Chapter 2

Optimality Conditions

2.1 Introduction

This introductory chapter is intended to set the stage for the optimization and discretization methods to come in the subsequent chapters. On the one hand, it presents the necessary notation and recalls well-known optimality conditions in differential optimization where we borrow a lot of ideas from [339]. On the other hand, it tries to provide also a certain point of view on optimization such that the subsequent discussions can be understood more easily. In particular, the point of view of the choice of functional spaces and corresponding scalar products is of importance for the efficiency as well as efficacy of the optimization methods to be employed. This topic is tightly intertwined with spectral Hessian analysis as well as preconditioning of one-shot methods.

2.2 Optimality Conditions

In this book, we will consider optimization problems of the form

$$\min_{y,u} J(y,u), \tag{2.1}$$

$$c(y,u) = 0, \tag{2.2}$$

$$u \in U_{ad}. \tag{2.3}$$

We assume that the constraint model equation $c(y,u) = 0$ uniquely defines the state y if the decision variable u is given, which means that c_y is invertible. This particular separation of the variables in the constrained optimization problem is usually called the separability framework. The objective $J(.,.)$ is a scalar-valued function. All functions are supposed to be as smooth as required. The decision variable is constrained to some admissible set U_{ad}. For ease of presentation, we deliberately exclude state constraints $y \in Y(u)$ at the moment and refer to Section 5.7.1 for practical consequences. In this chapter, we give a brief overview on the theoretical framework of existence of solutions and optimality conditions. We will always try to frame the results in the finite-dimensional setting as well as in the function space setting. The function space setting is of importance if one tries to solve optimization with increasing space and/or time discretization refinement.

The finite-dimensional setting—although a simplified consequence of the general function space setting—has its own justification, since it is easier to understand and may thus help to obtain a better intuitive understanding of the more general concepts.

Let us first take the point of view of black-box methods; that is, let us exploit the implicit function theorem and consider the state variable y as a function depending on u, that is, $y = y(u)$ such that $c(y(u),u) = 0$. That means the constrained optimization problem is equivalently formulated as

$$\min_{u \in U_{ad}} \hat{J}(u),$$

where $\hat{J}(u) := J(y(u),u)$ is the so-called reduced objective.

Now, we state the existence of an optimal solution in a finite-dimensional setting. We consider a generic function $f : U_{ad} \to \mathbb{R}$ having in mind $\hat{J}$ as a formal particular instance.

Theorem 2.1. *If $U_{ad} \subset \mathbb{R}^n$, for some $n \in \mathbb{N}, n < \infty$, the function $f : U_{ad} \to \mathbb{R}$ is continuous and bounded from below, and the level set*

$$L_t := \{u \in U_{ad} \mid f(u) \le t\}$$

is a nonempty, closed, and bounded set for some $t < \infty$, then there exists an optimal solution $\hat{u} \in U_{ad}$.

Proof. Since f is bounded from below, there exists

$$j := \inf_{u \in U_{ad}} f(u).$$

Therefore, we can define a sequence $\{u_m\}_{m=1}^{\infty}$ with $f(u_m) \to j$ $(m \to \infty)$. For all m greater than some number $M < \infty$, all u_m have to lie in L_t with $t > j$. Since L_t is closed and bounded and therefore compact, we can select a converging subsequence $\{u_{m_k}\}_{k=1}^{\infty}$ such that $u_{m_k} \to \hat{u}$ $(k \to \infty)$. Since f is continuous, we know that $\hat{u}$ is a minimum, because

$$j = \lim_{k \to \infty} f(u_{m_k}) = f(\hat{u}). \quad \square$$

The general case, where U_{ad} is a subset of a function space, is more complicated, since closeness and boundedness no longer yield compactness. However, the situation is only slightly more complicated if $U_{ad} \subset H$, where H is a Hilbert space, that is, a complete linear space with a scalar product $(.,.)_H$. Then, the only property in addition to the finite-dimensional case is convexity to ensure existence of a solution.

Definition 2.2. *A function $f : U_{ad} \to \mathbb{R}$ is called convex if*

$$f(t\,u + (1-t)\,v) \le t\,f(u) + (1-t)\,f(v) \quad \forall u, v \in U_{ad}, \forall t \in (0,1).$$

It is called strictly convex, if the inequality is strict (i.e., $<$ instead of $\le$).

Theorem 2.3. *If $U_{ad} \subset H$, for some Hilbert space H, the function $f : U_{ad} \to \mathbb{R}$ is continuous, convex, and bounded from below, and the level set*

$$L_t := \{u \in U_{ad} \mid f(u) \le t\}$$

is a nonempty, closed, and bounded set for some $t < \infty$, then there exists an optimal solution $\hat{u} \in U_{ad}$.

Proof. See [339] or other books. □

If one aims at computing an optimal solution numerically, one should hope for some argument for uniqueness of the optimal solution in addition to existence—or at least local uniqueness is a prerequisite to make numerical computations manageable. In both cases, the finite-dimensional as well as the function space case, we observe that strict convexity of f ensures uniqueness of the optimal solution.

Theorem 2.4. *If f is strictly convex and it possesses a minimum, this minimum is unique, provided U_{ad} is a convex set.*

Proof. Let us assume the existence of two separate minima $u \neq v \in U_{ad}$ with $f(u) = j = f(v)$. Then we find that the vector $\frac{1}{2}(u+v)$ yields an even lower objective value

$$f\left(\frac{1}{2}(u+v)\right) < \frac{1}{2}(f(u)+f(v)) = j$$

because of strict convexity of f. This is in clear contrast to j being the infimum. □

Strict convexity overall in U_{ad} is a rather strong requirement. But if we have convexity in some region and can construct a minimizing sequence as above, which stays in that region, we can conclude at least local uniqueness, which is enough for numerical purposes. The minimizing sequence can be constructed by employing a variant of the steepest descent method. In order to be able to formulate this method, we need to define the gradient of a scalar function.

Definition 2.5.

(a) *For a differential function $f : \mathbb{R}^n \to \mathbb{R}$, we define the Euclidean gradient as the vector*

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial u_1} \\ \vdots \\ \frac{\partial f}{\partial u_n} \end{pmatrix}.$$

(b) *For a (Fréchet-)differentiable function $f : H \to \mathbb{R}$, where H is a Hilbert space, we define the gradient as the Riesz-representation of the derivative f', such that*

$$(\nabla f, v)_H = f'(v) \quad \forall v \in H,$$

where f' is defined as the linear operator

$$f'(u) : v \mapsto \left.\frac{d}{dt}\right|_{t=0} f(u+tv) \quad \forall v \in H.$$

With this definition, it is not possible to talk about the gradient without mentioning the specific scalar product employed to represent the derivative. Also, we do not have to deal with dual spaces, which can be consistently avoided by the usage of Riesz representations.

Let us study some examples in order to get used to this definition of the gradient.

Example 2.6. (a) First we study $H = \mathbb{R}^n$ endowed with the scalar product $(x, y) = x^\top Ay$, where A is a symmetric and positive definite matrix. The derivative of a differentiable function $f : H \to \mathbb{R}$ is defined as the mapping

$$f'(u) : v \mapsto \sum_{i=1}^{n} \frac{\partial f(u)}{\partial u_i} v_i.$$

Now

$$f'(u)v = \sum_{i=1}^{n} \frac{\partial f(u)}{\partial u_i} v_i = \left(A^{-1} \begin{pmatrix} \frac{\partial f(u)}{\partial u_1} \\ \vdots \\ \frac{\partial f(u)}{\partial u_n} \end{pmatrix}, v \right) \quad \forall v \in H.$$

Therefore

$$\nabla f(u) = A^{-1} \begin{pmatrix} \frac{\partial f(u)}{\partial u_1} \\ \vdots \\ \frac{\partial f(u)}{\partial u_n} \end{pmatrix}.$$

If we choose $A = I$, we obtain the Euclidean gradient as defined in Definition 2.5(a). If we choose it differently, we obtain a different vector as the gradient. As we will see below, a steepest descent method based on this definition of the gradient profits from a good choice of the metric A. Basically, A acts as a scaling of the variables.

(b) Let us now consider the function space $H = L^2([0,1])$ endowed with the scalar product $(x, y) := \int_0^1 x(t)\, y(t)\, \omega(t) dt$ with some weight function $\omega \in L^\infty([0,1], \mathbb{R}_+)$. We consider the function $f : H \to \mathbb{R}$ defined by

$$f(u) := \int_0^1 \phi(u(t)) dt$$

for some given twice differentiable function $\phi : \mathbb{R} \to \mathbb{R}$ with bounded second derivative. This is the integral over the so-called Nemytskii operator, whose first derivative exists in $L^2([0,1])$ but whose second derivative is a subject of subtle discussion in [339]. Now, the directional derivative is

$$f'(u)v = \frac{d}{d\alpha}\bigg|_{\alpha=0} f(u + \alpha\, v) = \frac{d}{d\alpha}\bigg|_{\alpha=0} \int_0^1 \phi(u(t) + \alpha v(t)) dt = \int_0^1 \frac{1}{\omega(t)} \phi'(u(t)) v(t) \omega(t) dt.$$

Therefore

$$\nabla f(u) = \frac{1}{\omega} \phi'(u).$$

Note that $\nabla f(u)$ itself is a function so that the long version of this L^2 gradient is as follows

$$\nabla f(u)(t) = \frac{1}{\omega(t)} \phi'(u(t)).$$

Here, the weight function ω defines the metric in H.

(c) Now, let us assume that we are only interested in gradient representations in the subspace

$$H = \{u \in H^1([0,1]) \,|\, u(0) = 0, u(1) = 0\} \subset L^2([0,1])$$

endowed with the scalar product

$$(x, y) = \varepsilon \int_0^1 x(t)y(t)dt + \int_0^1 \dot{x}(t)\dot{y}(t)dt\,,\ \varepsilon > 0.$$

If we want to determine the gradient of the function f in example (b), where we express the derivative of f in terms of the scalar product of H, we have to find a function g such that for all $v \in L^2([0,1])$

$$\int_0^1 \phi'(u(t))v(t)dt = \varepsilon \int_0^1 g(t)v(t)dt + \int_0^1 \dot{g}(t)\dot{v}(t)dt = \int_0^1 (\varepsilon g(t) - \ddot{g}(t))v(t)dt,$$

which means we have to solve the second-order ODE

$$-\ddot{g}(t) = \phi'(u(t)) - \varepsilon g(t)$$

with boundary conditions $g(0) = 0$, $g(1) = 0$. Thus the H^1 gradient given by

$$\nabla f(u) = g$$

is a smoother function than $\phi'(u)$. Here, we see that the choice of the scalar product can have a smoothing effect on the gradient. The effect that we observe here is sometimes called Sobolev smoothing, which can be naturally explained as expressing the gradient in a particular scalar product. ■

Since now the definition of the gradient depends on the choice of the scalar product, we have to reassure ourselves of well-known properties of the gradient.

Theorem 2.7. $-\nabla f(u_0)$ *is orthogonal to the level set of* f *in* u_0 *and it defines the direction of steepest descent of* f *if we consider the scalar product used for the definition of the gradient.*

Proof. The level set is locally characterized by curves $\gamma : (-\varepsilon;\varepsilon) \ni t \mapsto \gamma(t) \in H$ with $f(\gamma(t)) \equiv f(u)$ for all t and $\gamma(0) = u_0$. Orthogonality pertains to the tangent space which is spanned by the velocity vectors $\dot{\gamma}(0)$ of all of these curves. Therefore, we obtain

$$0 = \left.\frac{d}{dt}\right|_{t=0} f(\gamma(t)) = (\nabla f(u_0), \dot{\gamma}(0)).$$

The steepest descent direction can be found as the solution of

$$\min_h f(u_0) + f'(u_0)h,$$
$$(h,h) = 1.$$

We abbreviate $g := \nabla f(u_0)/\|\nabla f(u_0)\|$. Each vector h in the sphere $S := \{h \,|\, (h,h) = 1\}$ can be uniquely represented as

$$h = \alpha g + \beta g^{\perp}, \quad \alpha^2 + \beta^2 = 1,$$

where $g^{\perp}$ is some vector in H with $(g, g^{\perp}) = 0$ and $(g^{\perp}, g^{\perp}) = 1$. Now

$$f'(u_0)h = (\nabla f(u_0), h) = \alpha \|\nabla f(u_0)\|$$

is minimized at the minimum possible value for α, which is $\alpha = -1$, which means also $\beta = 0$. This concludes the proof. □

We can also use the gradient, defined as a Riesz representation, in order to characterize solutions to optimization problems by necessary conditions.

Lemma 2.8 (necessary condition of first order). *If $\hat{u} = \arg\min_{u \in U(\hat{u})} f(u)$ for some open neighborhood $U(\hat{u})$ of $\hat{u}$, then $\nabla f(\hat{u}) = 0$.*

Proof. Let us consider the mapping $\varphi : (-\varepsilon; \varepsilon) \ni t \mapsto f(\hat{u} + th)$ for an arbitrary perturbation vector h. Then, $\varphi(0)$ is the obvious minimum of that function and from elementary calculus, we obtain

$$0 = \frac{d}{dt}\bigg|_{t=0} \varphi(t) = (\nabla f(\hat{u}), h) \quad \forall h \in U.$$

In particular, we conclude that

$$0 = (\nabla f(\hat{u}), \nabla f(\hat{u}))$$

and therefore

$$\nabla f(\hat{u}) = 0. \quad \square$$

Analogously to the definition of the gradient, we define the Hessian operator in the following way.

Definition 2.9. *We assume that U is a real Hilbert space. Let $f : U \to \mathbb{R}$ be twice differentiable in $u \in U$ with the second derivative as a symmetric bilinear form denoted by*

$$D^2 f(u)[h_1, h_2] := \frac{d}{dt_1}\bigg|_{t_1=0} \frac{d}{dt_2}\bigg|_{t_2=0} f(u + t_1 h_1 + t_2 h_2) \quad \forall h_1, h_2 \in U.$$

Then for each $h_1 \in U$ there exists a vector $v(h_1) \in U$ which is the Riesz representation of the linear form $D^2 f(u)[h_1, .]$ such that

$$D^2 f(u)[h_1, h_2] = (v(h_1), h_2) \quad \forall h_2 \in U.$$

We call the linear mapping

$$\begin{array}{rcl} \mathit{Hess}\, f(u) : U & \to & U \\ h_1 & \mapsto & v(h_1) \end{array}$$

the Hessian operator at u.

Obviously, the operator is linear and self-adjoint. Let us see how the Hessian operators in Example 2.6 look.

Example 2.10. We investigate the Hessian operators of Example 2.6(a)–(c).

(a) We study $H = \mathbb{R}^n$ endowed with the scalar product $(x, y) = x^\top Ay$, where A is a symmetric and positive definite matrix. We are interested in the Hessian operator of a twice differentiable function $f : H \to \mathbb{R}$. Its Hessian operator is

$$Hess\, f(u) = A^{-1}\begin{bmatrix} \frac{\partial^2 f(u)}{\partial u_1^2} & \cdots & \frac{\partial^2 f(u)}{\partial u_1 \partial u_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(u)}{\partial u_n \partial u_1} & \cdots & \frac{\partial^2 f(u)}{\partial u_n \partial u_n} \end{bmatrix},$$

since

$$D^2 f(u)[h_1, h_2] = h_1^\top \begin{bmatrix} \frac{\partial^2 f(u)}{\partial u_1^2} & \cdots & \frac{\partial^2 f(u)}{\partial u_1 \partial u_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(u)}{\partial u_n \partial u_1} & \cdots & \frac{\partial^2 f(u)}{\partial u_n \partial u_n} \end{bmatrix} h_2 = (Hess\, f(u)h_1, h_2).$$

(b) Now, we consider $H = L^2([0,1])$ endowed with the scalar product $(x, y) := \int_0^1 x(t)\, y(t)\, \omega(t) dt$ and the function $f : H \to \mathbb{R}$ defined by

$$f(u) := \int_0^1 \phi(u(t))dt.$$

As pointed out in Section 4.9.4 of [339], this integrated Nemytskii operator is not twice differentiable in $L^2([0,1])$, but rather in $L^\infty([0,1])$. However, the second derivative can be represented as a linear operator $L^2([0,1]) \to L^2([0,1])$, which can be obtained by a formal derivation and which we also call the Hessian. It is defined by

$$Hess\, f(u)h = \frac{1}{\omega}\phi''(u)h$$

because

$$D^2 f(u)[h_1, h_2] = \int_0^1 \frac{1}{\omega(t)}\phi''(u(t))h_1(t)h_2(t)dt = (Hess\, f(u)h_1, h_2).$$

More details on the necessity to work with two norms, when dealing with second derivatives, can be found in [339].

(c) We investigate the same objective as in (b) but in a different space

$$H = \{u \in H^1([0,1]) \,|\, u(0) = 0 u(1) = 0\}$$

endowed with the (Sobolev-smoothing) scalar product

$$(x, y) = \varepsilon \int_0^1 x(t)y(t)dt + \int_0^1 \dot{x}(t)\dot{y}(t)dt,\ \varepsilon > 0.$$

The Hessian in this scalar product is

$$Hess\, f(u) = \left(\varepsilon I - \frac{d^2}{dt^2}\right)^{-1} \phi''(u).$$

In many shape optimization problems, $\phi''(u)$ is spectrally equivalent to $(\varepsilon I - d^2/dt^2)$—c.f. Example 2.17. Then, we observe that the Hessian operator as defined in (c) is spectrally equivalent to the identity, which gives us an excellent performance of a gradient descent method. In those cases Sobolev smoothing is advantageous. ■

As a first consequence of these definitions, we can derive a Taylor series expansion.

Theorem 2.11. *We assume for the function $f : U \to \mathbb{R}$ defined on a Hilbert space H, where the subset $U \subset H$ is open and convex and f is twice differentiable, the property*

$$\|Hess\ f(u) - Hess\ f(v)\| \leq L\|u - v\| \quad \forall u, v \in U$$

with a constant $L < \infty$. Then, we achieve the estimation

$$\left| f(u) - f(v) + (\nabla f(u), v - u) + \frac{1}{2}(Hess\ f(u)(v - u), v - u) \right| \leq \frac{L}{6}\|u - v\|^3.$$

Proof. Let us consider the mapping $\varphi : [0,1] \ni t \mapsto f(\hat{u} + t(v - u)) \in H$. We note that for all differentiable functions and in particular for φ it yields

$$\int_0^1 \int_0^t \varphi''(s) - \varphi''(0) ds\, dt = \varphi(1) - \varphi(0) - \varphi'(0) - \frac{1}{2}\varphi''(0).$$

Since

$$\varphi(1) = f(v), \quad \varphi(0) = f(u), \quad \varphi'(0) = (\nabla f(u), v - u), \qquad \varphi''(0) = (Hess\ f(u)(v - u), v - u),$$

we observe

$$\left| f(u) - f(v) + (\nabla f(u), v - u) + \frac{1}{2}(Hess\ f(u)(v - u), v - u) \right| = \int_0^1 \int_0^t |\varphi''(s) - \varphi''(0)| ds\, dt$$

$$= \int_0^1 \int_0^t |((Hess\ f(u + s(v - u)) - Hess\ f(u))(v - u), v - u)|\, ds\, dt$$

$$\leq \int_0^1 \int_0^t \|Hess\ f(u + s(v - u)) - Hess\ f(v)\|\, \|u - v\|^2 ds\, dt$$

$$\leq \int_0^1 \int_0^t s L \|u - v\|^3 ds\, dt = \frac{L}{6}\|u - v\|^3. \quad \square$$

Now, we can exploit the Taylor expansion of Theorem 2.11 for necessary and sufficient optimality conditions.

Theorem 2.12. *Under the assumptions of Theorem* 2.11 *we obtain the following:*

(a) *If $\hat{u}$ is an optimal solution, then* Hess $f(\hat{u}) \geq 0$*, i.e.,* (Hess $f(\hat{u})h, h) \geq 0$ *for all $h \in H$.*

(b) *If $\hat{u}$ satisfies $\nabla f(\hat{u}) = 0$, and* Hess $f(\hat{u})$ *is coercive, i.e.,* (Hess $f(\hat{u})h, h) \geq c\|h\|^2$*, for all $h \in H$ and for some $c > 0$, then $\hat{u}$ is a local minimum, provided* Hess$_f(u)$ *satisfies a Lipschitz condition as in Theorem* 2.11.

Proof.

(a) Let us consider the mapping $\varphi : (-\varepsilon;\varepsilon) \ni t \mapsto f(\hat{u}+th)$ for an arbitrary perturbation vector h. Then, $\varphi(0)$ is the minimum, $\nabla f(\hat{u}) = 0$, and φ is convex, i.e., *Hess* $f(\hat{u}) \geq 0$.

(b) Choose a neighborhood U of $\hat{u}$ small enough so that $\|u-\hat{u}\| < 6c/L$ for all $u \in U$. Then for $h := u - \hat{u}$

$$f(u) - f(\hat{u}) \geq (\nabla f(\hat{u}),h) + (\mathit{Hess}\, f(\hat{u})h,h) - \frac{L}{6}\|h\|^3 \geq \left(c - \frac{L}{6}\|h\|\right)\|h\|^2 > 0. \quad \square$$

From the Taylor expansion, we observe at some point u close to $\hat{u}$

$$0 = \nabla f(\hat{u}) = \nabla f(u) + \mathit{Hess}\, f(u)(\hat{u}-u) + \mathcal{O}(\|\hat{u}-u\|^2)$$

and therefore

$$\hat{u} = u - \mathit{Hess}\, f(u)^{-1}\nabla f(u) + \mathcal{O}(\|\hat{u}-u\|^2).$$

That means that a Newton method of the form

$$u^{k+1} = u^k - \mathit{Hess}\, f(u^k)^{-1}\nabla f(u^k)$$

is locally quadratically convergent. Since often we can only afford a variant of the linearly converging steepest descent method

$$u^{k+1} = u^k - \nabla f(u^k),$$

in practice it would be highly profitable to choose the scalar product in H so that *Hess* $f(\hat{u})$ is close to identity, because that choice improves the convergence properties of the steepest descent method. At least, one should include the basic characteristics of the second derivative in the scalar product, e.g., the order of the operator.

Remark. Again, one should note that in general optimization problems the assumption of differentiability as in Theorem 2.11 is in question, even in simple cases. Also, the Hilbert space setting may not be applicable. Then, more refined techniques as in [339, 194, 207] for guaranteeing optimality conditions are to be applied which for many problems are still under investigation. Those issues are delicate, difficult, and important but beyond the scope of this book.

Now we come back to the constrained problem (2.1)–(2.3) and employ the reduced formulation

$$\min_u \hat{J}(u) := J(y(u),u).$$

The implicit function theorem gives us the expression

$$\hat{J}'(u) = \frac{\partial J(y,u)}{\partial u} - \frac{\partial J(y,u)}{\partial y}\left(\frac{\partial c(y,u)}{\partial y}\right)^{-1}\frac{\partial c(y,u)}{\partial u},$$

where y is fixed at $y = y(u)$. Note that the partial derivatives are meant only with respect to the explicit occurrence of the respective variable. The implicit dependency of $y = y(u)$ is taken care of by the formula itself. Since $\partial J(y,u)/\partial y$ is only one (dual) vector in contrast to the whole operator $\partial c(y,u)/\partial u$, it is computationally more convenient to compute this derivative in the order indicated by the brackets:

$$\hat{J}'(u) = \frac{\partial J(y,u)}{\partial u} - \left[\frac{\partial J(y,u)}{\partial y} \left(\frac{\partial c(y,u)}{\partial y} \right)^{-1} \right] \frac{\partial c(y,u)}{\partial u}. \tag{2.4}$$

This leads to the so-called adjoint approach to the gradient computation. As the name indicates, it is associated with adjoint operators, which will enable us to write the expression within the bracket in the form of a linear system of equations. In order to introduce this concept in a concise way, we make some simplifying assumptions on the vector spaces involved.

The PDE model $c(y,u) = 0$ is defined by a mapping

$$c : Y \times U \to Z,$$

where we assume that U is a Hilbert space. For the sake of clarity, we also assume for Y and Z that they are Hilbert spaces, where we denote by $(.,.)_U$, $(.,.)_Y$, and $(.,.)_Z$ the respective scalar products. The derivatives

$$\frac{\partial J(y,u)}{\partial y} \in Y^* \quad \text{and} \quad p^* := \frac{\partial J(y,u)}{\partial y} \left(\frac{\partial c(y,u)}{\partial y} \right)^{-1} \in Z^*$$

are elements of the dual spaces which can be pulled back to the primal space by the Riesz representation theorem. By use of adjoint operators (also denoted by *) corresponding to the scalar products of the Hilbert spaces, such that

$$\left(p, \left(\frac{\partial c(y,u)}{\partial u} \right) u \right)_Z = \left(\left(\frac{\partial c(y,u)}{\partial u} \right)^* p, u \right)_U \qquad \forall p \in Z, u \in U$$

and

$$\left(p, \left(\frac{\partial c(y,u)}{\partial y} \right) y \right)_Z = \left(\left(\frac{\partial c(y,u)}{\partial y} \right)^* p, y \right)_Y \qquad \forall p \in Z, y \in Y,$$

we can write the gradient in the form

$$\nabla \hat{J}(u) = \nabla_u J(y,u) + \left(\frac{\partial c(y,u)}{\partial u} \right)^* p, \tag{2.5}$$

where p solves

$$\left(\frac{\partial c(y,u)}{\partial y} \right)^* p = -\nabla_y J(y,u). \tag{2.6}$$

The engineering literature usually prefers the following derivation of this expression: Since $y(u)$ satisfies $c(y(u),u) = 0$, we can add it to the objective function in the following way, by using the scalar product in Z

$$\hat{J}(u) = J(y(u),u) = J(y(u),u) + (p, c(y(u),u))_Z$$

for any $p \in Z$. The gradient now is by chain rule

$$\begin{aligned}
\nabla \hat{J}(u) &= \nabla \left[J(y(u),u) + (p, c(y(u),u))_Z \right] \\
&= \nabla_u J + \left(\frac{\partial y}{\partial u}\right)^* \nabla_y J + \left(\frac{\partial c}{\partial u}\right)^* p + \left(\frac{\partial y}{\partial u}\right)^* \left(\frac{\partial c}{\partial y}\right)^* p \\
&= \nabla_u J + \left(\frac{\partial c}{\partial u}\right)^* p + \left(\frac{\partial y}{\partial u}\right)^* \left[\nabla_y J + \left(\frac{\partial c}{\partial y}\right)^* p \right].
\end{aligned}$$

If we define p by $\nabla_y J + (\partial c/\partial y)^* p = 0$, we can get rid of the expression in the brackets and arrive at the formula we have already obtained above

$$\nabla \hat{J}(u) = \nabla_y J + \left(\frac{\partial c}{\partial u}\right)^* p.$$

The expression of the gradient of $\hat{J}$ as derived above brings us to the necessary conditions of first order for equality constrained problems in the separability framework.

Theorem 2.13. *Define for the equality constrained problem* (2.1)–(2.3), *where $U_{ad} = U$, the Lagrangian as*

$$L(y,u,p) := J(y,u) + (p, c(y,u))_Z.$$

Then the necessary conditions for optimality are

$$\begin{aligned}
0 &= \nabla_y L(y,u,p) = \nabla_y J + \left(\frac{\partial c}{\partial y}\right)^* p, \\
0 &= \nabla_u L(y,u,p) = \nabla_u J + \left(\frac{\partial c}{\partial u}\right)^* p, \\
0 &= \nabla_p L(y,u,p) = c(y,u).
\end{aligned}$$

Proof. This follows immediately from the condition $\nabla \hat{J}(u) = 0$ and the derivation of a computational expression for $\nabla \hat{J}(u)$ above. □

Remark 2.14. *The choice of the sign of the adjoint variable in the definition of the Lagrangian is arbitrary and chosen only for aesthetic reasons. One can as well define*

$$L(y,u,p) := J(y,u) - (p, c(y,u))_Z$$

with obvious sign switches in all corresponding formulae.

Necessary and sufficient optimality conditions of second order follow from Theorem 2.12 by the observation that

$$\mathit{Hess}\, \hat{J}(u) = \begin{bmatrix} -(\partial c/\partial y)^{-1}(\partial c/\partial u) \\ I_U \end{bmatrix}^* \begin{bmatrix} \mathit{Hess}_{yy} L & \mathit{Hess}_{yu} L \\ \mathit{Hess}_{uy} L & \mathit{Hess}_{uu} L \end{bmatrix} \begin{bmatrix} -(\partial c/\partial y)^{-1}(\partial c/\partial u) \\ I_U \end{bmatrix}, \quad (2.7)$$

which can be seen from writing the gradient in the form

$$\nabla \hat{J}(u) = \begin{bmatrix} -(\partial c/\partial y)^{-1}(\partial c/\partial u) \\ I_U \end{bmatrix}^* \begin{pmatrix} \nabla_y L(y,u,p) \\ \nabla_u L(y,u,p) \end{pmatrix},$$

where p is defined by (2.6).

Theorem 2.15.

(a) *If* $(\hat{y},\hat{u})$ *is an optimal solution for the equality constrained problem* (2.1)–(2.3), *where* $U_{ad} = U$, *and* $\hat{p}$ *is defined by*

$$\left(\frac{\partial c(\hat{y},\hat{u})}{\partial y}\right)^* \hat{p} = \nabla_y J(\hat{y},\hat{u}),$$

then the operator

$$\mathit{Hess}\, L(\hat{y},\hat{u},\hat{p}) = \begin{bmatrix} \mathit{Hess}_{yy}\, L(\hat{y},\hat{u},\hat{p}) & \mathit{Hess}_{yu}\, L(\hat{y},\hat{u},\hat{p}) \\ \mathit{Hess}_{uy}\, L(\hat{y},\hat{u},\hat{p}) & \mathit{Hess}_{uu}\, L(\hat{y},\hat{u},\hat{p}) \end{bmatrix}$$

is positive semidefinite on the kernel of the linearized constraints, in the sense that

$$\left(\mathit{Hess}\, L(\hat{y},\hat{u},\hat{p})\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}\right)_{Y\times U} \geq 0$$

for all $(h_1,h_2) \in Y \times U$ *with*

$$\left[(\partial c/\partial y)\ (\partial c/\partial u)\right]\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = 0.$$

(b) *If* $(\hat{y},\hat{u},\hat{p})$ *satisfy the necessary conditions of first order as stated in Theorem* 2.13, *and* Hess L *is coercive on the kernel of the equality constraints, that is, there is a constant* $c < \infty$ *such that*

$$\left(\mathit{Hess}\, L(\hat{y},\hat{u},\hat{p})\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}\right)_{Y\times U} \geq c\left(\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}\right)_{Y\times U}$$

for all $(h_1,h_2) \in Y \times U$ *with*

$$\left[(\partial c/\partial y)\ (\partial c/\partial u)\right]\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = 0,$$

then $(\hat{y},\hat{u})$ *is a local minimum, provided* Hess L *satisfies a Lipschitz condition as in Theorem* 2.11.

Proof. The proposition is an immediate consequence of Theorem 2.12 and (2.7). □

The expression on the right-hand side of (2.7) is often called the reduced Hessian, because it is derived from a reduction (i.e., projection) of the full Hessian onto the nullspace of the constraints. We see immediately that the well-posedness of the optimization problem is equivalent to the coercivity of the reduced Hessian (if there are no further constraints on the control).

For illustration of the concepts above, we will derive specific necessary optimality conditions in a finite-dimensional case and for the basic elliptic optimal control problem of PDE constrained optimization.

Example 2.16. (a) We study the finite-dimensional optimization problem

$$\begin{gathered} \min J(y,u), \\ c(y,u) = 0, \end{gathered}$$

where $u \in \mathbb{R}^{n_u}$, $y \in \mathbb{R}^{n_y}$, $c(y,u) \in \mathbb{R}^{n_y}$, and the functions J, c are at least twice differentiable. We use in $\mathbb{R}^{n_u}$ and $\mathbb{R}^{n_y}$ the Euclidean scalar product $(x,y)_0 = x^\top y$. With that definition of the spaces, the necessary optimality conditions of first order can be written as

$$\begin{aligned}
0 &= \nabla_y L(y,u,p) = \left(\frac{\partial J}{\partial y}\right)^\top + \left(\frac{\partial c}{\partial y}\right)^\top p,\\
0 &= \nabla_u L(y,u,p) = \left(\frac{\partial J}{\partial u}\right)^\top + \left(\frac{\partial c}{\partial y}\right)^\top p,\\
0 &= \nabla_p L(y,u,p) = c(y,u),
\end{aligned}$$

where the Lagrangian is

$$L(y,u,p) := J(y,u) + p^\top c(y,u).$$

(b) In many cases, one studies PDE optimization problems in variational form, e.g., of the type

$$\begin{gathered}
\min J(y,u),\\
(y,p)_Y = b(u,p) \quad \forall p \in Y,
\end{gathered}$$

where $(.,.)_Y$ is a scalar product in Y and $b(.,.) : U \times Y \to \mathbb{R}$ is a bilinear form. We can use the engineering approach to arrive at a concise formulation of the necessary conditions. Since $b(u,p) - (y,p)_Y = 0$ for all $p \in Y$, we can add it to the objective

$$\hat{J}(u) = J(y(u),u) = J(y(u),u) - (y(u),p)_Y + b(u,p).$$

The necessary condition is

$$\begin{aligned}
0 &= \frac{d}{dt}\bigg|_{t=0} f(u+th)\\
&= \frac{d}{dt}\bigg|_{t=0} (J(y(u+th),u+th) - (y(u+th),p)_Y + b((u+th),p))\\
&= \left(\frac{\partial J}{\partial u}\right)h + \left(\frac{\partial J}{\partial y}\right)\left(\frac{\partial y}{\partial u}\right)h - \left(\left(\frac{\partial y}{\partial u}\right)h, p\right)_Y + b(h,p)\\
&= \left(\frac{\partial J}{\partial u}\right)h + \left[\left(\frac{\partial J}{\partial y}\right)\left(\frac{\partial y}{\partial u}\right)h - \left(\left(\frac{\partial y}{\partial u}\right)h, p\right)_Y\right] + b(h,p)\\
&= (\nabla_u J, h) + b(h,p) \quad \forall h \in U
\end{aligned}$$

if we can find $p \in Y$ such that

$$\left(\frac{\partial J}{\partial y}\right) y - (y,p)_Y = 0 \quad \forall y \in Y$$

and therefore $p = \nabla_y J$ in the Y-scalar product. Now, by use of the Lagrangian

$$L(y,u,p) := J(y,u) - (y,p)_Y + b(u,p)$$

we state the necessary conditions

$$\begin{array}{llllll} \text{(adjoint eq.)} & 0 = \nabla_y L & \Leftrightarrow & (h,p)_Y & = (\partial J/\partial y)h & \forall h \in Y, \\ \text{(design eq.)} & 0 = \nabla_u L & \Leftrightarrow & (\nabla_u J, w)_U & = -b(w,p) & \forall w \in U, \\ \text{(state eq.)} & 0 = \nabla_p L & \Leftrightarrow & (y,q)_Y & = b(u,q) & \forall q \in Y. \end{array}$$

(c) In this example, we derive necessary optimality conditions for a basic elliptic optimal control given by

$$\min J(y,u) = \frac{1}{2}\int_\Omega (y(x) - y_\Omega(x))^2 dx + \frac{\nu}{2}\int_\Omega u(x)^2 dx \tag{2.8a}$$

$$\begin{array}{rcl} -\triangle y(x) & = & u(x) \quad \forall x \in \Omega, \\ y(x) & = & 0 \quad \forall x \in \Gamma := \partial\Omega \end{array} \tag{2.8b}$$

with the computational domain $\Omega := [0,1]^2$ and $y_\Omega : \Omega \to \mathbb{R}$ a given function, and $\nu > 0$. It is well known that the solution y of the state equation is in $H^1(\Omega)$ if $u \in L^2(\Omega)$. The homogeneous Dirichlet boundary condition can be taken into account by reduction to the space $Y := H_0^1(\Omega) = \{h \in H^1(\Omega) \,|\, h|_\Gamma = 0\}$. We note that the bilinear forms

$$(y,p)_1 := \int_\Omega \nabla y(x)^\top \nabla p(x) dx \quad \text{and} \quad (y,p)_0 := \int_\Omega y(x)\, p(x) dx$$

are both scalar products on $H_0^1(\Omega)$, and $(u,w)_0$ is a scalar product in $U := L^2(\Omega)$. With these definitions, we can write the problem in the equivalent variational form

$$\min J(y,u) = \frac{1}{2}(y - y_\Omega, y - y_\Omega)_0 + \frac{\nu}{2}(u,u)_0,$$
$$(y,q)_1 = (u,q)_0 \quad \forall q \in H_0^1(\Omega).$$

Here, we can apply the results of example (b) above and arrive at the necessary conditions in variational form: determine $(y,u,p) \in H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega)$ which satisfy

$$\begin{array}{rcll} (h,p)_1 & = & (h, y - y_\Omega)_0 & \forall h \in Y, \\ \nu(u,w)_0 & = & -(p,w)_0 & \forall w \in U, \\ (y,q)_1 & = & (u,q)_0 & \forall q \in Y. \end{array}$$

We notice that the adjoint variational equation is equivalent to the boundary value problem

$$-\triangle p(x) = y(x) - y_\Omega(x) \quad \forall x \in \Omega,$$
$$p(x) = 0 \quad \forall x \in \Gamma$$

so that we can rewrite the necessary conditions in differential form

$$\begin{array}{rcllrcll} -\triangle p(x) & = & y(x) - y_\Omega(x) & \forall x \in \Omega, & p(x) & = & 0 & \forall x \in \Gamma, \\ \nu u(x) + p(x) & = & 0 & \forall x \in \Omega, & & & & \\ -\triangle y(x) & = & u(x) & \forall x \in \Omega, & y(x) & = & 0 & \forall x \in \Gamma. \end{array}$$

It might be illustrative to compute the reduced Hessian in this example. First, we have to compute

$$D^2L\left[\begin{pmatrix}\tilde{y}_1\\ \tilde{u}_1\end{pmatrix},\begin{pmatrix}\tilde{y}_2\\ \tilde{u}_2\end{pmatrix}\right]=\frac{d}{dt_1}\bigg|_{t_1=0}\frac{d}{dt_2}\bigg|_{t_2=0}L(y+t_1\tilde{y}_1+t_2\tilde{y}_2,u+t_1\tilde{u}_1+t_2\tilde{u}_2,p)$$

for perturbations which satisfy

$$\begin{aligned}(\tilde{y}_1,q)_1&=(\tilde{u}_1,q)_0\quad\forall q\in Y,\\ (\tilde{y}_2,q)_1&=(\tilde{u}_2,q)_0\quad\forall q\in Y.\end{aligned}$$

Obviously, we have

$$D^2L\left[\begin{pmatrix}\tilde{y}_1\\ \tilde{u}_1\end{pmatrix},\begin{pmatrix}\tilde{y}_2\\ \tilde{u}_2\end{pmatrix}\right]=(\tilde{y}_1,\tilde{y}_2)_0+\nu(\tilde{u}_1,\tilde{u}_2)_0.$$

Since

$$(\tilde{y},q)_1=(-\Delta\tilde{y},q)_0\quad\forall\tilde{y},q\in H_0^1(\Omega),$$

we observe that $\tilde{y}_1=(-\Delta)^{-1}\tilde{u}_1$ and $\tilde{y}_2=(-\Delta)^{-1}\tilde{u}_2$ in $H_0^1(\Omega)$, which means that

$$\begin{aligned}D^2L\left[\begin{pmatrix}\tilde{y}_1\\ \tilde{u}_1\end{pmatrix},\begin{pmatrix}\tilde{y}_2\\ \tilde{u}_2\end{pmatrix}\right]&=((-\Delta)^{-1}\tilde{u}_1,(-\Delta)^{-1}\tilde{u}_1)_0+\nu(\tilde{u}_1,\tilde{u}_2)_0\\ &=(((-\Delta)^{-2}+\nu\,I_U)\tilde{u}_1,\tilde{u}_2)_0\end{aligned}$$

and, therefore, the reduced Hessian in the $(.,.)_0$ scalar product is

$$Hess\,f(u)=(-\Delta)^{-2}+\nu\,I_U.$$

Note that $(-\Delta)^{-2}$ is a compact operator and, in particular, not coercive. System solution with these operators is ill-posed. The reduced Hessian, however, is a compact perturbation of the operator $\nu\,I_U$, and the complete reduced Hessian is obviously coercive for $\nu>0$. This structure of the reduced Hessian operator is typical in PDE-constrained optimization problems and shows that, in general, regularization (here $\nu>0$) is necessary for the sake of well-posedness. However, there are exceptions, like in shape optimization (see later). ■

2.3 The Formal Lagrangian Approach

In [339] a formal and straightforward approach to the derivation of necessary optimality conditions is given. We briefly sketch this approach and apply it to determine necessary conditions for slightly more complex problems as in the section above.

The key essence in deriving necessary conditions is the proper definition of a Lagrangian function. Different forms of this Lagrangian function are possible, as we have seen above. In [339] a canonical but formal Lagrangian approach is given, where, after formally deriving the necessary conditions, one has to decide whether the expressions derived make any sense at all. In optimization problems which are similar to the basic elliptic optimal control example above, this formal Lagrangian approach means using extensions of the $(.,.)_0$ scalar product in the image space of the differential operator and then playing with the resulting formulas. We use and demonstrate this approach in the following two examples.

Example 2.17. (a) For the sake of getting acquainted with the formal Lagrangian approach, we derive again the necessary conditions for the basic elliptic optimal control problem

$$\min J(y,u) = \frac{1}{2}\int_\Omega (y(x) - y_\Omega(x))^2 dx + \frac{\nu}{2}\int_\Omega u(x)^2 dx$$
$$\begin{aligned} -\triangle y(x) &= u(x) \quad \forall x \in \Omega, \\ y(x) &= 0 \quad \forall x \in \Gamma := \partial\Omega. \end{aligned}$$

According to the formal Lagrangian approach, we define an adjoint variable p in the domain Ω and an adjoint variable p_Γ on the boundary $\Gamma = \partial\Omega$ such that we get the formal expression

$$L(y,u,p,p_\Gamma) = J(y,u) + \int_\Omega p\,(-\triangle y - u)dx + \int_\Gamma p_\Gamma\, y ds.$$

We will have to recognize expressions in the $(.,.)_0$ with y alone, which means without a differential operator. Therefore, we first swap the differential operator over to the adjoint variable by the use of Green's formula

$$L(y,u,p,p_\Gamma) = J(y,u) + \int_\Omega (-\triangle p)\,y - p\,u dx + \int_\Gamma p\frac{\partial y}{\partial \vec{n}} - y\frac{\partial p}{\partial \vec{n}} ds + \int_\Gamma p_\Gamma\, y ds.$$

The adjoint boundary value problem is derived from the expression for perturbations $\tilde{y}$

$$\begin{aligned} 0 &= \frac{d}{dt}\bigg|_{t=0} L(y + t\,\tilde{y}, u, p, p_\Gamma) \\ &= \underbrace{(y - y_\Omega, \tilde{y})_0 + ((-\triangle p), \tilde{y})_0}_{(1)} + \underbrace{\int_\Gamma p_\Gamma\, \tilde{y} ds - \int_\Gamma \tilde{y}\frac{\partial p}{\partial \vec{n}} ds}_{(2)} + \underbrace{\int_\Gamma p\frac{\partial \tilde{y}}{\partial \vec{n}} ds}_{(3)}. \end{aligned}$$

This equation should hold, in particular, for all perturbations $\tilde{y} \in C^\infty(\Omega)$, that is, all infinitely often differential functions with compact support. Focusing on these perturbations eliminates the expressions (2) and (3) for the moment, which results in expression (1) in

$$y - y_\Omega + (-\triangle p) = 0.$$

This result eliminates expression (1) permanently. Next, we focus on perturbations $\tilde{y}$ which satisfy the Dirichlet boundary condition ($\tilde{y}|_\Gamma = 0$) but have arbitrary Neumann values $\partial\tilde{y}/\partial\vec{n}$. This eliminates expression (2) temporarily and gives us from expression (3)

$$p(x) = 0 \quad \forall x \in \Gamma.$$

This eliminates expression (3) permanently. Now, we have already derived the adjoint boundary value problem. For the sake of completeness, we also try to give an expression for p_Γ. We do this by leaving the perturbations $\tilde{y}$ completely free. Thus we conclude from (2) that

$$p_\Gamma(x) - \frac{\partial \tilde{y}}{\partial \vec{n}} = 0 \quad \forall x \in \Gamma.$$

The design equation is found out analogously by stating

$$0 = \frac{d}{dt}\bigg|_{t=0} L(y, u + t\,\tilde{u}, p, p_\Gamma) = \nu\,(u,\tilde{u})_0 - (p,\tilde{u})_0$$

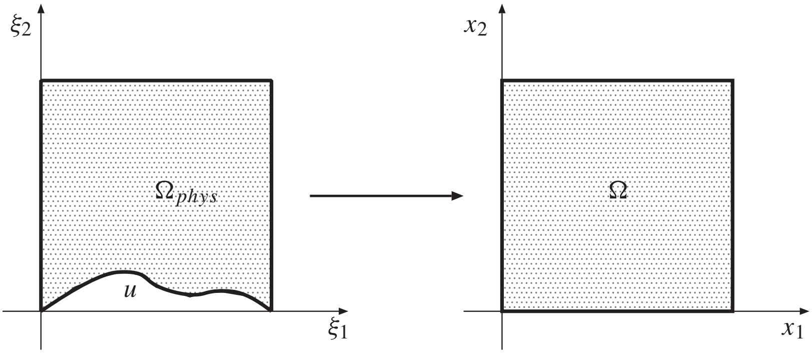

Figure 2.1. *Method of mapping in shape optimization.*

for arbitrary perturbations $\tilde{u} \in L^2(\Omega)$, which gives us the already known design equation

$$\nu u - p = 0.$$

(b) Now, we use the formal Lagrangian approach to derive necessary conditions for optimality of a typical model problem for aerodynamic shape optimization methods. The left-hand side of Figure 2.1 is imagined as the physical domain, which is thought as part of the free air stream along a geometric detail of an airplane which is aligned with the ξ_1-axis. The free stream is supposed to go from left to right and is modeled by the potential flow $-\triangle y = 0$ with a Neumann condition at the shape boundary. The shape boundary is supposed to be described by the function $u(\xi_1)$. In order to simplify the problem, we map the physical domain onto the computational domain on the right-hand side of Figure 2.1 by stretching it appropriately in the vertical direction, such that we can write

$$(x_1, x_2) = \left(\xi_1, \frac{\xi_2 - u(\xi_1)}{1 - u(\xi_1)}\right)$$

and $y(\xi)$ is thought of as $y(x(\xi))$. That means that the Neumann condition in the physical plane at the shape boundary $\{(\xi_1, \xi_2) \,|\, \xi_2 = u(\xi_1)\}$

$$0 = \frac{\partial y}{\partial \vec{n}} = \frac{\partial y}{\partial \xi_1}\dot{u} - \frac{\partial y}{\partial \xi_2}, \quad \dot{u} := \frac{\partial u}{\partial \xi_1}$$

is mapped to the boundary condition

$$0 = \dot{u}\frac{\partial y}{\partial \xi_1} - \frac{1 + \dot{u}^2}{1 - u^2}\frac{\partial y}{\partial x_2}.$$

For the purpose of simplification, we assume that the shape u is almost a straight line $u(x_1) \approx 0$, such that squared expressions like u^2 and $\dot{u}^2$ can be neglected. Furthermore, the tangential velocity $\frac{\partial y}{\partial x_1}$ can be assumed to be constant (e.g., 1). This results in the boundary condition

$$\frac{\partial y}{\partial x_2} = \frac{\partial u}{\partial x_1}.$$

The goal of the problem is to match a certain pressure profile at the shape boundary. By the gas law, pressure is inversely correlated to the tangential velocity, which means we aim at matching a certain tangential velocity profile at the shape boundary. The complete problem formulation in the computational domain is then

$$\min J(y,u) = \int_{\Gamma_1} \left(\frac{\partial y(x_1,x_2)}{\partial x_1} - v(x_1) \right)^2 dx_1, \tag{2.9}$$

$$\begin{aligned} -\Delta y(x) &= 0 \quad \forall x \in \Omega, \\ \frac{\partial y}{\partial \vec{n}}(x) &= \frac{\partial u}{\partial x_1}(x) \quad \forall x \in \Gamma_1, \\ y(x) &= x_1 \quad \forall x \in \Gamma_2 := \partial\Omega \backslash \Gamma_1, \end{aligned} \tag{2.10}$$

where $\Omega = (0,1) \times (0,1)$, $\Gamma_1 := [0,1] \times \{0\}$, and $v : [0,1] \to \mathbb{R}$ is a given function. We build up the formal Lagrangian as follows

$$\begin{aligned} L(y,u,p,p_1,p_2) &= J(y,u) + \int_\Omega p(-\triangle y)dx + \int_{\Gamma_1} p_1 \left(\frac{\partial y}{\partial \vec{n}} - \frac{\partial u}{\partial \xi_1} \right) ds + \int_{\Gamma_2} p_2 (y - x_1) ds \\ &= \int_{\Gamma_1} \left(\frac{\partial y(x_1,x_2)}{\partial x_1} - v(x_1) \right)^2 dx_1 + \int_\Omega (-\triangle p) y dx \\ &\quad + \int_{\Gamma_1 \cup \Gamma_2} p \frac{\partial y}{\partial \vec{n}} - y \frac{\partial p}{\partial \vec{n}} ds + \int_{\Gamma_1} p_1 \left(\frac{\partial y}{\partial \vec{n}} - \frac{\partial u}{\partial \xi_1} \right) ds + \int_{\Gamma_2} p_2 (y - x_1) ds. \end{aligned}$$

The adjoint boundary value problem is again derived from the expression for perturbations $\tilde{y}$

$$\begin{aligned} 0 &= \frac{d}{dt}\bigg|_{t=0} L(y + t\,\tilde{y}, u, p, p_1, p_2) \\ &= \int_{\Gamma_1} \left(\frac{\partial y}{\partial x_1} - v(x_1) \right) \frac{\partial \tilde{y}}{\partial x_1} dx_1 + \int_\Omega (-\triangle p)\, \tilde{y} dx \\ &\quad + \int_{\Gamma_1 \cup \Gamma_2} p \frac{\partial \tilde{y}}{\partial \vec{n}} - \tilde{y} \frac{\partial p}{\partial \vec{n}} ds + \int_{\Gamma_1} p_1 \frac{\partial \tilde{y}}{\partial \vec{n}} ds + \int_{\Gamma_2} p_2\, \tilde{y} ds \\ &= \left[\left(\frac{\partial y}{\partial x_1} - v(x_1) \right) \tilde{y} \right]_0^1 - \int_{\Gamma_1} \tilde{y} \frac{\partial}{\partial x_1} \left(\frac{\partial y}{\partial x_1} - v(x_1) \right) dx_1 + \int_\Omega (-\triangle p)\, \tilde{y} dx \\ &\quad + \int_{\Gamma_1 \cup \Gamma_2} p \frac{\partial \tilde{y}}{\partial \vec{n}} - \tilde{y} \frac{\partial p}{\partial \vec{n}} ds + \int_{\Gamma_1} p_1 \frac{\partial \tilde{y}}{\partial \vec{n}} ds + \int_{\Gamma_2} p_2\, \tilde{y} ds. \end{aligned}$$

(1) We focus on perturbations $\tilde{y} \in C^\infty(\Omega)$, which results in

$$-\triangle p = 0.$$

(2) Next we focus on perturbations with fixed Dirichlet value 0 and variable Neumann values, which gives

$$p|_{\Gamma_1} = -p_1|_{\Gamma_1}, \qquad p|_{\Gamma_2} = 0.$$

(3) Finally, we perturb in all respects (with the exception of the corner points), which gives

$$\frac{\partial p}{\partial \vec{n}}\bigg|_{\Gamma_1} = \frac{\partial}{\partial x_1} \left(\frac{\partial y}{\partial x_1} - v \right), \qquad p_2 = \frac{\partial p}{\partial \vec{n}}\bigg|_{\Gamma_2}.$$

The adjoint boundary value problem now reads completely as

$$\begin{aligned} -\triangle p &= 0 \quad \text{in } \Omega, \\ \left.\frac{\partial p}{\partial \vec{n}}\right|_{\Gamma_1} &= \frac{\partial}{\partial x_1}\left(\frac{\partial y}{\partial x_1} - v\right), \\ p|_{\Gamma_2} &= 0. \end{aligned}$$

For the design condition, we have to perturb the Lagrangian also in the control direction, which means

$$\begin{aligned} 0 = \left.\frac{d}{dt}\right|_{t=0} L(y, u + t\,\tilde{u}, p, p_1, p_2) &= -\int_{\Gamma_1} p_1 \frac{\partial \tilde{u}}{\partial x_1} ds \\ = -\left[p_1 \tilde{u}\right]_0^1 + \int_{\Gamma_1} \frac{\partial p_1}{\partial x_1} \tilde{u} dx_1 &= \int_{\Gamma_1} \frac{\partial p_1}{\partial x_1} \tilde{u} dx_1. \end{aligned}$$

Free variation of $\tilde{u}$ gives the design equation on Γ_1,

$$0 = \frac{\partial p_1}{\partial x_1} = -\frac{\partial p}{\partial x_1},$$

which also gives the gradient of the objective with respect to u in the L^2 scalar product on Γ_1. In order to decide about well-posedness of the problem, we might be interested in the reduced Hessian of $\hat{J}(u) = J(y(u,u))$. Here it seems impossible to perform analytical manipulations of the optimality conditions in order to come up with an explicit expression. We note that

$$\nabla \hat{J}(u) = -\left.\frac{\partial p(y(u))}{\partial x_1}\right|_{\Gamma_1} = -\left.\frac{\partial}{\partial x_1}\right|_{\Gamma_1}\left(\left.\frac{\partial y(u)}{\partial x_1}\right|_{\Gamma_1} - v\right),$$

where the mapping $u \mapsto y$ is described by the forward problem and the mapping $y \mapsto p$ by the adjoint problem.

All these mappings are affine linear. Therefore, a Fourier analysis of the homogeneous parts gives us the highest differential order of the Hessian operator. If this order is zero or greater, we can conclude well-posedness of the shape optimization problem. For the Fourier analysis, we neglect the boundary Γ_2 and stretch the boundary Γ_1 infinitely wide. We start with an arbitrary Fourier mode for u, as follows

$$u = e^{i\omega_u x_1} \Rightarrow \frac{\partial u}{\partial x_1} = i\omega_u e^{i\omega_u x_1}.$$

The solution of the forward problem is assumed to be of the form

$$y(x_1, x_2) = r\, e^{i\omega_1 x_1 + i\omega_2 x_2} \Rightarrow \left.\frac{\partial y}{\partial \vec{n}}\right|_{\Gamma_1} = i\omega_2 r\, e^{i\omega_1 x_1} \quad \text{and} \quad \left.\frac{\partial^2 y}{\partial x_1^2}\right|_{\Gamma_1} = -\omega_1^2 r\, e^{i\omega_1 x_1}.$$

From the boundary condition of the flow problem, we obtain

$$i\omega_u e^{i\omega_u x_1} = i\omega_2 r\, e^{i\omega_1 x_1}, \qquad x_1 \in \mathbb{R}.$$

For $x_1 = 0$ this implies $\omega_2 = \omega_u / r$ and thus also $\omega_1 = \omega_u$. The differential equation $-\triangle y = 0$ gives

$$0 = (\omega_1^2 + \omega_2^2) r\, e^{i\omega_1 x_1 + i\omega_2 x_2} \Leftrightarrow \frac{\omega_u^2}{r^2} + \omega_u^2 = 0,$$

which means $r = \pm i$. The adjoint solution is assumed to be of similar form

$$p(x_1,x_2) = s\, e^{i\phi_1 x_1 + i\phi_2 x_2} \Rightarrow \left.\frac{\partial p}{\partial \vec{n}}\right|_{\Gamma_1} = i\phi_2 s\, e^{i\phi_1 x_1}.$$

The Neumann boundary condition of the adjoint problem now gives

$$i\phi_2 s\, e^{i\phi_1 x_1} = -\omega_1^2 r\, e^{i\omega_1 x_1}, \qquad x_1 \in \mathbb{R};$$

with arguments similar to those above, we conclude that

$$\begin{aligned} \phi_2 &= -\frac{\omega_1^2 r}{is} = \mp\frac{\omega_u^2}{s}, \\ \phi_1 &= \omega_1 = \omega_u. \end{aligned}$$

From the adjoint differential equation, we obtain

$$0 = \phi_1^2 + \phi_2^2 \Rightarrow s = \pm i\omega_u.$$

Thus

$$\left.\frac{\partial p}{\partial x_1}\right|_{\Gamma_1}(x_1) = \mp\omega_u^2 u.$$

From the necessary conditions of optimality, the Hessian has to be positive (semi-) definite and therefore we have for large ω_u

$$\left.\frac{\partial p}{\partial x_1}\right|_{\Gamma_1}(x_1) = \omega_u^2 u.$$

From that, we conclude by Fourier analysis that the highest order of the Hessian operator is 2, such that the Hessian can be approximated by

$$Hess\, \hat{J} \approx -\frac{\partial^2}{\partial x_1^2},$$

which means that the problem is essentially well posed and needs no further regularization. A very similar structure of the Hessian arises in Section 7.2.1, where we will discuss realistic aerodynamic shape optimization problems. ■

2.4 Control Constraints

In the previous sections, we have omitted inequality constraints. State constraints are a highly complicated issue theoretically. Therefore, we consider only the practical treatment of them in the applications section. Control constraints, however, are easily integrated in the context of this chapter. Let us again look at the generic problem (2.1)–(2.3). Now we assume that we allow only $u \in U_{ad}$, where $U_{ad} \subset U$ is a convex subset of the Hilbert space U. We consider the following

$$\min f(u), \tag{2.11}$$

$$u \in U_{ad}. \tag{2.12}$$

Theorem 2.18. *If $U_{ad} \subset U$ is a convex subset of the Hilbert space U and $f : U_{ad} \to \mathbb{R}$ is a differentiable function, then the solution $\hat{u}$ of the optimization problem* (2.11)–(2.12) *satisfies the variational inequality*

$$(\nabla f(\hat{u}), u - \hat{u}) \geq 0 \quad \forall u \in U_{ad}. \tag{2.13}$$

Proof. The proof follows the lines of the proof of Lemma 2.20 in [339]. We choose an arbitrary $u \in U_{ad}$ and consider the convex combination

$$u_s := \hat{u} + s\,(u - \hat{u}), \quad s \in [0,1].$$

We observe $u_s \in U_{ad}$, because of the convexity of U_{ad}. The optimality of $\hat{u}$ gives us $f(u_s) \geq f(\hat{u})$, from which we conclude that

$$\frac{1}{s}(f(u_s) - f(\hat{u})) = \frac{1}{s}(f(\hat{u} + s\,(u - \hat{u})) - f(\hat{u})) \geq 0$$

and therefore

$$0 \leq \lim_{s \to 0} \frac{1}{s}(f(\hat{u} + s\,(u - \hat{u})) - f(\hat{u})) = f'(\hat{u})(u - \hat{u}) = (\nabla f(\hat{u}), u - \hat{u}). \quad \square$$

Corollary 2.19. *If f in Theorem* 2.18 *is a convex function, then the condition* (2.13) *is sufficient for optimality of $\hat{u}$.*

Proof. f is assumed to be convex. Therefore

$$f(t\,u + (1-t)\,\hat{u}) \leq t\,f(u) + (1-t)\,f(\hat{u}) \quad \forall t \in [0,1].$$

Differentiating with respect to t, that is, applying $\frac{d}{dt}\big|_{t=0}$ on both sides, gives

$$0 \leq f'(\hat{u})(u - \hat{u}) \leq f(u) - f(\hat{u})$$

for arbitrary $u \in U_{ad}$, which means that $\hat{u}$ is optimal. $\square$

We study consequences of Theorem 2.18 in two frequently appearing instances: box constraints for control functions in L^2, and finite-dimensional controls and constraints. We first treat box constraints in L^2.

Theorem 2.20. *We consider the space $U = L^2(\Omega)$ for some bounded and open computational domain $\Omega \subset \mathbb{R}^n$ and the subset $U_{ad} \subset U$, defined in the following way*

$$U_{ad} := \left\{ u \in U \,|\, \underline{u}(x) \leq u(x) \leq \bar{u}(x) \textit{ for almost all } x \in \Omega \right\},$$

where $\underline{u}, \bar{u} \in U$ with $\underline{u}(x) \leq \bar{u}(x)$ for almost all $x \in \Omega$ are functions defining lower and upper bounds for the control u. For $\lambda, \mu \in U$, we define the Lagrangian

$$L(u,\lambda,\mu) := f(u) + (\lambda, \underline{u} - u)_U + (\mu, u - \bar{u})_U.$$

If $\hat{u}$ is the solution to the problem

$$\min f(u),$$
$$u \in U_{ad},$$

then there exist $\lambda, \mu \in U$ *with* $\lambda(x), \mu(x) \geq 0$*, for almost all* $x \in \Omega$*, such that*

$$\nabla_u L(\hat{u}) = 0,$$
$$\lambda(x)(\hat{u}(x) - \underline{u}(x)) = 0,\, \mu(x)(\bar{u}(x) - \hat{u}(x)) = 0 \quad \textit{for almost all } x \in \Omega.$$

Proof. The convexity of the set U_{ad} is obvious. Therefore, we can apply Theorem 2.18. First, we define

$$\lambda(x) := (\nabla f(\hat{u}(x)))^+, \qquad \mu(x) := (\nabla f(\hat{u}(x)))^- \quad \forall x \in \Omega,$$

where we define for $r \in \mathbb{R}$

$$r^+ := (|r| + r)/2, \qquad r^- := (|r| - r)/2.$$

From that definition of λ, μ we obtain immediately $\lambda, \mu \in U$ and

$$\nabla_u L(\hat{u}) = \nabla_u f(\hat{u}) - \lambda + \mu = 0.$$

From Theorem 2.18, we now obtain

$$(\lambda - \mu, u - \hat{u})_U \geq 0 \quad \forall u \in U_{ad}.$$

We define

$$\Omega^+ := \{x \in \Omega \,|\, \nabla f(\hat{u}(x)) > 0\}.$$

Since $L^2(\Omega) \subset L_1(\Omega)$, this is a measurable subset of Ω, and we have $\mu|_{\Omega^+} = 0$. Now, we show that

$$\int_M \lambda(x)(\hat{u}(x) - \underline{u}(x))dx = 0$$

for all subsets $M \subset \Omega$ with nonzero measure: if we would find a measurable subset $M \subset \Omega$ with $\int_M \lambda(x)(\hat{u}(x) - \underline{u}(x))dx > 0$, then we can assume without loss of generality that $M \subset \Omega^+$, because the integrand is zero on $M \setminus (\Omega^+ \cap M)$. We define the function $\tilde{u}$ by

$$\tilde{u}(x) = \begin{cases} \underline{u}(x), & x \in M, \\ \hat{u}(x), & x \in \Omega \backslash M. \end{cases}$$

Then, $\tilde{u} \in U_{ad}$ and

$$0 \leq \int_M \lambda(x)(\hat{u}(x) - \underline{u}(x))dx = \int_M (\lambda(x) - \mu(x))(\hat{u}(x) - \underline{u}(x))dx = -(\lambda - \mu, \tilde{u} - \hat{u})_U \leq 0$$

and therefore $\lambda(x)(\hat{u}(x) - \underline{u}(x)) = 0$ for almost all $x \in \Omega$ and analogously $\mu(x)(\bar{u}(x) - \hat{u}(x)) = 0$. □

Remark. As we know from Section 2.2, the existence of a solution in U_{ad} can be guaranteed only if f is convex, which is the case in many model problems which are typically linear-quadratic. In practice, convexity can be assured only from the Taylor series locally in the vicinity of the optimal solution.

More general inequality constraints require the introduction of certain constraint qualifications, which are beyond the scope of this book. Instead, we consider general inequality conditions for $u \in \mathbb{R}^n$ together with the Euclidean scalar product.

Theorem 2.21. *For $u \in \mathbb{R}^n$, we consider the constrained optimization problem* (2.11), (2.12)*, where U_{ad} is defined by*

$$U_{ad} = \left\{ u \in \mathbb{R}^n \mid h(u) \le 0 \right\},$$

where the inequality is meant componentwise. The mapping $h : \mathbb{R}^n \to \mathbb{R}^m$ is supposed to be twice differentiable, and its components h_i are supposed to be convex such that

$$U_{ad} = \bigcap_{i=1}^{m} \left\{ u \in \mathbb{R}^n \mid h_i(u) \le 0 \right\}$$

is a convex set. We assume that $\hat{u}$ is the solution of problem (2.11), (2.12)*. At $\hat{u}$, we define the active set*

$$\mathcal{A} := \{ i \in \{1,\dots,m\} \,|\, h_i(\hat{u}) = 0 \}.$$

Furthermore, we assume that the functions h_i satisfy the linear independence constraint qualification (LICQ)

$$\{\nabla h_i(\hat{u}) \,|\, i \in \mathcal{A}\} \text{ is a linear independent set of vectors.}$$

Then, there is an adjoint vector $\lambda \in \mathbb{R}^m$ with components $\lambda_i \ge 0$, for all $i = 1,\dots,m$, such that the Lagrangian function

$$L(u,\lambda) := f(u) + \lambda^\top h(u)$$

satisfies

$$\nabla_u L(\hat{u}) = 0, \tag{2.14}$$

$$\lambda_i \, h_i(\hat{u}) = 0 \quad \forall i = 1,\dots,m. \tag{2.15}$$

Proof. The solution $\hat{u}$ is by definition also the solution to the problem

$$\begin{gathered} \min f(u), \\ h_i(u) = 0 \quad \forall i \in \mathcal{A}. \end{gathered}$$

Because of the LICQ, we can separate the variables locally such that

$$u = \begin{pmatrix} u_d \\ u_f \end{pmatrix}, \quad u_d \in \mathbb{R}^{\#\mathcal{A}}, \quad \frac{\partial h_{\mathcal{A}}}{\partial u_d} \text{ nonsingular,}$$

where $h_{\mathcal{A}} = (h_i)_{i \in \mathcal{A}}$. We obtain the locally equivalent problem

$$\begin{gathered} \min f(u_d, u_f), \\ h_{\mathcal{A}}(u_d, u_f) = 0, \quad \frac{\partial h_{\mathcal{A}}}{\partial u_d} \text{ nonsingular.} \end{gathered}$$

We know this problem structure already from Section 2.2 and conclude from Theorem 2.13 that there is $\mu \in \mathbb{R}^{\#\mathcal{A}}$ with

$$\begin{aligned} 0 &= \nabla_{u_d} f + \left(\frac{\partial h_{\mathcal{A}}}{\partial u_d} \right)^\top \mu, \\ 0 &= \nabla_{u_f} f + \left(\frac{\partial h_{\mathcal{A}}}{\partial u_f} \right)^\top \mu; \end{aligned}$$

if we define $\lambda \in \mathbb{R}^m$ by

$$\lambda_i = \begin{cases} \mu_i, & i \in \mathcal{A}, \\ 0, & i \notin \mathcal{A}, \end{cases}$$

we can write this equivalently as

$$0 = \nabla_u f + \left(\frac{\partial h}{\partial u}\right)^\top \lambda = \nabla_u L.$$

The complementarity conditions (2.15) are now obvious: either $i \in \mathcal{A}$ and then $h_i(\hat{u}) = 0$, or $i \notin \mathcal{A}$ and then $\lambda_i = 0$. From a Taylor series expansion, we obtain now

$$h_i(u) = h_i(\hat{u}) + \nabla h_i(\hat{u})^\top (u - \hat{u}) + \mathcal{O}(\|u - \hat{u}\|^2) \quad \forall i = 1, \dots, m,$$

from which we conclude that

$$\nabla h_i(\hat{u})^\top (u - \hat{u}) \leq \eta \quad \forall i \in \mathcal{A},$$

where $\eta = \mathcal{O}(\|u - \hat{u}\|^2)$. We choose $u^i \in U_{ad}$ close to $\hat{u}$ such that

$$\nabla h_i(\hat{u})^\top (u^i - \hat{u}) < 0, \text{ and } \nabla h_j(\hat{u})^\top (u^i - \hat{u}) = 0 \quad \forall i \in \mathcal{A}.$$

Now, we conclude from condition (2.13) that

$$0 \leq \nabla f(\hat{u})^\top (u^i - \hat{u}) = -\sum_{j=1}^m \lambda_j \nabla h_j(\hat{u})^\top (u^i - \hat{u}) = -\lambda_i \nabla h_i(\hat{u})^\top (u^i - \hat{u})$$

and therefore $\lambda_i \geq 0$ for all $i = 1, \dots, m$. □

We observe that the necessary conditions for inequalities are very similar to the necessary conditions for equality constraints: The differences lie in the predefined sign of the adjoint variables ($\lambda \geq 0$) and in the complementarity conditions. These complementarity conditions are the starting point for active set methods and for interior point methods for the treatment of inequalities.