp)^\perp}R = P_{(AT^\perp)^\perp}AT$. As $(AT^\perp)^\perp = \operatorname{span}\left\{A^{-T}T\right\}$, it follows that

$$g = \frac{\left\langle A^{-T}T, AT\right\rangle}{\langle A^{-T}T, R\rangle} = \frac{\langle T, A^{-1}AT\rangle}{\langle T, A^{-1}R\rangle} = \frac{\|T\|^2}{\langle T, A^{-1}R\rangle}. \qquad \square$$

B.8. Smooth singular value decomposition

The singular value decomposition (SVD) of a smooth (i.e. sufficiently differentiable) family of matrices is only smooth if the singular values do not cross each other or 0. However, a smooth SVD can sometimes be obtained if the requirements $\sigma_k \geq 0$, $\sigma_k > \sigma_{k+1}$ are given up. Such decompositions have already been studied extensively, see e.g. [BGBMN91], [Wri92]. We will give here a concise proof for the existence of a differentiable SVD of a square matrix with isolated singular values, where one singular value crosses 0, which is sufficient for our purposes.

Let $A = U\Sigma V$ be the SVD of $A \in \mathbb{R}^{n\times n}$ and $B := A^T A$. For an eigenpair (λ, v) of B, consider the variation of its defining equations

$$Bv = \lambda v, \tag{B.7}$$
$$\|v\| = 1. \tag{B.8}$$

We have from (B.7)

$$\begin{aligned}(B + dB)(v + dv) &= (\lambda + d\lambda)(v + dv)\\ \Rightarrow dB \cdot v &= -(B - \lambda)dv + d\lambda \cdot v.\end{aligned} \tag{B.9}$$

From (B.8) it follows that $v \perp dv$. As the eigenspace is one-dimensional, this implies $v \perp Bdv$ and hence (B.9) can be decomposed into

$$d\lambda = \langle v, dB \cdot v\rangle \tag{B.10}$$
$$P_v^\perp dB \cdot v = (\lambda - B)dv, \tag{B.11}$$

where $P_v^\perp$ is the orthogonal projector onto $\operatorname{span}\{v\}^\perp$. If λ is an isolated eigenvalue of B, then $\mathcal{N}(B - \lambda)$=span $\{v\}$ and the unique solution of (B.11) under the constraint $v \perp dv$ is given by

$$dv = (\lambda - B)^+ dB \cdot v, \tag{B.12}$$

where $^+$ denotes the Moore-Penrose pseudoinverse.

Appendix B. Step size control for Newton's method in the presence of singularities

If $\lambda \neq 0$, we obtain smooth functions for the corresponding singular value σ of A and a left singular vector u with $\|u\| = 1$ by taking

$$\sigma = \sqrt{\lambda}, \tag{B.13}$$

$$u = \frac{Av}{\sigma}. \tag{B.14}$$

From $d\lambda = \langle v, dB \cdot v\rangle = \langle v, (A^T dA + dA^T A)v\rangle = 2\langle Av, dA \cdot v\rangle = \frac{2}{\sigma}\langle u, dA \cdot v\rangle$ and $\sigma = \sqrt{\lambda}$ we obtain

$$d\sigma = \langle u, dA \cdot v\rangle . \tag{B.15}$$

If $\lambda = 0$, we have (using $Av = 0$)

$$\begin{aligned}
(A + dA)(v + dv) &= Av + Adv + dA \cdot v \\
&= A(0 - B)^+ dB \cdot v + dA \cdot v \\
&= -AB^+(dA^T A + A^T dA)v + dA \cdot v \\
&= (Id - A(A^T A)^+ A^T)dA \cdot v \\
&= P_{\mathcal{R}(A)^\perp} dA \cdot v \\
&\overset{!}{=} (\sigma + d\sigma)(u + du) \\
&= d\sigma \cdot u.
\end{aligned}$$

In this case, u is given (up to a sign change) by $\mathcal{R}(A)^\perp = \operatorname{span}\{u\}$ and we obtain again

$$d\sigma = \langle u, dA \cdot v\rangle . \tag{B.16}$$

As u is an eigenvector of $AA^T = B^T$, a calculation similar to (B.9) yields (for all σ)

$$du = (\lambda - B^T)^+ dB^T \cdot u. \tag{B.17}$$

Bibliography

[AH09] Kendall Atkinson and Weimin Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Texts in Applied Mathematics. Springer, New York, NY, 2009.

[ASG01] Athanasios C Antoulas, Danny C Sorensen, and Serkan Gugercin. A survey of model reduction methods for large-scale systems. *Contemporary mathematics*, 280:193–220, 2001.

[Bel54] Richard Bellman. Dynamic programming and a new formalism in the calculus of variations. *Proceedings of the National Academy of Sciences of the United States of America*, 40(4):231–235, 1954.

[BG04] Hans-Joachim Bungartz and Michael Griebel. Sparse Grids. *Acta Numerica*, 13:147–269, May 2004.

[BGBMN91] Angelika Bunse-Gerstner, Ralph Byers, Volker Mehrmann, and Nancy K. Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numer. Math*, 60:1–40, 1991.

[BH69] Arthur E. Bryson and Yu-Chi Ho. *Applied optimal control: optimization, estimation, and control*. Blaisdell book in the pure and applied sciences. Blaisdell Pub. Co., 1969.

[Bol02] Oskar Bolza. Some instructive examples in the calculus of variations. *Bull. Amer. Math. Soc.*, 9(1):1–10, 10 1902.

[Bow81] A. Bowyer. Computing dirichlet tessellations. *The Computer Journal*, 24(2):162–166, 1981.

[BP84] Hans Georg Bock and Karl-Josef Plitt. A multiple shooting algorithm for direct solution of optimal control problems. Proceedings of the IFAC World Congress, 1984.

[Bre84] Karl Breitung. Asymptotic approximations for multinormal integrals. *Journal of Engineering Mechanics*, 110(3):357–366, 1984.

[CGC06] S.K. Choi, R.V. Grandhi, and R.A. Canfield. *Reliability-based Structural Design*. Springer London, 2006.

[CH87] D.A. Carlson and A. Haurie. *Infinite horizon optimal control: theory and applications*. Lecture notes in economics and mathematical systems. Springer, 1987.

[Cla13] F. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Graduate Texts in Mathematics. Springer London, 2013.

[Cop78] W.A. Coppel. *Dichotomies in stability theory*. Lecture notes in mathematics. Springer, 1978.

[Dac07] Bernard Dacorogna. *Direct methods in the calculus of variations*, volume 78 of *Applied Mathematical Sciences*. Second edition, 2007.

[Deu74] Peter Deuflhard. A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting. *Numer. Math.*, 22:289 – 315, 1974.

[Deu04] Peter Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer Series in Computational Mathematics. Springer, 2004.

[Eis01] Nathalie Eisenbaum. On Ito's formula of Föllmer and Protter. *Séminaire de probabilités de Strasbourg*, 35:390–395, 2001.

[Est95] Donald Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM Journal on Numerical Analysis*, 32(1):pp. 1–48, 1995.

[Eva10] L.C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. Second edition, 2010.

[FF94] Marizio Falcone and Roberto Ferretti. Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations. *Numer. Math.*, 67:315–344, 1994.

[GG73] M. Golubitsky and V. Guillemin. *Stable Mappings and Their Singularities*. Graduate Texts in Mathematics. Springer New York, 1973.

[GJ05] Lars Grüne and Oliver Junge. A set oriented approach to optimal feedback stabilization. *Systems & control letters*, 54(2):169–180, 2005.

[GR84] Andreas Griewank and George W. Reddien. Characterization and computations of generalized turning points. *SIAM Journal of Numerical Analysis*, 21(1):176–185, 1984.

[Grü97] Lars Grüne. An Adaptive Grid Scheme for the discrete Hamilton-Jacobi-Bellman Equation. *Numer. Math.*, 75:319–337, 1997.

[HL74] Abraham M Hasofer and Niels C Lind. Exact and invariant second-moment code format. *Journal of the Engineering Mechanics division*, 100(1):111–121, 1974.

[IW89] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Kodansha scientific books. North-Holland, 1989.

[JO04] Oliver Junge and Hinke M Osinga. A set oriented approach to global optimal control. *ESAIM: Control, optimisation and calculus of variations*, 10(2):259–270, 2004.

[JS15] Oliver Junge and Alex Schreiber. Dynamic programming using radial basis functions. *Discrete and Continuous Dynamical Systems*, 35(9):4439–4453, 2015.

[Kan01] Yu.S. Kan. Control optimization by the quantile criterion. *Automation and Remote Control*, 62(5):746–757, 2001.

[Kat58] Tosio Kato. Perturbation theory for nullity, deficiency and other quantities of linear operators. *Journal d'Analyse Mathématique*, 6:261–322, 1958.

[Kie06] Hansjörg Kielhöfer. *Bifurcation theory: An introduction with applications to PDEs*, volume 156. Springer Science & Business Media, 2006.

[Kri12] Wikimedia Commons (User Krishnavedala). Schematic drawing of an inverted pendulum on a cart., 2012.

[Kuc72] Vladimir Kucera. A contribution to matrix quadratic equations. *Automatic Control, IEEE Transactions on*, 17(3):344–347, 1972.

[KVX04] K. Kunisch, S. Volkwein, and L. Xie. HJB-POD-based feedback design for the optimal control of evolution problems. *SIAM Journal on Applied Dynamical Systems*, 3(4):701–722, 2004.

[LM98] Andrzej Lasota and Michael C Mackey. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, volume 97. Springer Science & Business Media, 1998.

[Mor25] Marston Morse. Relations between the critical points of a real function of n independent variables. *Transactions of the American Mathematical Society*, 27(3):345–396, 1925.

[NW92] Ulrich Nowak and Lutz Weimann. A family of newton codes for systems of highly nonlinear equations. Technical Report TR-91-10, Zuse Institute Berlin, 1992.

[OH06] Hinke M. Osinga and John Hauser. The geometry of the solution set of nonlinear optimal control problems. *Journal of Dynamics and Differential Equations*, 18(4):881–900, 2006.

[Pon87] Lev Semenovich Pontryagin. *Mathematical theory of optimal processes.* CRC Press, 1987.

[Rac76] R. Rackwitz. *Practical Probabilistic Approach to Design.* Bulletin d'information. Technical University of Munich, Institut für Bauingenieurwesen III, 1976.

[Ste11] Tim Steinhoff. *Approximate and Projected Natural Level Functions for Newton-type Iterations.* Dissertation, Technische Universität Hamburg-Harburg, 2011.

[Tes12] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[Tre08] Lloyd N Trefethen. Is gauss quadrature better than clenshaw-curtis? *SIAM review*, 50(1):67–87, 2008.

[vdS96] Arjan van der Schaft. *L2-Gain and Passivity Techniques in Nonlinear Control*, volume 218 of *Lecture Notes in Control and Information Sciences.* Springer, 1996.

[Wal06] Jörg Waldvogel. Fast construction of the fejér and clenshaw–curtis quadrature rules. *BIT Numerical Mathematics*, 46(1):195–202, 2006.

[Wat81] David F Watson. Computing the n-dimensional delaunay tessellation with application to voronoi polytopes. *The computer journal*, 24(2):167–172, 1981.

[Wri92] K. Wright. Differential equations for the analytic singular value decomposition of a matrix. *Numer. Math*, 63:283–295, 1992.

[Zei95] E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications.* Applied Mathematical Sciences. Springer New York, 1995.