**Heedong Goh**

## Lecture Notes for

# Inverse Problems

Department of Civil, Urban and Environmental Engineering
Seoul National University
2026

*"Once men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."*

Frank Herbert, *Dune* (1965)

# Contents


**1 Preliminaries** **1**
- 1.1 Index notation . . . 1
- 1.2 Inner product and induced norm . . . 1
- 1.3 Function space . . . 2
  - 1.3.1 Banach space . . . 2
  - 1.3.2 Hilbert space . . . 3
- 1.4 Differentiation . . . 5
- 1.5 Integration by parts . . . 6
- 1.6 Implicit function theorem . . . 7
- 1.7 Weak form . . . 9

**2 Calculus of Variations** **11**
- 2.1 Functionals . . . 11
- 2.2 Functional derivatives . . . 11
- 2.3 Necessary and sufficient conditions for local extrema . . . 13
- 2.4 Euler equation . . . 15
- 2.5 Canonical form of the Euler equations . . . 18
- 2.6 Noether's theorem . . . 21
- 2.7 General variation of a functional . . . 24
- 2.8 Functionals depending on higher-order derivatives . . . 26
- 2.9 Lagrange multipliers . . . 28
  - 2.9.1 Constrained finite-dimensional minimization problem . . . 28
  - 2.9.2 Constrained variational problems . . . 31

**3 Source Inversion** **35**
- 3.1 Inverse problem . . . 35
- 3.2 Deblurring problem . . . 36
- 3.3 SVD-based approach . . . 40
- 3.4 Tikhonov regularization . . . 41
- 3.5 Regularization parameter choice . . . 42
  - 3.5.1 Discrepancy (Morozov) principle . . . 43
  - 3.5.2 L-curve criterion . . . 44
- 3.6 Image deblurring in two dimension . . . 45

**4 Parameter Field Inversion** **49**
- 4.1 Synopsis . . . 49
- 4.2 Elliptic problem . . . 50
  - 4.2.1 Problem definition . . . 50

4.2.2 Lagrangian and necessary condition . . . . . 51
4.2.3 Gradient-based optimization and the adjoint method . . . 52
4.2.4 Step length criteria . . . . . 54
4.2.5 Newton method . . . . . 55
4.3 Hyperbolic problem . . . . . 58
4.4 Eigenvalue problem . . . . . 60

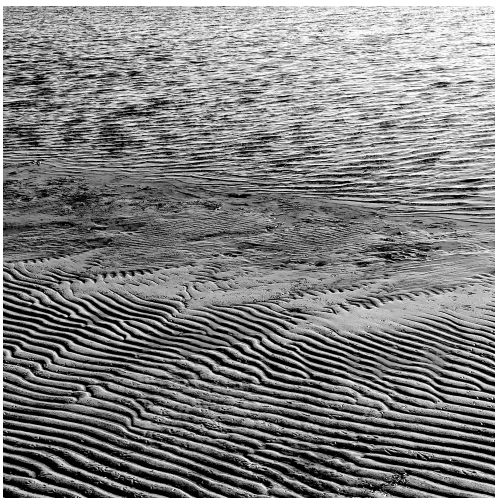
**5 Refinements** **63**
5.1 Partial observations and data filtering . . . . . 63
5.1.1 Partial observations . . . . . 63
5.1.2 Data filtering . . . . . 64
5.2 Discontinuous parameter fields . . . . . 65
5.3 Bounded and discrete parameter fields . . . . . 68
5.3.1 Logarithmic barrier functional . . . . . 68
5.3.2 Nonlinear reparametrization . . . . . 69

# Preface

This monograph is written for the course *Introduction to Computational Method for Inverse Problems* offered in the Department of Civil, Urban and Environmental Engineering at Seoul National University.

Inverse problems arise in nondestructive testing, imaging, system identification, parameter estimation, and optimal design; the mathematical focus of these notes is PDE-constrained optimization, in which model parameters or design variables are recovered by minimizing an objective functional subject to a governing partial differential equation. This framework connects directly to modern machine learning, where neural network training is a large-scale instance of parameter estimation.

In preparing these notes, I borrowed many parts from the following books and lecture notes: [Ghattas, 2015, Stone and Goldbart, 2009, Gelfand et al., 2000, Petra and Stadler, 2011, Vogel, 2002]. These lecture notes are a working manuscript, subject to ongoing refinement and enhancement; I welcome and greatly appreciate any reports regarding errors, typos, or any other inaccuracies found within the notes.



Wave-induced waves at Assateague Island, MD, USA

The author gratefully acknowledges the support of the Korean Ministry of Education and the Seoul Metropolitan Government's RISE (Regional Innovation System & Education) project in the preparation of these lecture notes.

# Chapter 1

# Preliminaries

## 1.1 Index notation

We follow *Einstein notation* and use regular font for both scalars and tensors, whose types are to be inferred from context. For example, a vector $v$ expressed in terms of a basis $g_i$ is written as

$$v = v^i g_i. \tag{1.1}$$

In most cases, we make no distinction between *vectors* and *covectors*, as a Cartesian basis is assumed unless stated otherwise. Accordingly, we also write $v = v^i g_i = v_i g^i = v_i g_i$. Similarly, the *inner product* between two geometric vectors is given by

$$(a, b) \coloneqq a \cdot \overline{b} = a^i \overline{b_i} = a_i \overline{b_i}, \tag{1.2}$$

where $\cdot$ denotes *(single) contraction* and $\overline{(\,)}$ indicates complex conjugation of the subtended quantity.

## 1.2 Inner product and induced norm

A *set* equipped with inner product is called an *inner product space*. An inner product is a *bilinear form* (or a *sesquilinear form* for complex vectors) such that

$$(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{F}. \tag{1.3}$$

Here $\mathbb{F}$ is either a real number $\mathbb{R}$ or a complex number $\mathbb{C}$. Thus, an inner product takes two vectors in $\mathcal{V}$ and returns a scalar.

In addition, an inner product must satisfy the following properties [Oden and Demkowicz, 2017]:

- *Linearity* with respect to the first argument

$$(\alpha u + \beta v, w) = \alpha\,(u, w) + \beta\,(v, w) \quad \forall \alpha, \beta \in \mathbb{F},\ \forall u, v, w \in \mathcal{V}. \tag{1.4}$$

- *Conjugate symmetry*

$$(u, v) = \overline{(v, u)} \quad \forall u, v \in \mathcal{V}. \tag{1.5}$$

- *Positive definiteness*

$$(u, u) > 0 \quad \forall u \neq 0,\ u \in \mathcal{V}. \tag{1.6}$$

Note that an inner product is *anti-linear* with respect to the second argument, i.e.,

$$(u, \alpha v) = \overline{\alpha}\,(u, v) \quad \forall \alpha \in \mathbb{F},\ \forall u, v \in \mathcal{V}. \tag{1.7}$$

*Orthogonality* is defined such that

$$(u, v) = 0. \tag{1.8}$$

Any inner product space is equipped with an *induced norm* such that

$$\|u\| \coloneqq \sqrt{(u, u)}. \tag{1.9}$$

**Theorem 1.2.1 (Cauchy–Schwarz Inequality)** *Let $(\cdot, \cdot)$ be an inner product on a vector space $\mathcal{V}$. Then, $\forall u, v \in \mathcal{V}$,*

$$|(u, v)| \leq \|u\|\,\|v\|\,. \tag{1.10}$$

*Equality holds when $u$ and $v$ are linearly dependent.*

*Triangle inequality* can be proved from the Cauchy-Schwarz inequality as

$$\begin{aligned} \|u + v\|^2 &= (u + v, u + v) \\ &= (u, u) + (u, v) + (v, u) + (v, v) \\ &= (u, u) + 2\mathrm{Re}\,\{(u, v)\} + (v, v) \\ &\leq \|u\|^2 + 2\,|(u, v)| + \|v\|^2 \\ &\leq \|u\|^2 + 2\,\|u\|\,\|v\| + \|v\|^2 \\ &= (\|u\| + \|v\|)^2\,. \end{aligned} \tag{1.11}$$

Thus, we have $\|u + v\| \leq \|u\| + \|v\|$.

## 1.3 Function space

### 1.3.1 Banach space

- *Linear space*: A set $X$ is a linear space over a field $\mathbb{F}$ (typically $\mathbb{R}$ or $\mathbb{C}$) if its elements $x, y, \ldots \in X$ satisfy the following axioms under addition and scalar multiplication by $\alpha, \beta, \ldots \in \mathbb{F}$:

  1. $x + y = y + x$ (commutativity of addition);
  2. $(x + y) + z = x + (y + z)$ (associativity of addition);
  3. There exists a zero element $0 \in X$ such that $x + 0 = x$ for all $x \in X$;
  4. For each $x \in X$, there exists an additive inverse $-x \in X$ such that $x + (-x) = 0$;
  5. $1 \cdot x = x$ for all $x \in X$ (scalar identity);

6. $\alpha(\beta x) = (\alpha\beta)x$ (associativity of scalar multiplication);
7. $(\alpha + \beta)x = \alpha x + \beta x$ (distributivity over scalar addition); *and*
8. $\alpha(x + y) = \alpha x + \alpha y$ (distributivity over vector addition).

- *Normed linear space*: A linear space $X$ is said to be *normed* if each element $x \in X$ is assigned a nonnegative real number $\|x\|$, called the *norm* of $x$, such that

  1. $\|x\| = 0$ if and only if $x = 0$ (positive definiteness);
  2. $\|\alpha x\| = |\alpha|\|x\|$ for all $\alpha \in \mathbb{F}$ and $x \in X$ (absolute homogeneity); *and*
  3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

- *Banach space*: A normed linear space $X$ is said to be a *Banach space* if it is complete with respect to its norm. That is, every Cauchy sequence $(x_n)$ in $X$ converges to a limit $x \in X$ such that $\lim_{n\to\infty} \|x_n - x\| = 0$.

A Cauchy sequence is a sequence whose elements become arbitrarily close to each other as the sequence progresses. Namely, a sequence $(x_n)$ is Cauchy if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $\|x_m - x_n\| < \varepsilon$ for all $m, n \geq N$.

- $C^0$ *space*: A function $u : \Omega \to \mathbb{R}$ belongs to $C^0(\Omega)$ if it is continuous on $\Omega$, i.e.,

$$\lim_{x\to x_0} u(x) = u(x_0), \quad \forall x_0 \in \Omega. \tag{1.12}$$

  When equipped with the supremum norm,

$$\|u\|_{C^0(\Omega)} = \sup_{x\in\Omega} |u(x)|, \tag{1.13}$$

  the space $C^0(\Omega)$ is a Banach space.

- $C^1$ *space*: A function $u : \Omega \to \mathbb{R}$ belongs to $C^1(\Omega)$ if the function and all its first-order partial derivatives are continuous on $\Omega$. The space $C^1(\Omega)$ is a Banach space when equipped with the $C^1$ norm,

$$\|u\|_{C^1(\Omega)} = \|u\|_{C^0(\Omega)} + \|\operatorname{grad} u\|_{C^0(\Omega)}. \tag{1.14}$$

  Here, $\|\operatorname{grad} u\|_{C^0(\Omega)} = \sup_{x\in\Omega} |\operatorname{grad} u(x)|$. Note that the gradient should be understood in the strong (classical) sense.

- $C^k$ *space*: A function $u : \Omega \to \mathbb{R}$ belongs to $C^k(\Omega)$ if all partial derivatives up to order $k$ are continuous.

- $C^\infty$ *space*: A function $u : \Omega \to \mathbb{R}$ belongs to $C^\infty(\Omega)$ if all partial derivatives of all orders are continuous. Such functions are called *smooth* or *infinitely differentiable*.

### 1.3.2 Hilbert space

- *Inner product space*: A linear space $X$ is said to be an *inner product space* if each pair of elements $x, y \in X$ is assigned a scalar $(x, y) \in \mathbb{F}$, called the *inner product* of $x$ and $y$, such that

  1. $(x, x) \geq 0$ and $(x, x) = 0$ if and only if $x = 0$ (positive definiteness);
  2. $(x, y) = \overline{(y, x)}$ for all $x, y \in X$ (conjugate symmetry);

3. $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ for all $\alpha, \beta \in \mathbb{F}$ and $x, y, z \in X$ (linearity in the first argument).

The inner product induces a norm via $\|x\| = \sqrt{(x, x)}$.

- *Hilbert space*: An inner product space $X$ is said to be a *Hilbert space* if it is complete with respect to the norm induced by its inner product, $\|x\| = \sqrt{(x, x)}$. Equivalently, a Hilbert space is a Banach space whose norm derives from an inner product.

- $L^2$ *space*: For a domain $\Omega \subset \mathbb{R}^n$, the space $L^2(\Omega)$ is defined as

$$L^2(\Omega) = \left\{ u : \Omega \to \mathbb{R} \text{ (or } \mathbb{C}) : \int_\Omega |u|^2 \, dx < \infty \right\}. \tag{1.15}$$

The space $L^2(\Omega)$ is a Hilbert space equipped with the inner product

$$(u, v)_{L^2(\Omega)} = \int_\Omega u \cdot \overline{v} \, dx \tag{1.16}$$

and induced norm

$$\|u\|_{L^2(\Omega)} = \sqrt{(u, u)_{L^2(\Omega)}} = \left( \int_\Omega |u|^2 \, dx \right)^{1/2}. \tag{1.17}$$

- *Sobolev space of the first order*:

$$H^1(\Omega) \coloneqq \left\{ u \in L^2(\Omega) \ : \ \operatorname{grad} u \in L^2(\Omega)^3 \right\}. \tag{1.18}$$

The corresponding inner product is

$$\begin{aligned} (u, v)_{H^1(\Omega)} &= (u, v) + (\operatorname{grad} u, \operatorname{grad} v) \\ &= \int_\Omega u\overline{v} + \int_\Omega \operatorname{grad} u \cdot \overline{\operatorname{grad} v} \end{aligned} \tag{1.19}$$

and the induced norm is

$$\|u\|_{H^1(\Omega)} = \sqrt{(u, u)_{H^1(\Omega)}}. \tag{1.20}$$

Note that the gradient must be understood in the weak (distributional) sense.

- *Sobolev space of the k-th order*:

$$H^k(\Omega) \coloneqq \left\{ u \in L^2(\Omega) \ : \ D^\alpha u \in L^2(\Omega), \ \forall |\alpha| \leq k \right\}. \tag{1.21}$$

Here, $D^\alpha = \partial^{|\alpha|} / \left( \partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n} \right)$ denotes partial derivative of order $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$ in the weak sense.

- $H$ (div) space:

$$H(\operatorname{div}, \Omega) = \left\{ u \in L^2(\Omega)^3 \ : \ \operatorname{div} u \in L^2(\Omega) \right\}. \tag{1.22}$$

- $H$ (curl) space:

$$H(\operatorname{curl}, \Omega) = \left\{ u \in L^2(\Omega)^3 \ : \ \operatorname{curl} u \in L^2(\Omega)^3 \right\}. \tag{1.23}$$

**Exercise 1.3.1** *Let $f(x) = C$ for some constant $C \neq 0$. Determine whether $f$ belongs to each of the following function spaces on $\Omega = \mathbb{R}$:*

1. *$C^0(\mathbb{R})$ and $C^\infty(\mathbb{R})$,*
2. *$L^2(\mathbb{R})$,*
3. *$H^k(\mathbb{R})$ for $k \geq 1$.*

## 1.4 Differentiation

The *directional derivative* of a map $u : X \to \mathbb{R}$ at a point $x \in X$ in the direction $h \in X$ is defined as

$$d_h u(x) \coloneqq \lim_{\epsilon \to 0} \frac{u(x + \epsilon h) - u(x)}{\epsilon}. \tag{1.24}$$

This limit probes $u$ along the single direction $h$ and need not be linear in $h$ in general.

If the map $h \to d_h u(x)$ is linear and bounded, then $d_h u(x)$ is called the *Gâteaux differential* of $u$ at $x$ in the direction $h$, and we write

$$d_h u(x) = du(x)[h]. \tag{1.25}$$

**Theorem 1.4.1 (Riesz Representation Theorem)** *Let $H$ be a Hilbert space with inner product $(\cdot, \cdot)_H$. For every continuous linear functional $\varphi : H \to \mathbb{R}$, there exists a unique element $f \in H$ such that*

$$\varphi(x) = (f,\, x)_H, \quad \forall x \in H. \tag{1.26}$$

*Moreover, $\|\varphi\|_{H^*} = \|f\|_H$. The element $f$ is called the* Riesz representation *of $\varphi$.*

The Riesz Representation Theorem guarantees the existence of a unique element $g \in X$ such that

$$du(x)[h] = (g,\, h)_X, \quad \forall h \in X. \tag{1.27}$$

This element $g$ is called the *gradient* of $u$ at $x$, denoted $\nabla u(x)$ or $\operatorname{grad} u(x)$. It is the unique Riesz representation of the Gâteaux derivative $du(x)[h]$ in $X$, and depends on the choice of inner product on $X$.

**Definition 1.4.1 (Weak Derivative)** *Let $u \in L^2(\Omega)$. A function $v \in L^2(\Omega)$ is called the* weak derivative *of $u$ with respect to $x_i$, denoted $v = \partial u / \partial x_i$, if*

$$(v,\, \phi) = -\left(u,\, \frac{\partial \phi}{\partial x_i}\right), \quad \forall \phi \in C_0^\infty(\Omega). \tag{1.30}$$

*More generally, for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ with $|\alpha| = \alpha_1 + \cdots + \alpha_n$, the weak derivative $D^\alpha u$ is the function $v \in L^2(\Omega)$ satisfying*

$$(v,\, \phi) = (-1)^{|\alpha|}\,(u,\, D^\alpha \phi), \quad \forall \phi \in C_0^\infty(\Omega), \tag{1.31}$$

*where $(\,\cdot\,,\,\cdot\,) \coloneqq (\,\cdot\,,\,\cdot\,)_{L^2(\Omega)}$.*

The function $u$ is said to be *Fréchet differentiable* if

$$u(x + h) - u(x) - d_h u(x) = o(\|u\|), \tag{1.28}$$

i.e.,

$$\lim_{\|h\| \to 0} \frac{|u(x+h) - u(x) - d_h u(x)|}{\|h\|} = 0. \tag{1.29}$$

The functional $d_h u(x)$ is called the *Fréchet differential* of $u$ at $x$.

When $u \in C^1(\Omega)$, the weak derivative coincides with the classical derivative, so the weak derivative is an extension of classical differentiation to functions that may not be pointwise differentiable. This extension is fundamental to Sobolev space theory: the Sobolev space $H^k(\Omega)$ consists precisely of those functions in $L^2(\Omega)$ whose weak derivatives up to order $k$ also belong to $L^2(\Omega)$.

**Example 1.4.1 (Weak Derivative of the Absolute Value Function)** *Consider $u(x) = |x|$ on $\mathbb{R}$. For any $\phi \in C_0^\infty(\mathbb{R})$,*

$$\begin{aligned}\int_{-\infty}^{\infty} |x|\,\phi'(x)\,dx &= \int_{-\infty}^{0} (-x)\,\phi'(x)\,dx + \int_{0}^{\infty} x\,\phi'(x)\,dx \\ &= [-x\phi]_{-\infty}^{0} - \int_{-\infty}^{0} -\phi(x)\,dx + [x\phi]_{0}^{\infty} - \int_{0}^{\infty} \phi(x)\,dx \\ &= -\int_{-\infty}^{\infty} \operatorname{sgn}(x)\,\phi(x)\,dx, \end{aligned} \tag{1.32}$$

*where*

$$\operatorname{sgn}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases} \tag{1.33}$$

*The boundary terms vanish because $\phi$ has compact support and the contributions at $x = 0$ cancel. Thus, the weak derivative is $v(x) = \operatorname{sgn}(x)$.*

**Exercise 1.4.1 (Heaviside Step Function)** *Consider the Heaviside function on $\mathbb{R}$:*

$$H(x) = \begin{cases} 0, & x \le 0, \\ 1, & x > 0. \end{cases} \tag{1.34}$$

*Find the weak derivative of the Heaviside step function.*

## 1.5 Integration by parts

Integration by parts in one dimension reads

$$\int_0^1 \frac{du}{dx} v = [uv]_0^1 - \int_0^1 u \frac{dv}{dx}. \tag{1.35}$$

The multi-dimensional generalization of the above equation reads

$$\int_\Omega \frac{\partial u}{\partial x^i} v = \int_{\partial\Omega} uvn_i - \int_\Omega u \frac{\partial v}{\partial x^i}. \tag{1.36}$$

In the above, $\Omega \in \mathbb{R}^N$ is the domain in $N$ dimension and its boundary is denoted by $\partial\Omega$, and $n_i$ is the $i$-th component of the outward normal vector on the boundary.

Integration by parts for divergence and curl operators are (respectively)

$$\int_\Omega (\operatorname{div} u)\, v = \int_{\partial\Omega} (u \cdot n)\, v - \int_\Omega u \cdot \operatorname{grad} v \quad \text{and} \tag{1.37}$$

$$\int_\Omega \operatorname{curl} E \cdot F = \int_{\partial\Omega} (n \times E) \cdot F + \int_\Omega E \cdot \operatorname{curl} F. \tag{1.38}$$

In the above, $u$, $v$, $E$, and $F$ are vector-valued functions and $n$ is unit outward normal vector. Integration by parts involves with the divergence of a *two tensor*, $A \in \mathbb{R}^{N\times N}$, reads

$$\int_\Omega (\operatorname{div} A) \cdot v = \int_{\partial\Omega} (An) \cdot v - \int_\Omega A : \operatorname{grad} v, \tag{1.39}$$

where : denotes *double contraction* such that $A : B = A_{ij}B^{ij}$.

## 1.6 Implicit function theorem

The Implicit Function Theorem is a fundamental result in nonlinear analysis with wide-ranging applications, including the derivation of first-order optimality conditions via Lagrange multipliers and sensitivity analysis of PDE-constrained optimization problems. We state the theorem first in the finite-dimensional setting and then in the general Banach space setting.

**Theorem 1.6.1 (Implicit Function Theorem)** *Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$ be an open set and let $g : U \to \mathbb{R}^m$ be a $C^1$ function. Suppose there exists a point $(x_0, y_0) \in U$ such that*

$$g(x_0, y_0) = 0. \tag{1.40}$$

*If the partial Jacobian of $g$ with respect to $y$, evaluated at $(x_0, y_0)$, is non-singular, i.e.,*

$$\det\left(\frac{\partial g_i}{\partial y_j}(x_0, y_0)\right) \neq 0, \quad i, j = 1, 2, \ldots, m, \tag{1.41}$$

*then there exist open neighborhoods $B \subset \mathbb{R}^n$ of $x_0$ and $V \subset \mathbb{R}^m$ of $y_0$, and a unique $C^1$ function $\varphi : B \to V$ such that $\varphi(x_0) = y_0$ and*

$$g(x, \varphi(x)) = 0, \quad \forall x \in B. \tag{1.42}$$

*Moreover, the Jacobian of $\varphi$ is given by*

$$D\varphi(x) = -\left[D_y g(x, \varphi(x))\right]^{-1} D_x g(x, \varphi(x)), \quad \forall x \in B, \tag{1.43}$$

*or equivalently in component form,*

$$\frac{\partial \varphi_i}{\partial x_j}(x) = -\sum_{k=1}^{m} F^{-1}_{ik}(x, \varphi(x)) \frac{\partial g_k}{\partial x_j}(x, \varphi(x)), \quad \forall x \in B, \tag{1.44}$$

*where $F_{ij} \coloneqq \partial g_i / \partial y_j$.*

$D_y g(x_0, y_0)$ denotes the Fréchet derivative of $g$ with respect to $y$, evaluated at $(x_0, y_0)$.

**Example 1.6.1 (Unit Circle)** *Consider the unit circle defined implicitly by*

$$g(x, y) = x^2 + y^2 - 1 = 0. \tag{1.45}$$

*This is a classic example because the circle cannot be expressed as a single global function $y = \varphi(x)$ over all of $\mathbb{R}$, yet the implicit function theorem (IFT) guarantees a local function exists near any point where the hypotheses are satisfied.*

Here $g : \mathbb{R}^2 \to \mathbb{R}$, so $n = 1$ and $m = 1$ in Theorem 1.6.1. The partial derivative with respect to $y$ is

$$\frac{\partial g}{\partial y}(x, y) = 2y. \tag{1.46}$$

The IFT applies at any point $(x_0, y_0)$ on the circle where

$$\frac{\partial g}{\partial y}(x_0, y_0) = 2y_0 \neq 0, \tag{1.47}$$

i.e., at any point except $(\pm 1, 0)$.

Take the point $(x_0, y_0) = (0, 1)$. Since $\partial g/\partial y = 2 \neq 0$ there, the IFT guarantees the existence of a neighborhood $B \subset \mathbb{R}$ of $x_0 = 0$ and a unique $C^1$ function $\varphi : B \to \mathbb{R}$ such that $\varphi(0) = 1$ and

$$g(x, \varphi(x)) = x^2 + \varphi(x)^2 - 1 = 0, \quad \forall x \in B. \tag{1.48}$$

Explicitly, this function is the upper semicircle

$$\varphi(x) = \sqrt{1 - x^2}, \quad x \in (-1, 1). \tag{1.49}$$

Similarly, near $(x_0, y_0) = (0, -1)$, the IFT gives the lower semicircle $\varphi(x) = -\sqrt{1 - x^2}$.

The IFT formula gives the derivative of $\varphi$:

$$\varphi'(x) = -\left[\frac{\partial g}{\partial y}(x, \varphi(x))\right]^{-1} \frac{\partial g}{\partial x}(x, \varphi(x)) = -\frac{2x}{2\varphi(x)} = -\frac{x}{\varphi(x)}, \tag{1.50}$$

which for $\varphi(x) = \sqrt{1 - x^2}$ gives

$$\varphi'(x) = -\frac{x}{\sqrt{1 - x^2}}. \tag{1.51}$$

At $x_0 = 0$, we obtain $\varphi'(0) = 0$, reflecting the horizontal tangent at the top of the circle.

At the points $(\pm 1, 0)$, the condition $\partial g/\partial y \neq 0$ fails and the IFT does not apply. Indeed, the circle has a vertical tangent at these points and cannot be expressed as a $C^1$ function of $x$ near $x = \pm 1$.

The finite-dimensional theorem extends to Banach spaces, with the non-singularity of the partial Jacobian replaced by the bounded invertibility of the partial Fréchet derivative. This generalization is essential for applications to PDE-constrained optimization, where the state variable lives in an infinite-dimensional function space such as a Sobolev space.

**Theorem 1.6.2 (Implicit Function Theorem in Banach Spaces)** *Let $X$, $Y$, $Z$ be Banach spaces, $U \subset X \times Y$ open, and $g : U \to Z$ a $C^1$ map. Suppose $(x_0, y_0) \in U$ satisfies $g(x_0, y_0) = 0$ and the partial Fréchet derivative*

$$A := D_y g(x_0, y_0) : Y \to Z \tag{1.52}$$

*is a Banach space isomorphism, i.e., $A$ is a bounded bijection with bounded inverse $A^{-1} : Z \to Y$. Then there exist open neighborhoods $B \subset X$ of $x_0$ and $V \subset Y$ of $y_0$, and a unique $C^1$ map $\varphi : B \to V$ such that $\varphi(x_0) = y_0$,*

An isomorphism between two Banach spaces is a bounded bijective linear map whose inverse is also bounded.

$$g(x, \varphi(x)) = 0, \quad \forall x \in B, \tag{1.53}$$

*and*

$$D\varphi(x) = -\left[D_y g(x, \varphi(x))\right]^{-1} D_x g(x, \varphi(x)), \quad \forall x \in B. \tag{1.54}$$

## 1.7 Weak form

Consider a model Poisson's equation:

$$(\mathrm{S}) \begin{cases} \dfrac{d^2u}{dx^2} + f = 0 \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} . \tag{1.55}$$

The left-hand side of the governing equation vanishes for all $x \in (0,1)$ when $u$ is a true solution. Thus, the governing equation gives a pointwise requirement of the solution, and such formulation is called a *strong form*.

Alternatively, a *weak form* focuses on the definition of the 0 on the right-hand side; like the left-hand side of the equation, 0 is a function, or a vector. Specifically, the function 0 returns a number 0 when inner-producted with any other function. Thus, we can equivalently write

$$\int v \left( \frac{d^2u}{dx^2} + f \right) = 0, \quad \forall v, \tag{1.56}$$

which is identified as a weak form. Here, $u$ is called a *trial function* and the arbitrary function $v$ is called a *test function*.

The above weak form becomes more useful when we take an integration by parts:

$$\begin{aligned} 0 &= \int v \left( \frac{d^2u}{dx^2} + f \right) \\ &= \left[ v \frac{du}{dx} \right]_0^1 - \int \frac{dv}{dx} \frac{du}{dx} + \int vf. \end{aligned} \tag{1.57}$$

Here, the boundary term vanishes when the test function satisfies homogeneous Dirichlet boundary conditions, i.e., $v(0) = v(1) = 0$. Then, we have

$$(\mathrm{W}) \begin{cases} u \in \mathcal{U} \\ b(u,v) = l(v), \quad \forall v \in \mathcal{V} \end{cases} . \tag{1.58}$$

In the above, the bilinear and linear forms are defined by

$$b(u,v) = \int \frac{dv}{dx}\frac{du}{dx} \quad \text{and} \tag{1.59}$$

$$l(v) = \int vf. \tag{1.60}$$

$u$ and $v$ may be elements of different function spaces $\mathcal{U}$ and $\mathcal{V}$. However, in the given model problem, we derived a symmetric, or Hermitian, bilinear form by integration by parts, which allowed to use the same function space for both trial and test functions:

$$\mathcal{U} = \mathcal{V} = \left\{ w \in H^1(0,1) \ : \ w(0) = w(1) = 0 \right\}. \tag{1.61}$$

The space $H^1(0,1)$ is chosen to ensure the bilinear form is integrable, i.e.,

$$\left| \left( \frac{du}{dx}, \frac{dv}{dx} \right) \right| < \infty, \tag{1.62}$$

where the Cauchy-Schwarz inequality gives

$$\left| \left( \frac{du}{dx}, \frac{dv}{dx} \right) \right| \leq \left\| \frac{du}{dx} \right\| \left\| \frac{dv}{dx} \right\|. \tag{1.63}$$

# Chapter 2

# Calculus of Variations

## 2.1 Functionals

A *functional* is a map $J : V \to \mathbb{R}$, where $V$ is a function space. That is, a functional is a function that takes functions as its arguments. If $J$ satisfies $J[\alpha u + \beta v] = \alpha J[u] + \beta J[v]$ for all $u, v \in V$ and scalars $\alpha, \beta$, it is called a *linear functional*.

For example, consider the total potential energy functional $J[w]$ of an Euler-Bernoulli beam:

$$J[w] = \underbrace{\frac{1}{2}\int_0^L EI\left(\frac{d^2w}{dx^2}\right)^2 dx}_{=U[w]} - \underbrace{\int_0^L qw dx}_{=V[w]}, \tag{2.1}$$

where $w$ is the deflection, $q$ is the distributed load, $EI$ is the flexural rigidity, and $L$ is the beam length. Here, both the strain energy $U[w]$ and the work done by an external load $V[w]$ are functionals, while $V[w]$ is linear with respect to $w$.

**Definition 2.1.1 (Continuity)** *The functional $J : V \to \mathbb{R}$ is said to be* continuous *at the point $\hat{u} \in V$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$|J[u] - J[\hat{u}]| < \epsilon, \tag{2.2}$$

*provided that*

$$\|u - \hat{u}\| < \delta. \tag{2.3}$$

## 2.2 Functional derivatives

Let $J[u]$ be a functional defined on some normed linear space, where its *increment* is defined as

$$\Delta J[h] = J[u + h] - J[u]. \tag{2.4}$$

Note that for a fixed $u$, $\Delta J$ is a (generally nonlinear) functional of $h$. Suppose that the increment can be decomposed as

$$\Delta J[h] = \varphi[h] + \epsilon \|h\|, \tag{2.5}$$

where $\varphi[h]$ is a linear functional. If $\epsilon \to 0$ as $\|h\| \to 0$, the functional $J[u]$ is said to be *differentiable*. The linear functional $\varphi[h]$ is called the *variation* (or *differential*) of the functional $J[u]$ and is denoted by $d_h J$ or $\delta_h J$. To make the base point explicit, we may also write $dJ(u)[h]$ or $dJ[u;h]$; similarly, $\delta J(u)[h]$ or $\delta J[u;h]$.

In practice, the variation can be computed as a Gâteaux differential:

$$d_v J[u] = \lim_{\epsilon \to 0} \frac{J[u+\epsilon v] - J[u]}{\epsilon} = \frac{d}{d\epsilon} J[u+\epsilon v]\bigg|_{\epsilon=0}. \tag{2.6}$$

**Example 2.2.1 (Uniqueness of the Differential)** *Prove that a differentiable functional has a unique differential.*

Suppose a differentiable functional $J[u]$ admits two differentials such that

$$\Delta J[h] = \varphi_1[h] + \epsilon_1 \|h\| \quad \text{and} \tag{2.7}$$
$$\Delta J[h] = \varphi_2[h] + \epsilon_2 \|h\|, \tag{2.8}$$

where $\epsilon_1, \epsilon_2 \to 0$ as $\|h\| \to 0$. Then, subtracting the two equations, we have

$$\varphi_1[h] - \varphi_2[h] = (\epsilon_2 - \epsilon_1)\|h\|. \tag{2.9}$$

Set $h = h_0/n$ for a fixed nonzero $h_0$. By linearity, we have

$$\frac{\varphi_1[h_0] - \varphi_2[h_0]}{n} = (\epsilon_2 - \epsilon_1)\frac{\|h_0\|}{n}, \tag{2.10}$$

which simplifies to

$$\varphi_1[h_0] - \varphi_2[h_0] = (\epsilon_2 - \epsilon_1)\|h_0\|. \tag{2.11}$$

Taking $n \to \infty$ implies $\|h\| \to 0$ so that $\epsilon_1, \epsilon_2 \to 0$. Thus, the right-hand-side of the equation vanishes, giving $\varphi_1[h_0] = \varphi_2[h_0]$. Since $h_0$ is arbitrary, we conclude $\varphi_1 = \varphi_2$.

The gradient $g$ arising from the Riesz representation theorem, i.e.,

$$d_v J[u] = (g, v) \tag{2.12}$$

is called the *functional derivative* and is denoted by $\nabla J$, $\operatorname{grad} J$ or $\frac{\delta J}{\delta u}$.

A functional $J\colon V \to \mathbb{R}$ is said to be *twice differentiable* at $u \in V$ if its increment can be written in the form

$$\Delta J[h] = \varphi_1[h] + \varphi_2[h] + \epsilon\|h\|^2, \tag{2.13}$$

where $\varphi_1[h]$ is a linear functional in $h$ (i.e., the first variation), $\varphi_2[h]$ is a quadratic functional in $h$, and $\epsilon \to 0$ as $\|h\| \to 0$. The quadratic functional $\varphi_2$ is called the *second variation* (or *second differential*) and is denoted by

$$\delta^2 J[u;h] \quad \text{or equivalently} \quad d^2 J(u)[h]. \tag{2.14}$$

**Example 2.2.2 (First and Second Variations)** *Consider the functional* $J\colon C^1([a,b]) \to \mathbb{R}$ *defined by*

$$J[u] = \int_a^b F(x,u,u')dx, \tag{2.15}$$

*where* $F(x,u,u') = \frac{1}{2}(u')^2 + \frac{1}{2}u^2$ *is sufficiently smooth. Find its first and second variations.*

We compute the increment $\Delta J[h] \coloneqq J[u+h] - J[u]$ by expanding $F(x,u+h,u'+h')$ in a Taylor series about $(u,u')$:

$$\begin{aligned} F\left(x,u+h,u'+h'\right) = F\left(x,u,u'\right) &+ \frac{\partial F}{\partial u}h + \frac{\partial F}{\partial u'}h' \\ &+ \frac{1}{2}\left(\frac{\partial^2 F}{\partial u \partial u}h^2 + 2\frac{\partial^2 F}{\partial u \partial u'}hh' + \frac{\partial^2 F}{\partial u' \partial u'}h'^2\right) \\ &+ \cdots . \end{aligned} \tag{2.16}$$

Then, we have the first and the second variations as

$$\begin{aligned} \Delta J[h] = &\underbrace{\int_a^b \left(\frac{\partial F}{\partial u}h + \frac{\partial F}{\partial u'}h'\right)dx}_{=\varphi_1[h]} \\ &+ \underbrace{\frac{1}{2}\int_a^b \left(\frac{\partial^2 F}{\partial u \partial u}h^2 + 2\frac{\partial^2 F}{\partial u \partial u'}hh' + \frac{\partial^2 F}{\partial u' \partial u'}h'^2\right)dx}_{=\varphi_2[h]} \\ &+ \cdots . \end{aligned} \tag{2.17}$$

Here, $\partial F/\partial u = u$, $\partial F/\partial u' = u'$, $\partial^2 F/\partial u \partial u = 1$, $\partial^2 F/\partial u \partial u' = 0$, and $\partial^2 F/\partial u' \partial u' = 1$. Then, the first and second variations are

$$dJ[u;h] = \int_a^b \left(uh + u'h'\right)dx = \int_a^b \left(u - u''\right)hdx \quad \text{and} \tag{2.18}$$

$$d^2J[u;h] = \frac{1}{2}\int_a^b \left(h^2 + h'^2\right)dx. \tag{2.19}$$

## 2.3 Necessary and sufficient conditions for local extrema

Here, we define local extrema (i.e., minima and maxima) of a functional and state the necessary and sufficient conditions for their existence.

**Definition 2.3.1 (Extremum)** *Let* $J[u]$ *be a differentiable functional. Then,*

*$J[u]$ is said to have a local* extremum *at $\hat{u}$ if*

$$\Delta J \coloneqq J[u] - J[\hat{u}] \tag{2.20}$$

*does not change its sign in some neighborhood of $\hat{u}$. $J[\hat{u}]$ is a* minimum *if $\Delta J \geq 0$ and a* maximum *if $\Delta J \leq 0$.*

We say $J[u]$ has a *weak extremum* at $u = \hat{u}$ if there exists $\epsilon > 0$ such that $\Delta J$ has constant sign for all $u$ satisfying $\|u - \hat{u}\|_{C^1} < \epsilon$. We say $J[u]$ has a *strong extremum* at $u = \hat{u}$ if the neighborhood is defined by $\|u - \hat{u}\|_{C^0} < \epsilon$.

The above definitions use Banach spaces $C^1(\Omega)$ and $C^0(\Omega)$. Alternatively, for problems involving weak solutions of PDEs, we can define extrema in Hilbert (Sobolev) spaces by replacing the neighborhoods with $\|u - \hat{u}\|_{H^1(\Omega)} < \epsilon$ (weak extremum) or $\|u - \hat{u}\|_{L^2(\Omega)} < \epsilon$ (strong extremum).

**Theorem 2.3.1 (First-Order Necessary Condition for Extrema)** *A necessary condition for a differentiable functional $J[u]$ to have a local extremum at $u = \hat{u}$ is that its first variation vanishes at $u = \hat{u}$:*

$$dJ(\hat{u})[h] = 0, \quad \forall h, \tag{2.21}$$

*where $h$ is any admissible direction.*

**Theorem 2.3.2 (Second-Order Necessary Condition for Extrema)** *Let $J[u]$ be a twice differentiable functional. A necessary condition for $J[u]$ to have a local minimum at $u = \hat{u}$ is that*

$$d^2J(\hat{u})[h] \geq 0, \quad \forall h, \tag{2.22}$$

*where $h$ is any admissible direction. For a local maximum, $d^2J(\hat{u})[h] \leq 0$.*

**Theorem 2.3.3 (Sufficient Condition for a Local Minimum)** *Let $J[u]$ be a twice differentiable functional, and suppose that the first-order necessary condition*

$$dJ(\hat{u})[h] = 0, \quad \forall h, \tag{2.23}$$

*is satisfied. A sufficient condition for $J[u]$ to have a local minimum at $u = \hat{u}$ is that its second variation be* strongly positive*, i.e., there exists a constant $\alpha > 0$ such that*

$$d^2J(\hat{u})[h] \geq \alpha\|h\|^2, \quad \forall h. \tag{2.24}$$

Since $\hat{u}$ satisfies the first-order necessary condition $dJ[h] = 0$ for all admissible $h$, the increment of the functional can be written as

$$\Delta J[h] \coloneqq J[\hat{u} + h] - J[\hat{u}] = d^2J[h] + \epsilon(h)\|h\|^2, \tag{2.25}$$

where $\epsilon(h) \to 0$ as $\|h\| \to 0$. Suppose that the second variation is strongly positive, i.e., there exists a constant $\alpha > 0$ such that

$$d^2J[h] \geq \alpha\|h\|^2, \quad \forall h. \tag{2.26}$$

Since $\epsilon(h) \to 0$ as $\|h\| \to 0$, there exists $\epsilon_1 > 0$ such that $|\epsilon(h)| < \frac{1}{2}\alpha$ whenever $\|h\| < \epsilon_1$. It follows that, for all admissible $h$ with $0 < \|h\| < \epsilon_1$,

$$\Delta J[h] = d^2J[h] + \epsilon(h)\|h\|^2 > \frac{1}{2}\alpha\|h\|^2 > 0. \tag{2.27}$$

Therefore, $J[\hat{u} + h] > J[\hat{u}]$ for all admissible $h$ with $0 < \|h\| < \epsilon_1$, which establishes that $\hat{u}$ is a strict local minimum.

## 2.4 Euler equation

Let $F(x, u, u')$ be a function with continuous first and second derivatives with respect to all of its arguments. Among all continuously differentiable functions $u$ defined on $[a, b]$ satisfying the boundary conditions

$$u(a) = A \quad \text{and} \quad u(b) = B, \tag{2.28}$$

we seek the function for which the functional

$$J[u] = \int_a^b F(x, u, u')dx \tag{2.29}$$

has a weak extremum.

Suppose we give $u(x)$ an increment $h(x)$. In order for the perturbed function $u + h$ to satisfy the boundary conditions, we require

$$h(a) = 0 \quad \text{and} \quad h(b) = 0. \tag{2.30}$$

The corresponding increment of the functional is then

$$\begin{aligned}\Delta J &= J[u + h] - J[u] \\ &= \int_a^b F(x, u + h, u' + h')dx - \int_a^b F(x, u, u')dx \\ &= \int_a^b \left[F(x, u + h, u' + h') - F(x, u, u')\right] dx. \end{aligned} \tag{2.31}$$

Expanding the integrand by Taylor's theorem yields

$$\Delta J = \int_a^b \left[\frac{\partial F}{\partial u}h + \frac{\partial F}{\partial u'}h'\right] dx + \cdots . \tag{2.32}$$

The leading term on the right-hand side is the principal linear part of the increment $\Delta J$, which we identify as the first variation. A necessary condition for the weak extremum is that this first variation vanish for all admissible $h$:

$$d_h J = \int_a^b \left[\frac{\partial F}{\partial u}h + \frac{\partial F}{\partial u'}h'\right] dx = 0. \tag{2.33}$$

Integrating the second term by parts and using $h(a) = h(b) = 0$, we obtain

$$d_h J = \int_a^b \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'}\right] h dx. \tag{2.34}$$

Since $h$ is an arbitrary continuously differentiable function vanishing at the endpoints, we have

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} = 0. \tag{2.35}$$

This is the *Euler equation* (also called the *Euler-Lagrange equation*).

**Theorem 2.4.1 (Euler Equation)** *Let $J[u]$ be a functional of the form*

$$J[u] = \int_a^b F(x, u, u')dx, \tag{2.36}$$

*defined on the set of functions $u(x)$ having continuous first derivatives on $[a, b]$ and satisfying the boundary conditions $u(a) = A$ and $u(b) = B$. A necessary condition for $J[u]$ to have a weak extremum at a given function $u(x)$ is that $u(x)$ satisfy the* Euler equation*:*

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} = 0. \tag{2.37}$$

A functional $J[u] = \int_a^b F dx$ is called *autonomous* if the integrand has the form $F = F(u, u')$, with no explicit dependence on $x$. Then, the Euler equation admits a first integral known as the *Beltrami identity*:

$$F - u'\frac{\partial F}{\partial u'} = c, \quad c = \text{const.} \tag{2.38}$$

This follows from the observation that, when the Euler equation is satisfied,

$$\frac{d}{dx}\left(F - u'\frac{\partial F}{\partial u'}\right) = u'\left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'}\right) = 0. \tag{2.39}$$

**Example 2.4.1 (Minimal Surface of Revolution)** *Consider finding a curve $u(x) > 0$ on the interval $[a, b]$ that generates a surface of revolution of minimal area when rotated about the $x$-axis, subject to the boundary conditions $u(a) = A > 0$ and $u(b) = B > 0$. The minimization problem reads:*

$$\begin{cases} \text{Find } u(x),\ x \in [a, b],\ \text{such that} \\ u(a) = A,\ u(b) = B \\ J[u] = \inf\limits_{w \in U} J[w] \end{cases} \tag{2.40}$$

*where the admissible function space is*

$$U \coloneqq \left\{ w \in C^1([a, b]) : w(a) = A,\ w(b) = B \right\} \tag{2.41}$$

*and the cost functional is*

$$J[w] = 2\pi \int_C w ds = 2\pi \int_a^b w\sqrt{1 + (w')^2}dx. \tag{2.42}$$

*Here, $ds = \sqrt{1 + (w')^2}dx$ denotes the arc length element and $C$ is the curve parameterized by $x \in [a, b]$.*

The integrand $F(w, w') = 2\pi w\sqrt{1 + (w')^2}$ is autonomous, where we have the Beltrami identity:

$$F - w'\frac{\partial F}{\partial w'} = c, \tag{2.43}$$

where $c$ is a constant. Computing

$$w' \frac{\partial F}{\partial w'} = 2\pi w' \cdot \frac{w w'}{\sqrt{1 + (w')^2}} = \frac{2\pi w (w')^2}{\sqrt{1 + (w')^2}}, \tag{2.44}$$

the Beltrami identity yields

$$2\pi w \sqrt{1 + (w')^2} - \frac{2\pi w (w')^2}{\sqrt{1 + (w')^2}} = c, \tag{2.45}$$

which simplifies to

$$w = \frac{c}{2\pi} \sqrt{1 + (w')^2}. \tag{2.46}$$

Setting $c = 2\pi a_0$ for a positive constant $a_0$, we have a general solution (*catenary*)

$$u(x) = a_0 \cosh\left(\frac{x - x_0}{a_0}\right), \tag{2.47}$$

where $a_0$ and $x_0$ are determined by the boundary conditions $u(a) = A$ and $u(b) = B$.

This problem models the shape of a soap film stretched between two coaxial circular rings. When the constant $2\pi$ is replaced by the weight per unit arc length $\rho g$, the functional becomes the gravitational potential energy of a hanging cable, and the catenary solution then describes the equilibrium shape of a chain or cable suspended under its own weight.

The Beltrami identity is the variational counterpart of energy conservation in classical mechanics. For example, consider a mechanical system with generalized coordinate $q(t)$ and *Lagrangian* $L(q, \dot{q}) = T - V$, where $T$ is the kinetic energy and $V$ is the potential energy. The *action functional* is

$$S[q] = \int_{t_1}^{t_2} L(q, \dot{q}) dt, \tag{2.48}$$

and the Euler equation (i.e., Newton's second law in generalized form) is

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} = 0. \tag{2.49}$$

When the Lagrangian does not depend explicitly on time $t$ (i.e., $\partial L / \partial t = 0$), the Beltrami identity applied with $t$ as the independent variable gives

$$L - \dot{q} \frac{\partial L}{\partial \dot{q}} = -c, \tag{2.50}$$

where $c$ is a constant along solutions of (2.49). The *Hamiltonian* (total energy) is defined as

$$H \coloneqq \dot{q} \frac{\partial L}{\partial \dot{q}} - L, \tag{2.51}$$

For a mechanical system with generalized coordinates $q = (q_1, \ldots, q_n)$ and generalized velocities $\dot{q} = (\dot{q}_1, \ldots, \dot{q}_n)$, the *Lagrangian* $L \colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is defined by

$$L(q, \dot{q}, t) \coloneqq T(q, \dot{q}, t) - V(q, t),$$

where $T$ is the kinetic energy and $V$ is the potential energy of the system.

The *principle of stationary action* states that the equations of motion are obtained by requiring the *action functional*

$$S[q] \coloneqq \int_{t_1}^{t_2} L(q, \dot{q}, t) dt$$

to be *stationary* (i.e., $dS = 0$), not necessarily minimized.

which is constant as suggested by (2.50). This is the *conservation of energy*: the Hamiltonian $H$ is constant along trajectories satisfying the Euler equation, provided $L$ has no explicit time dependence.

**Example 2.4.2 (Conservation of Energy of a Moving Particle)** *For a particle of mass $m$ moving in one dimension under a potential $V(q)$, the Lagrangian is*

$$L(q, \dot{q}) = \underbrace{\tfrac{1}{2} m \dot{q}^2}_{=T} - \underbrace{V(q)}_{=V}. \tag{2.52}$$

*Derive the conservation of energy.*

Since $L$ does not depend explicitly on time $t$, the system is autonomous and the Beltrami identity guarantees the existence of a first integral. The canonical momentum is

$$p \coloneqq \frac{\partial L}{\partial \dot{q}} = m\dot{q}. \tag{2.53}$$

The Hamiltonian is

$$\begin{aligned} H &= p\dot{q} - L(q, \dot{q}) \\ &= m\dot{q} \cdot \dot{q} - \left(\tfrac{1}{2} m \dot{q}^2 - V(q)\right) \\ &= \tfrac{1}{2} m \dot{q}^2 + V(q) \\ &= T + V. \end{aligned} \tag{2.54}$$

The Beltrami identity states that $H$ is constant along trajectories satisfying the Euler equation, i.e.,

$$T + V = \text{const}, \tag{2.55}$$

which is the classical conservation of mechanical energy.

## 2.5 Canonical form of the Euler equations

Consider a functional of several dependent variables $u_1, u_2, \ldots, u_n$:

$$J[u_1, u_2, \ldots, u_n] = \int_a^b F\left(x, u_1, u_2, \ldots, u_n, u_1', u_2', \ldots, u_n'\right) dx. \tag{2.56}$$

The corresponding Euler equations are

$$\frac{\partial F}{\partial u_i} - \frac{d}{dx} \frac{\partial F}{\partial u_i'} = 0, \quad i = 1, 2, \ldots, n, \tag{2.57}$$

which form a system of $n$ second-order ordinary differential equations for the unknown functions $u_i(x)$. This system can be rewritten as $2n$ first-order equations

by introducing new variables for $u_i'$:

$$\frac{du_i}{dx} = u_i', \quad \frac{d}{dx}\frac{\partial F}{\partial u_i'} = \frac{\partial F}{\partial u_i}, \quad i = 1, 2, \ldots, n. \tag{2.58}$$

We define the *canonical momenta* $p_i$ as

$$p_i \coloneqq \frac{\partial F}{\partial u_i'}, \quad i = 1, 2, \ldots, n. \tag{2.59}$$

Provided the map $(u_1', \ldots, u_n') \mapsto (p_1, \ldots, p_n)$ is invertible, each $u_i'$ can be expressed as a function of $(x, u_1, \ldots, u_n, p_1, \ldots, p_n)$.

We introduce the *Hamiltonian* $H(x, u_1, \ldots, u_n, p_1, \ldots, p_n)$, defined via the *Legendre transform*:

$$H \coloneqq \sum_{i=1}^{n} p_i u_i' - F, \tag{2.60}$$

where each $u_i'$ on the right-hand side is understood as a function of $(x, u_i, p_i)$ obtained by inverting (2.59). Together with the generalized coordinates $u_i$, the canonical momenta $p_i$ form the *canonical variables*

$$(x, u_1, \ldots, u_n, p_1, \ldots, p_n)\,. \tag{2.61}$$

We now examine the differential of $H$ to express the first-order system (2.58) in terms of the canonical variables. Applying the product rule to (2.60) gives

$$dH = \sum_{i=1}^{n} p_i \, du_i' + \sum_{i=1}^{n} u_i' \, dp_i - dF. \tag{2.62}$$

Expanding $dF$ by the chain rule yields

$$dH = \sum_{i=1}^{n} u_i' \, dp_i + \sum_{i=1}^{n} p_i \, du_i' - \frac{\partial F}{\partial x}dx - \sum_{i=1}^{n} \frac{\partial F}{\partial u_i}du_i - \sum_{i=1}^{n} \frac{\partial F}{\partial u_i'}du_i'. \tag{2.63}$$

Since $p_i = \partial F/\partial u_i'$, the $du_i'$ terms cancel, leaving

$$dH = -\frac{\partial F}{\partial x}dx - \sum_{i=1}^{n} \frac{\partial F}{\partial u_i}du_i + \sum_{i=1}^{n} u_i' \, dp_i. \tag{2.64}$$

On the other hand, since $H$ is regarded as a function of $(x, u_i, p_i)$, its total differential is

$$dH = \frac{\partial H}{\partial x}dx + \sum_{i=1}^{n} \frac{\partial H}{\partial u_i}du_i + \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}dp_i. \tag{2.65}$$

Comparing coefficients of $dx$, $du_i$, and $dp_i$ between (2.64) and (2.65), we obtain

$$\frac{\partial H}{\partial x} = -\frac{\partial F}{\partial x}, \ \frac{\partial H}{\partial u_i} = -\frac{\partial F}{\partial u_i}, \text{ and } \frac{\partial H}{\partial p_i} = u_i'. \tag{2.66}$$

Substituting (2.66) into the first-order system (2.58) yields the *canonical form of the Euler equations* (Hamilton's equations):

$$\frac{du_i}{dx} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dx} = -\frac{\partial H}{\partial u_i}, \quad i = 1, 2, \ldots, n. \tag{2.67}$$

Let us revisit the case of an autonomous problem, i.e., $F$ does not depend explicitly on $x$, so that $H$ likewise has no explicit $x$-dependence. Along an extremal, the total derivative of $H$ with respect to $x$ is

$$\frac{dH}{dx} = \frac{\partial H}{\partial x} + \sum_{i=1}^{n} \left( \frac{\partial H}{\partial u_i} \frac{du_i}{dx} + \frac{\partial H}{\partial p_i} \frac{dp_i}{dx} \right). \tag{2.68}$$

Since $\partial H/\partial x = 0$ by hypothesis, substituting the canonical equations (2.67) gives

$$\frac{dH}{dx} = \sum_{i=1}^{n} \left( \frac{\partial H}{\partial u_i} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial u_i} \right) = 0. \tag{2.69}$$

Thus, the Hamiltonian is a first integral of the canonical system: $H$ is constant along every extremal of an autonomous problem.

The following definition restates the Hamiltonian in the standard language of analytical mechanics:

**Definition 2.5.1 (Hamiltonian via the Legendre Transform)** *Given a Lagrangian $L(q, \dot{q}, t)$ with generalized coordinates $q = (q_1, \ldots, q_n)$ and generalized velocities $\dot{q} = (\dot{q}_1, \ldots, \dot{q}_n)$, the canonical momenta are defined by*

$$p_i \coloneqq \frac{\partial L}{\partial \dot{q}_i}, \quad i = 1, \ldots, n. \tag{2.70}$$

*Assuming the map $\dot{q} \mapsto p$ is invertible, so that $\dot{q}_i$ can be expressed as a function of $(q, p, t)$, the Hamiltonian $H \colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is defined by the Legendre transform of $L$ with respect to $\dot{q}$:*

$$H(q, p, t) \coloneqq \sum_{i=1}^{n} p_i \dot{q}_i - L(q, \dot{q}, t), \tag{2.71}$$

*where it is understood that all $\dot{q}_i$ on the right-hand side are replaced by their expressions in terms of $(q, p, t)$.*

The invertibility of the map $\dot{q} \mapsto p$ requires

$$\det \left[ \frac{\partial p_i}{\partial \dot{q}_j} \right] = \det \left[ \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} \right] \neq 0. \tag{2.72}$$

When $L = T - V$ with $T$ a positive-definite quadratic form in $\dot{q}$ and $V = V(q, t)$, the Hessian reduces to $\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} = \frac{\partial^2 T}{\partial \dot{q}_i \partial \dot{q}_j}$, which is positive definite and hence non-singular.

## 2.6 Noether's theorem

Consider a functional of the form

$$J[\mathtt{u}] = \int_a^b F(x, \mathtt{u}, \mathtt{u}')dx, \tag{2.73}$$

where $\mathtt{u} = (u_1, \ldots, u_n)$ and $\mathtt{u}' = (u_1', \ldots, u_n')$. Suppose the functional is invariant under the one-parameter family of transformations

$$x^* = \Phi(x, \mathtt{u}; \varepsilon) \quad \text{and} \tag{2.74}$$
$$u_i^* = \Psi_i(x, \mathtt{u}; \varepsilon),\ i = 1, \ldots, n, \tag{2.75}$$

where $\varepsilon = 0$ corresponds to the identity transformation $x^* = x$, $u_i^* = u_i$. Invariance means that for every admissible curve $u_i(x)$ and every $\varepsilon$ in a neighborhood of zero,

$$\int_{a^*}^{b^*} F\left(x^*, \mathtt{u}^*, \frac{d\mathtt{u}^*}{dx^*}\right) dx^* = \int_a^b F\left(x, \mathtt{u}, \frac{d\mathtt{u}}{dx}\right) dx, \tag{2.76}$$

where $a^* = \Phi(a, \mathtt{u}(a); \varepsilon)$ and $b^* = \Phi(b, \mathtt{u}(b); \varepsilon)$.

**Theorem 2.6.1 (Noether's theorem)** *If the functional $J$ is invariant under the one-parameter family of transformations*

$$x^* = \Phi(x, \mathtt{u}; \varepsilon), \quad u_i^* = \Psi_i(x, \mathtt{u}; \varepsilon), \quad i = 1, \ldots, n, \tag{2.77}$$

*with $\Phi(x, \mathtt{u}; 0) = x$ and $\Psi_i(x, \mathtt{u}; 0) = u_i$, and $u_1, \ldots, u_n$ satisfy the Euler equations, then*

$$\sum_{i=1}^n \frac{\partial F}{\partial u_i'}\psi_i + \left(F - \sum_{i=1}^n u_i' \frac{\partial F}{\partial u_i'}\right)\phi = const \tag{2.78}$$

*along each extremal of $J[\mathtt{u}]$, where the* infinitesimal variations*, or* generators*, are*

$$\phi(x, \mathtt{u}, \mathtt{u}') \coloneqq \left.\frac{\partial \Phi}{\partial \varepsilon}\right|_{\varepsilon=0} \quad \textit{and} \quad \psi_i(x, \mathtt{u}, \mathtt{u}') \coloneqq \left.\frac{\partial \Psi_i}{\partial \varepsilon}\right|_{\varepsilon=0}. \tag{2.79}$$

*In terms of the canonical variables,* (2.78) *reads*

$$\sum_{i=1}^n p_i\psi_i - H\phi = const. \tag{2.80}$$

Thus, Noether's theorem establishes a link between symmetries and conserved quantities.

The Beltrami identity is a special case of Noether's theorem. Consider a functional invariant under the following transformation

$$x^* = \Phi(x, u; \varepsilon) = x + \epsilon \quad \text{and} \quad u^* = \Psi(x, u; \varepsilon) = u, \tag{2.81}$$

which implies translational symmetry with respect to $x$. Then, we have

$$\phi = 1 \quad \text{and} \quad \psi = 0. \tag{2.82}$$

Thus, (2.80) yields

$$H = \text{const.} \tag{2.83}$$

Thus, the total energy is conserved when the Hamiltonian represents the total energy.

On the other hand, if $F$ is invariant under the transformation

$$x^* = \Phi(x, u; \varepsilon) = x \quad \text{and} \quad u^* = \Psi(x, u; \varepsilon) = u + \epsilon, \tag{2.84}$$

Noether's theorem reduces to

$$p = \text{const}, \tag{2.85}$$

which is identified as the *conservation of momentum.*

**Example 2.6.1 (Conservation of Total Momentum)** *Consider two particles with positions $q_1(t)$ and $q_2(t)$ interacting through a potential that depends only on their separation. The Lagrangian is*

$$L(q_1, q_2, \dot{q}_1, \dot{q}_2) = \tfrac{1}{2} m_1 \dot{q}_1^2 + \tfrac{1}{2} m_2 \dot{q}_2^2 - V(q_1 - q_2). \tag{2.86}$$

*Derive the conservation of total momentum.*

The potential $V$ depends on $q_1$ and $q_2$ only through the difference $q_1 - q_2$. Therefore the Lagrangian is invariant under simultaneous translation of both coordinates:

$$t^* = t, \ q_1^* = q_1 + \varepsilon, \text{ and } q_2^* = q_2 + \varepsilon. \tag{2.87}$$

Then, we have

$$\phi = 0, \ \psi_1 = 1, \text{ and } \psi_2 = 1. \tag{2.88}$$

Applying Noether's theorem, we have

$$p_1 \psi_1 + p_2 \psi_2 = \text{const}, \tag{2.89}$$

which is the total linear momentum of the system. Note that neither individual momentum $p_1$ nor $p_2$ is conserved.

**Example 2.6.2 (Conservation of Angular Momentum)** *Consider a particle of mass $m$ moving freely in the $(q_1, q_2)$-plane. The Lagrangian is*

$$L(\dot{q}_1, \dot{q}_2) = \tfrac{1}{2} m (\dot{q}_1^2 + \dot{q}_2^2). \tag{2.90}$$

*Derive the conservation of angular momentum.*

The kinetic energy $\frac{1}{2}m(\dot{q}_1^2 + \dot{q}_2^2)$ is invariant under rotations in the $(q_1, q_2)$-plane. Consider the one-parameter family of rotations

$$t^* = t,\ q_1^* = q_1 \cos\varepsilon - q_2 \sin\varepsilon,\ \text{and}\ q_2^* = q_1 \sin\varepsilon + q_2 \cos\varepsilon. \tag{2.91}$$

Here, we have $\dot{q}_1^{*2} + \dot{q}_2^{*2} = \dot{q}_1^2 + \dot{q}_2^2$; thus, $L$ is unchanged. Then, we have

$$\phi = 0,\ \psi_1 = \left.\frac{\partial q_1^*}{\partial \varepsilon}\right|_{\varepsilon=0} = -q_2,\ \text{and}\ \psi_2 = \left.\frac{\partial q_2^*}{\partial \varepsilon}\right|_{\varepsilon=0} = q_1. \tag{2.92}$$

Applying Noether's theorem, we have

$$p_1\,(-q_2) + p_2 q_1 = \text{const}, \tag{2.93}$$

which is the *angular momentum* of the particle about the origin.

**Example 2.6.3 (Free Relativistic Particle in $(1+1)$ Dimensions)** *In special relativity, the action $J$ for a free particle of mass $m$ moving in one spatial dimension is proportional to the proper time elapsed along the trajectory $C$:*

$$J = -mc \int_C ds, \tag{2.94}$$

*where $c$ is the speed of light and $ds = \sqrt{c^2 dt^2 - dx^2}$ is the infinitesimal spacetime interval.*

*By factoring out $dt$ and using coordinates $(t, q)$ where $t$ is the coordinate time and $q(t)$ is the spatial position, the action functional yields:*

$$J[q] = \int_{t_1}^{t_2} \left(-mc^2\sqrt{1 - \frac{\dot{q}^2}{c^2}}\right) dt. \tag{2.95}$$

*The corresponding Lagrangian is:*

$$L(\dot{q}) = -mc^2\sqrt{1 - \frac{\dot{q}^2}{c^2}}. \tag{2.96}$$

*This Lagrangian possesses two symmetries: it is autonomous (independent of $t$) and independent of $q$. Derive the associated conservation laws.*

*1. Spatial Symmetry:* Since $L$ does not depend on $q$, the Lagrangian is invariant under $t^* = t$, $q^* = q + \varepsilon$. By Noether's theorem, the conserved quantity is:

$$p = \frac{\partial L}{\partial \dot{q}} = \frac{m\dot{q}}{\sqrt{1 - \dot{q}^2/c^2}} = \gamma m \dot{q}, \tag{2.97}$$

where $\gamma := (1 - \dot{q}^2/c^2)^{-1/2}$ is the Lorentz factor. $p$ is the *relativistic momentum*. In the limit $\dot{q}/c \to 0$, we recover $p \approx m\dot{q}$.

*2. Temporal Symmetry:* Since $L$ does not depend on $t$, the Beltrami identity gives the conserved Hamiltonian:

$$\begin{aligned} H &= \dot{q}\frac{\partial L}{\partial \dot{q}} - L \\ &= \frac{m\dot{q}^2}{\sqrt{1-\dot{q}^2/c^2}} + mc^2\sqrt{1-\frac{\dot{q}^2}{c^2}} \\ &= \frac{mc^2}{\sqrt{1-\dot{q}^2/c^2}} = \gamma mc^2 = E, \end{aligned} \tag{2.98}$$

which is the *relativistic total energy.* In the low-velocity limit, $E \approx mc^2 + \frac{1}{2}m\dot{q}^2$, where the Taylor expansion of the Lorentz factor about $\dot{q} = 0$ reads

$$\gamma = \left(1 - \frac{\dot{q}^2}{c^2}\right)^{-1/2} = 1 + \frac{1}{2}\frac{\dot{q}^2}{c^2} + \dots. \tag{2.99}$$

## 2.7 General variation of a functional

We now derive the general variation of a functional, allowing simultaneous variation of the functions and the limits of integration. Consider a functional of the form

$$J[\mathtt{u}] = \int_{x_a}^{x_b} F(x, \mathtt{u}, \mathtt{u}')dx, \tag{2.100}$$

where $\mathtt{u} = (u_1, u_2, \dots, u_n)$ and $\mathtt{u}' = (u_1', u_2', \dots, u_n')$. For each function $u_i$, we consider a varied function $u_i^*$, and we simultaneously perturb the endpoints of the integration interval. The distance between the original and varied configurations is defined as

$$d(u_i, u_i^*) \coloneqq \max|u_i - u_i^*| + \max|u_i' - u_i'^*| + \|A_i - A_i^*\| + \|B_i - B_i^*\|, \tag{2.101}$$

where $\|\cdot\|$ denotes the Euclidean norm. The boundary points and their varied counterparts are

$$A_i = (x_a, u_i^a),\ A_i^* = (x_a + \delta x_a, u_i^a + \delta u_i^a) \quad \text{and} \tag{2.102}$$

$$B_i = (x_b, u_i^b),\ B_i^* = (x_b + \delta x_b, u_i^b + \delta u_i^b). \tag{2.103}$$

Here, $\delta x_a$ and $\delta x_b$ are the variations of the endpoints of the interval, $u_i^a = u_i(x_a)$ and $u_i^b = u_i(x_b)$ are the functions evaluated at the end points, and $\delta u_i^a$ and $\delta u_i^b$ are the corresponding variations of $u_i$ at the boundaries.

Let

$$\mathtt{h} = \mathtt{u}^* - \mathtt{u}. \tag{2.104}$$

Then, the increment reads

$$\begin{aligned}\Delta J &= J[\mathtt{u}+\mathtt{h}] - J[\mathtt{u}] \\ &= \int_{x_a+\delta x_a}^{x_b+\delta x_b} F(x, \mathtt{u}+\mathtt{h}, \mathtt{u}'+\mathtt{h}')\, dx - \int_{x_a}^{x_b} F(x, \mathtt{u}, \mathtt{u}')\, dx \\ &= \int_{x_a}^{x_b} [F(x, \mathtt{u}+\mathtt{h}, \mathtt{u}'+\mathtt{h}') - F(x, \mathtt{u}, \mathtt{u}')]\, dx \\ &\quad + \int_{x_b}^{x_b+\delta x_b} F(x, \mathtt{u}+\mathtt{h}, \mathtt{u}'+\mathtt{h}')\, dx \\ &\quad - \int_{x_a}^{x_a+\delta x_a} F(x, \mathtt{u}+\mathtt{h}, \mathtt{u}'+\mathtt{h}')\, dx \end{aligned} \tag{2.105}$$

The Taylor expansion up to the linear term, i.e., retaining only first-order terms relative to $\sum_{i=1}^{n} d(u_i, u_i^*)$, we have

$$\begin{aligned}\Delta J &\sim \int_{x_a}^{x_b} \sum_{i=1}^{n} \left( \frac{\partial F}{\partial u_i} h_i + \frac{\partial F}{\partial u_i'} h_i' \right) dx + F|_{x=x_b}\, \delta x_b - F|_{x=x_a}\, \delta x_a \\ &= \int_{x_a}^{x_b} \sum_{i=1}^{n} \left( \frac{\partial F}{\partial u_i} - \frac{d}{dx}\frac{\partial F}{\partial u_i'} \right) h_i dx + \left[ \sum_{i=1}^{n} \frac{\partial F}{\partial u_i'} h_i \right]_{x=x_a}^{x=x_b} \\ &\quad + F|_{x=x_b}\, \delta x_b - F|_{x=x_a}\, \delta x_a \end{aligned} \tag{2.106}$$

Similarly, $h_i$ at each boundary can be expressed as

$$h_i(x_a) \sim \delta u_i^a - u_i'(x_a)\, \delta x_a \quad \text{and} \tag{2.107}$$

$$h_i(x_b) \sim \delta u_i^b - u_i'(x_b)\, \delta x_b. \tag{2.108}$$

Then, plugging the above into (2.106) gives the general variation of the functional $J[\mathtt{u}]$:

$$\begin{aligned} dJ = &\int_{x_a}^{x_b} \sum_{i=1}^{n} \left( \frac{\partial F}{\partial u_i} - \frac{d}{dx}\frac{\partial F}{\partial u_i'} \right) h_i dx \\ &+ \left[ \sum_{i=1}^{n} \frac{\partial F}{\partial u_i'} \delta u_i + \left( F - \sum_{i=1}^{n} u_i' \frac{\partial F}{\partial u_i'} \right) \delta x \right]_{x=x_a}^{x=x_b}. \end{aligned} \tag{2.109}$$

In the above, $\delta x|_{x=x_k} = \delta x_k$ and $\delta u_i|_{x=x_k} = \delta u_i^k$, $k = a, b$.

Suppose the functional has an extremum for some curve $u_i$ joining the points $A_i$ and $B_i$. Then, the Euler equation is satisfied; thus, (2.109) becomes

$$dJ = \left[ \sum_{i=1}^{n} \frac{\partial F}{\partial u_i'} \delta u_i + \left( F - \sum_{i=1}^{n} u_i' \frac{\partial F}{\partial u_i'} \right) \delta x \right]_{x=x_a}^{x=x_b} \tag{2.110}$$

or, in terms of the canonical variables,

$$dJ = \left[ \sum_{i=1}^{n} p_i \delta u_i - H \delta x \right]_{x=x_a}^{x=x_b}. \tag{2.111}$$

Note that Noether's theorem is a special case of (2.111) when

$$\delta x = \varepsilon\phi \quad \text{and} \tag{2.112}$$
$$\delta u_i = \varepsilon\psi_i. \tag{2.113}$$

In the above, for small $\varepsilon$, Taylor expansion gives

$$x^* = \Phi(x, \mathtt{u}; \varepsilon) = x + \varepsilon\,\phi(x, \mathtt{u}, \mathtt{u}') + o(\varepsilon), \tag{2.114}$$
$$u_i^* = \Psi_i(x, \mathtt{u}; \varepsilon) = u_i + \varepsilon\,\psi_i(x, \mathtt{u}, \mathtt{u}') + o(\varepsilon). \tag{2.115}$$

From the assumption $J[\mathtt{u}]$ being invariant under the transformation, we have $dJ = 0$, i.e.,

$$dJ = \varepsilon \left[\sum_{i=1}^{n} p_i\psi_i - H\phi\right]_{x=x_a}^{x=x_b} = 0 \tag{2.116}$$

or

$$\left[\sum_{i=1}^{n} p_i\psi_i - H\phi\right]_{x=x_a} = \left[\sum_{i=1}^{n} p_i\psi_i - H\phi\right]_{x=x_b}. \tag{2.117}$$

Here, the arbitrariness of $x_a$ and $x_b$ yields Noether's theorem: $\sum_{i=1}^{n} p_i\psi_i - H\phi = \text{const.}$

## 2.8 Functionals depending on higher-order derivatives

We now consider finding an extremum of a functional depending on higher-order derivatives, e.g.,

$$J[u] = \int_a^b F\left(x, u, u', \ldots, u^{(n)}\right) dx. \tag{2.118}$$

The Taylor expansion of the corresponding the increment $\Delta J[h]$ reads

$$\begin{aligned}
\Delta J[h] &= J[u+h] - J[u] \\
&= \int_a^b \left[F\left(x, u+h, u'+h', \ldots, u^{(n)}+h^{(n)}\right) - F\left(x, u, u', \ldots, u^{(n)}\right)\right] dx \\
&= \underbrace{\int_a^b \left(\frac{\partial F}{\partial u}h + \frac{\partial F}{\partial u'}h' + \ldots + \frac{\partial F}{\partial u^{(n)}}h^{(n)}\right) dx}_{=:dJ[h]} + \ldots .
\end{aligned} \tag{2.119}$$

We seek a stationary point where the first variation vanishes, i.e., $dJ[h] = 0$. Applying integration by parts repeatedly yields

$$dJ[h] = \int_a^b \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} + \ldots + (-1)^n \frac{d^n}{dx^n}\frac{\partial F}{\partial u^{(n)}}\right) h\,dx, \tag{2.120}$$

where boundary terms vanish under appropriate boundary conditions on $h$ and its derivatives (e.g., $h^{(k)}(a) = h^{(k)}(b) = 0$ for $k = 0, 1, \ldots, n-1$). From the integration by parts

$$dJ[h] = \int_a^b \left( \frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} + \ldots + (-1)^n \frac{d^n}{dx^n}\frac{\partial F}{\partial u^{(n)}} \right) h dx \tag{2.121}$$

we have the Euler equation as

$$\begin{aligned} 0 &= \frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} + \frac{d^2}{dx^2}\frac{\partial F}{\partial u''} + \ldots + (-1)^n \frac{d^n}{dx^n}\frac{\partial F}{\partial u^{(n)}} \\ &= \sum_{k=0}^{n} (-1)^k \frac{d^k}{dx^k}\frac{\partial F}{\partial u^{(k)}}. \end{aligned} \tag{2.122}$$

For example, the total potential energy functional for the Euler-Bernoulli beam theory reads

$$J[w] = \int_0^L F(x, w, w', w'')\, dx, \tag{2.123}$$

where

$$F(x, w, w', w'') = \frac{1}{2} EI (w'')^2 - qw. \tag{2.124}$$

Here, $w(x)$ is the transverse deflection, $EI$ is the flexural rigidity, and $q(x)$ is the distributed transverse load. Then, the Euler equation

$$\frac{\partial F}{\partial w} - \frac{d}{dx}\frac{\partial F}{\partial w'} + \frac{d^2}{dx^2}\frac{\partial F}{\partial w''} = 0 \tag{2.125}$$

reduces to

$$\frac{d^2}{dx^2}\left( EI \frac{d^2 w}{dx^2} \right) = q. \tag{2.126}$$

**Example 2.8.1 (Displacement reconstruction)** *Suppose acceleration data $a_d(t)$ is measured over the time interval $[0, T]$, together with known displacements at the two endpoints $u(0) = u_0$ and $u(T) = u_T$. Find the displacement history $u(t)$ whose second derivative best fits the acceleration data in the least-squares sense.*

The corresponding variational problem is

$$\inf_u J[u] = \frac{1}{2}\int_0^T \left( \frac{\partial^2 u}{\partial t^2} - a_d \right)^2 dt, \tag{2.127}$$

subject to the boundary conditions

$$u(0) = u_0, \quad u(T) = u_T, \quad \frac{\partial^2 u}{\partial t^2}(0) = 0, \quad \frac{\partial^2 u}{\partial t^2}(T) = 0. \tag{2.128}$$

The conditions $\frac{\partial^2 u}{\partial t^2}(0) = \frac{\partial^2 u}{\partial t^2}(T) = 0$ express the assumption that the acceleration vanishes at both endpoints, which is physically appropriate when the motion begins and ends at rest under zero external forcing.

The integrand is $F = \frac{1}{2}\left(\frac{\partial^2 u}{\partial t^2} - a_d\right)^2$, which depends on $\frac{\partial^2 u}{\partial t^2}$ but on neither $u$ nor $\frac{\partial u}{\partial t}$. The Euler equation therefore reduces to

$$\frac{\partial^2}{\partial t^2}\left(\frac{\partial^2 u}{\partial t^2} - a_d\right) = 0. \tag{2.129}$$

Equation (2.129), together with the boundary conditions (2.128), has the same structure as the simply supported Euler-Bernoulli beam under distributed load and support settlements.

## 2.9 Lagrange multipliers

### 2.9.1 Constrained finite-dimensional minimization problem

Many optimization problems require the solution to satisfy one or more side conditions. Here we consider the case in which every such subsidiary condition takes the form of an equality constraint.

Let $f : D \to \mathbb{R}$ be a real-valued $C^1$ function defined on an open set $D \subset \mathbb{R}^n$, and let

$$g : \mathbb{R}^n \to \mathbb{R}^m, \quad m < n, \tag{2.130}$$

be a $C^1$ map whose $m$ component functions $g_1, \ldots, g_m$ define the equality constraints. The condition $m < n$ ensures that the feasible set is not empty. The constrained minimization problem then reads

$$\min_{\substack{x \in D \\ g(x)=0}} f(x). \tag{2.131}$$

That is, we seek a minimizer within the subset of $D$ defined by $g(x) = 0$.

**Theorem 2.9.1 (Lagrange Multiplier)** *Let $x_0 \in D$ be a local minimizer of problem* (2.131)*, and suppose*

$$\operatorname{rank}\left(\frac{\partial g_i}{\partial x_j}(x_0)\right) = m \tag{2.132}$$

*holds. Then there exists $\lambda_0 \in \mathbb{R}^m$ such that*

$$df(x_0)[h] + \lambda_{0,i} dg_i(x_0)[h] = 0, \quad \forall h \in \mathbb{R}^n. \tag{2.133}$$

*We call $\lambda_0$ the* Lagrange multiplier*.*

Since $g : \mathbb{R}^n \to \mathbb{R}^m$ is $C^1$ and the Jacobian has full rank at $x_0$, condition (2.132) guarantees that, after a possible relabeling of coordinates, the $m \times m$ submatrix

$$F_{ij} \coloneqq \frac{\partial g_i}{\partial y_j}(x_0), \quad i,j = 1,\dots,m, \tag{2.134}$$

is non-singular, i.e., $\det F \neq 0$. Here we have partitioned $x = (y,z)$ with $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n-m}$.

By the Implicit Function Theorem (Theorem 1.6.1), there exist a neighborhood $B \subset \mathbb{R}^{n-m}$ of $z_0$ and a unique $C^1$ map $\varphi : B \to \mathbb{R}^m$ such that $\varphi(z_0) = y_0$ and

$$g(\varphi(z), z) = 0, \quad \forall z \in B. \tag{2.135}$$

The constraint surface is thereby parametrized locally by the free variables $z$, and the constrained problem (2.131) reduces to the unconstrained problem

$$\min_{z \in B} \hat{f}(z), \quad \hat{f}(z) \coloneqq f(\varphi(z), z). \tag{2.136}$$

Since $x_0 = (y_0, z_0)$ is a local minimizer of the constrained problem, $z_0$ is a local minimizer of $\hat{f}$. A necessary condition is

$$\begin{aligned} 0 &= \frac{\partial \hat{f}}{\partial z_l}(z_0) \\ &= \frac{\partial f}{\partial y_k}(x_0)\frac{\partial \varphi_k}{\partial z_l}(z_0) + \frac{\partial f}{\partial z_l}(x_0), \quad l = 1,\dots,n-m. \end{aligned} \tag{2.137}$$

Differentiating (2.135) with respect to $z_l$ and applying the chain rule gives

$$\frac{\partial g_i}{\partial y_k}(x_0)\frac{\partial \varphi_k}{\partial z_l}(z_0) + \frac{\partial g_i}{\partial z_l}(x_0) = 0, \quad i = 1,\dots,m, \tag{2.138}$$

which, by the non-singularity of $\partial g_i / \partial y_k = F_{ik}$, yields the IFT sensitivity formula

$$\frac{\partial \varphi_k}{\partial z_l}(z_0) = -F^{-1}_{kj}(x_0)\,\frac{\partial g_j}{\partial z_l}(x_0). \tag{2.139}$$

Now define

$$\lambda_{0,j} \coloneqq -\frac{\partial f}{\partial y_j}(x_0) F^{-1}_{kj}(x_0). \tag{2.140}$$

Substituting (2.139) into (2.137) gives

$$\underbrace{-\frac{\partial f}{\partial y_k}(x_0) F^{-1}_{kj}(x_0)}_{=\lambda_{0,j}}\,\frac{\partial g_j}{\partial z_l}(x_0) + \frac{\partial f}{\partial z_l}(x_0) = 0, \tag{2.141}$$

which, by (2.140), reads

$$\frac{\partial f}{\partial z_l}(x_0) + \lambda_{0,j}\frac{\partial g_j}{\partial z_l}(x_0) = 0, \quad l = 1,\dots,n-m. \tag{2.142}$$

On the other hand, the definition (2.140) can be rewritten as

$$\frac{\partial f}{\partial y_k}(x_0) + \lambda_{0,j} \frac{\partial g_j}{\partial y_k}(x_0) = 0, \quad k = 1, \dots, m. \tag{2.143}$$

Combining (2.142) and (2.143) and recalling $x = (y, z)$, we obtain

$$\frac{\partial f}{\partial x_i}(x_0) + \lambda_{0,j} \frac{\partial g_j}{\partial x_i}(x_0) = 0, \quad i = 1, \dots, n, \tag{2.144}$$

or equivalently $df(x_0)[h] + \lambda_{0,i} dg_i(x_0)[h] = 0$, $\forall h \in \mathbb{R}^n$.

Now, define the *Lagrangian*

$$L(x, \lambda) \coloneqq f(x) + \lambda_i g_i(x), \tag{2.145}$$

where $\lambda = (\lambda_1, \dots, \lambda_m)^T \in \mathbb{R}^m$ are the Lagrange multipliers. The first-order optimality conditions (2.133) and the constraints are then equivalent to requiring that $L$ be stationary with respect to all of its arguments:

$$\frac{\partial L}{\partial x_i}(x_0, \lambda_0) = \frac{\partial f}{\partial x_i}(x_0) + \lambda_{0,j} \frac{\partial g_j}{\partial x_i}(x_0) = 0, \quad i = 1, \dots, n, \tag{2.146}$$

$$\frac{\partial L}{\partial \lambda_j}(x_0, \lambda_0) = g_j(x_0) = 0, \quad j = 1, \dots, m. \tag{2.147}$$

Equation (2.146) reproduces the Lagrange multiplier condition (2.133), and equation (2.147) recovers the original constraints. The Lagrangian therefore replaces the constrained minimization problem (2.131) by the problem of finding a stationary point of $L$ over $\mathbb{R}^n \times \mathbb{R}^m$.

Note that an alternative sign convention defines the Lagrangian as

$$\mathcal{L}(x, \mu) \coloneqq f(x) - \mu_i g_i(x), \tag{2.148}$$

where $\mu_i$ are the Lagrange multipliers. The two conventions are related by $\mu = -\lambda$ and yield the same minimizer $x_0$; only the sign of the multiplier changes.

**Example 2.9.1 (Closest point on a line)** *Find the point on the line $x_1 + x_2 = 1$ closest to the origin, i.e.,*

$$\min_{\substack{x \in \mathbb{R}^2 \\ g(x) = 0}} f(x) \coloneqq x_1^2 + x_2^2, \qquad g(x) \coloneqq x_1 + x_2 - 1 = 0. \tag{2.149}$$

The Lagrangian is

$$L(x, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1). \tag{2.150}$$

Setting the partial derivatives to zero gives three equations in three unknowns $(x_1, x_2, \lambda)$:

$$\frac{\partial L}{\partial x_1} = 2x_1 + \lambda = 0, \tag{2.151}$$

$$\frac{\partial L}{\partial x_2} = 2x_2 + \lambda = 0, \tag{2.152}$$

$$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - 1 = 0. \tag{2.153}$$

From (2.151) and (2.152), $x_1 = x_2 = -\lambda/2$. Substituting into (2.153):

$$-\frac{\lambda}{2} - \frac{\lambda}{2} = 1 \quad \Longrightarrow \quad \lambda_0 = -1,\ x_0 = \left(\tfrac{1}{2}, \tfrac{1}{2}\right). \tag{2.154}$$

For fixed $\lambda$, the Hessian of $L$ with respect to $(x, y)$ is

$$\nabla^2_{(x,y)} L = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}. \tag{2.155}$$

Thus, $L$ is strictly convex in the primal variables for every $\lambda$.

Substituting $x(\lambda) = y(\lambda) = -\lambda/2$ into $L$, we have

$$\begin{aligned} g(\lambda) &\coloneqq \inf_{x,y} L(x, y, \lambda) = L\left(-\tfrac{\lambda}{2}, -\tfrac{\lambda}{2}, \lambda\right) \\ &= 2 \cdot \frac{\lambda^2}{4} + \lambda\left(-\frac{\lambda}{2} - \frac{\lambda}{2} - 1\right) \\ &= -\frac{\lambda^2}{2} - \lambda. \end{aligned} \tag{2.156}$$

Since $g''(\lambda) = -1 < 0$, the dual function is strictly concave.

### 2.9.2 Constrained variational problems

We now extend the Lagrange multiplier method from finite-dimensional optimization to the variational setting, in which the objective is a functional and the constraint is imposed through another functional.

**Theorem 2.9.2 (Lagrange multiplier for the isoperimetric problem)** *Let the objective functional be*

$$J[u] = \int_a^b F(x, u, u')dx, \tag{2.157}$$

*where the admissible functions satisfy the boundary conditions*

$$u(a) = A \quad \textit{and} \quad u(b) = B \tag{2.158}$$

*and the isoperimetric constraint*

$$K[u] = \int_a^b G(x, u, u')dx = \ell, \tag{2.159}$$

*with $l$ a prescribed constant. Suppose $J[u]$ attains an extremum at an admissible function $u^*$ subject to* (2.159)*, and that $u^*$ is not an extremal of $K[u]$. Then there exists a constant $\lambda$ such that $u^*$ is an extremal of the augmented functional*

$$L[u] \coloneqq \int_a^b (F + \lambda G)\, dx. \tag{2.160}$$

*That is, $u^*$ satisfies the Euler equation*

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} + \lambda\left(\frac{\partial G}{\partial u} - \frac{d}{dx}\frac{\partial G}{\partial u'}\right) = 0. \tag{2.161}$$

The general solution of (2.161) contains two arbitrary constants of integration together with the multiplier $\lambda$. These three unknowns are determined by the two boundary conditions (2.158) and the subsidiary condition (2.159).

**Example 2.9.2 (Catenary prolem with fixed length)** *A chain of uniform linear mass density $\rho$ and prescribed arc length $\ell$ hangs under gravity $g$ between two supports at $(0, y_0)$ and $(L, y_1)$. Find the equilibrium shape by minimizing the gravitational potential energy subject to the arc-length constra int.*

The potential energy of the chain is

$$J[y] = \int_0^L \rho g y \sqrt{1 + y'^2} dx, \tag{2.162}$$

and the arc-length constraint is

$$C[y] = \int_0^L \sqrt{1 + y'^2} dx = \ell, \tag{2.163}$$

with boundary conditions $y(0) = y_0$ and $y(L) = y_1$.

Introducing a Lagrange multiplier $\lambda$ for the arc-length constraint, the augmented functional is

$$L[y, \lambda] = \int_0^L (\rho g y + \lambda) \sqrt{1 + y'^2} dx. \tag{2.164}$$

The integrand $F(y, y') = (\rho g y + \lambda)\sqrt{1 + y'^2}$ depends on $y$ and $y'$ but not explicitly on $x$. Then, the Beltrami identity gives

$$F - y' \frac{\partial F}{\partial y'} = (\rho g y + \lambda)\sqrt{1 + y'^2} - (\rho g y + \lambda) \frac{y'^2}{\sqrt{1 + y'^2}} = c_1. \tag{2.165}$$

Combining over a common denominator yields

$$\frac{\rho g y + \lambda}{\sqrt{1 + y'^2}} = c_1. \tag{2.166}$$

Define $a \coloneqq c_1/(\rho g)$ and $b \coloneqq -\lambda/(\rho g)$. Then (2.166) becomes

$$\frac{y - b}{\sqrt{1 + y'^2}} = a. \tag{2.167}$$

The above equation has a general solution of

$$y(x) = a \cosh\left(\frac{x - x_0}{a}\right) + b, \tag{2.168}$$

The three constants $a$, $b$, and $x_0$ are determined by the two boundary conditions $y(0) = y_0$, $y(L) = y_1$, and the isoperimetric constraint $C[y] = \ell$.

Next, consider a constraint imposed pointwise rather than in integral form.

**Theorem 2.9.3 (Finite Subsidiary Conditions)** *Let the objective be*

$$J[u_1, u_2] = \int_a^b F(x, u_1, u_2, u_1', u_2')dx, \tag{2.169}$$

*where the admissible functions satisfy the boundary conditions*

$$\begin{aligned} u_1(a) &= A_1, \quad u_1(b) = B_1, \\ u_2(a) &= A_2, \quad u_2(b) = B_2, \end{aligned} \tag{2.170}$$

*and the pointwise constraint*

$$g(x, u_1, u_2) = 0. \tag{2.171}$$

*Suppose $J[u_1, u_2]$ has an extremum at the functions*

$$u_1 = u_1(x) \quad \textit{and} \quad u_2 = u_2(x), \tag{2.172}$$

*and that $\partial g/\partial u_1$ and $\partial g/\partial u_2$ do not vanish simultaneously at any point of* (2.171)*. Then there exists a function $\lambda(x)$ such that* (2.172) *is an extremal of the augmented functional*

$$L[u_1, u_2] \coloneqq \int_a^b (F + \lambda(x)g)\, dx. \tag{2.173}$$

*That is, the extremal satisfies the Euler equations*

$$\frac{\partial F}{\partial u_1} + \lambda \frac{\partial g}{\partial u_1} - \frac{d}{dx}\frac{\partial F}{\partial u_1'} = 0, \tag{2.174}$$

$$\frac{\partial F}{\partial u_2} + \lambda \frac{\partial g}{\partial u_2} - \frac{d}{dx}\frac{\partial F}{\partial u_2'} = 0. \tag{2.175}$$

The condition that $\partial g/\partial u_1$ and $\partial g/\partial u_2$ do not vanish simultaneously is the analogue of the constraint qualification (2.132) in the finite-dimensional setting. If both partial derivatives were to vanish at some point $x = x_0$, the constraint $g(x_0, u_1, u_2) = 0$ would be insensitive to variations in $u_1$ and $u_2$ at that point, and the multiplier $\lambda(x_0)$ would be undetermined. More precisely, the constraint surface $g = 0$ defines $u_2$ as a function of $(x, u_1)$ (or $u_1$ as a function of $(x, u_2)$) locally by the Implicit Function Theorem, provided at least one of $\partial g/\partial u_1$ or $\partial g/\partial u_2$ is nonzero. When both vanish, the constraint surface has a singular point and the reduction to an unconstrained problem breaks down.

**Exercise 2.9.1 (Support settlement)** *The principle of minimum potential energy implies that the solution to a simple beam problem is the minimizer of the following problem:*

$$\min M, \quad M[w] = \frac{1}{2}\int_0^L \frac{d^2w}{dx^2} EI \frac{d^2w}{dx^2} dx - \int_0^L wq dx. \tag{2.176}$$

*Here, $q$ is an external load and $w$ is an admissible function satisfying boundary conditions:*

$$w(0) = w(L) = 0 \quad \textit{and} \tag{2.177}$$

$$EI\frac{d^2}{dx^2}w(0) = EI\frac{d^2}{dx^2}w(L) = 0. \tag{2.178}$$

*Now, consider a case when there is an additional support at $x = L/2$ with a settlement $\Delta \in \mathbb{R}$ downward. Then, the above minimization problem can be augmented by a Lagrange multiplier, where the Lagrangian $L$ reads*

$$L[\lambda, w] = M[w] - E[\lambda, w], \quad E[\lambda, w] = \lambda \int_0^L \delta\left(x - \frac{L}{2}\right)[w + \Delta]\, dx. \tag{2.179}$$

*Here, $\lambda \in \mathbb{R}$ is called a Lagrange multiplier and $\delta\left(x - \frac{L}{2}\right)$ is Dirac delta. The necessary conditions for the solution to the above problem are*

$$\begin{aligned} 0 &= d_\lambda L[\lambda, w]\left(\tilde{\lambda}\right) \\ &= -\tilde{\lambda}\int_0^L \delta\left(x - \frac{L}{2}\right)[w + \Delta]\, dx, \quad \forall\tilde{\lambda} \quad \textit{and} \end{aligned} \tag{2.180}$$

$$\begin{aligned} 0 &= d_w L[\lambda, w](\tilde{w}) \\ &= \int_0^L \frac{d^2\tilde{w}}{dx^2} EI \frac{d^2 w}{dx^2} dx - \int_0^L \tilde{w} q dx - \lambda \int_0^L \delta\left(x - \frac{L}{2}\right)\tilde{w} dx, \quad \forall\tilde{w}. \end{aligned} \tag{2.181}$$

*In the above, we have two equations, (2.180) and (2.181), and two unknowns: $w$ and $\lambda$. $\tilde{w}$ and $\tilde{\lambda}$ are the arbitrary directions or test functions.*

(a) *The strong form of the condition (2.180) is obtained by simply removing the dependency of $\tilde{\lambda}$, i.e.,*

$$w\left(\frac{L}{2}\right) + \Delta = 0. \tag{2.182}$$

*Perform integration by parts on the condition (2.181) to obtain the strong form:*

$$\begin{cases} \dfrac{d^2}{dx^2}\left[EI\dfrac{d^2 w}{dx^2}\right] = q + \lambda\delta\left(x - \dfrac{L}{2}\right), \quad x \in (0, L) \\ w(0) = w(L) = 0 \\ EI\dfrac{d^2}{dx^2}w(0) = EI\dfrac{d^2}{dx^2}w(L) = 0 \end{cases}. \tag{2.183}$$

(b) *Thus, the original problem reduces to solving (2.183) subject to (2.182). Interpret the conditions (2.182) and (2.183) in the framework of analyzing statically indeterminate structures via linear superposition of primary structures. Explain the physical meaning of $\lambda$.*

(c) *Determine $\lambda$ for the case $q = 0$.*

# Chapter 3

# Source Inversion

## 3.1 Inverse problem

In this chapter, we consider an image deblurring problem as an illustrative example of an inverse problem. Although the forward operator is not a partial differential equation, the problem shares the essential features of typical inverse problems: ill-posedness, the need for regularization, and the challenge of parameter selection. The presentation follows [Ghattas, 2015] and [Vogel, 2002] closely.

Consider the following abstract problem:

$$d = Kp. \tag{3.1}$$

Here, $K$ is an operator mapping $p$ to $d$, and typically represents a physical process, often expressed as a differential equation. The variable $d$ denotes the *data*, *measurements*, or *observables*. We classify the associated problems as follows:

- *Forward problem*: given $K$ and $p$, find $d$.
- *Inverse problem, Type A*: given $K$ and $d$, find $p$
  (e.g., seismic inversion, deblurring problem, and control problem).
- *Inverse problem, Type B*: given $p$ and $d$, find $K$
  (e.g., imaging, system identification, and design problem).

In the inverse problem of Type A, $p$ is the unknown to be reconstructed and is called the *parameter field*; it is typically a function or, after discretization, a finite-dimensional vector. If the inverse operator $K^{-1}$ exists and is readily available, the problem reduces to $p = K^{-1}d$, which is simply a forward solve and is not the focus of these notes. The inverse problems of interest are *ill-posed*, meaning that at least one of the following conditions of well-posedness, due to Hadamard, fails:

- *Existence*: for all data $d$, there exists a parameter $p$ satisfying $d = Kp$.
- *Uniqueness*: for all data $d$, the solution $p$ is unique.

- *Stability* (continuous dependence on data): small perturbations in $d$ result in small perturbations in $p$. Specifically, let $p$ and $q$ be parameters corresponding to data $d$ and $f$, respectively, i.e., $d = Kp$ and $f = Kq$. Stability requires the existence of a constant $C > 0$ such that

$$\|p - q\|_* \leq C\|d - f\|_{**}. \tag{3.2}$$

## 3.2 Deblurring problem

As a model problem, we begin with image deblurring in one dimension. Let $p\colon [0,1] \to \mathbb{R}$ denote the unknown (true) image and $d\colon [0,1] \to \mathbb{R}$ the observed, blurred, and noisy data. The forward model is a Fredholm integral equation of the first kind,

$$\begin{aligned} d(x) &= (Kp)(x) + n(x) \\ &\coloneqq \int_0^1 k(x - x')p(x')\,dx' + n(x), \qquad x \in [0,1], \end{aligned} \tag{3.3}$$

where $n$ denotes additive measurement noise and the convolution kernel $k$ is the Gaussian

$$k(x) \coloneqq C \exp\left(-\frac{x^2}{2\gamma^2}\right), \qquad C,\, \gamma > 0. \tag{3.4}$$

The parameter $\gamma$ controls the strength of blur: a larger $\gamma$ yields a wider kernel, so each output value averages $p$ over a broader neighborhood. Equivalently, the action of $K$ can be examined in the frequency domain. Since the Fourier transform of a Gaussian with standard deviation $\gamma$ is a Gaussian with standard deviation $1/\gamma$, a larger $\gamma$ narrows the transform of $k$ in frequency, thereby attenuating high-frequency components of $p$ more aggressively. The operator $K$ thus acts as a low-pass filter, and $\gamma$ governs the cutoff (Figure 3.1).

Fourier transform of a Gaussian $k(x)$ is another Gaussian:

$$\begin{aligned} &Fk(\omega) \\ &\coloneqq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} k(x)\, e^{i\omega x} dx \\ &= \gamma \exp\left(-\frac{\gamma^2\omega^2}{2}\right), \end{aligned}$$

where $k(x) = \exp\left[-x^2/(2\gamma^2)\right]$.

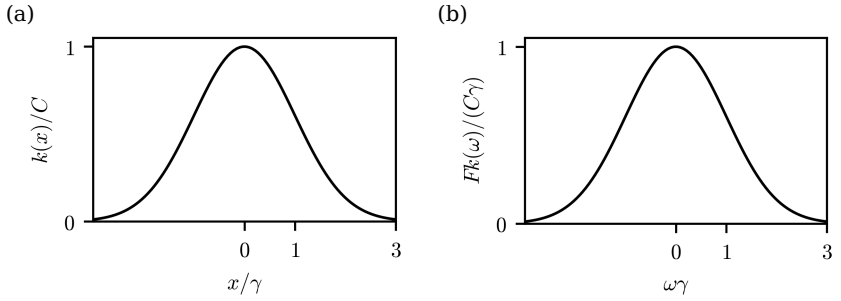



Figure 3.1: (a) Gaussian $k(x) = C \exp\left[-x^2/(2\gamma^2)\right]$ and (b) its Fourier transform $Fk(\omega) = C\gamma \exp\left[-\gamma^2\omega^2/2\right]$.

Next, we examine which of the three well-posedness conditions fails for the deblurring problem (3.3). The key structural property of the forward operator $K$ is compactness.

**Definition 3.2.1 (Compact operator)** *Let $X$ and $Y$ be Banach spaces. A bounded linear operator $K\colon X \to Y$ is* compact *if every bounded sequence $\{p_n\} \subset X$ has a subsequence whose image under $K$ converges in $Y$. Equivalently, $K$ maps bounded sets in $X$ to precompact sets in $Y$.*

Namely, the range space of $K$ is smaller than the $Y$.

A standard class of compact operators is provided by integral operators with square-integrable kernels.

**Definition 3.2.2 (Hilbert-Schmidt operator)** *Let $k \in L^2((0,1)\times(0,1))$. The integral operator*

$$(Kp)(x) \coloneqq \int_0^1 k(x,x')p(x')\,dx' \tag{3.5}$$

*is called a* Hilbert-Schmidt operator. *Every Hilbert-Schmidt operator is compact from $L^2(0,1)$ to $L^2(0,1)$.*

Since the Gaussian kernel (3.26) is smooth and bounded on $[0,1]\times[0,1]$, it belongs to $L^2((0,1)\times(0,1))$, and thus $K$ in (3.3) is a compact operator on $L^2(0,1)$.

**Lemma 3.2.1 (Riemann-Lebesgue)** *Let $f \in L^1(\mathbb{R})$. Then its Fourier transform vanishes at infinity:*

$$\hat{f}(\omega) \coloneqq \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx \to 0 \qquad \text{as } |\omega| \to \infty. \tag{3.6}$$

Now, we check the well-posedness of the given operator $K$:

- Existence: Since $K$ is compact, its range $\mathrm{R}(K)$ is not closed in $L^2(0,1)$. Consequently, there exist data $d \in L^2(0,1)$ for which $Kp = d$ has no solution; existence fails in general.
- Uniqueness: The Gaussian kernel (3.26) is strictly positive, and $K$ is injective on $L^2(0,1)$, i.e., $N(K) = \{0\}$. Hence, if a solution exists, it is unique; uniqueness holds.
- Stability: Consider a perturbed problem $K(p+\delta p) = d + \delta d$, where $\delta p$ is the high-frequency perturbation

  $$\delta p(x) \coloneqq \varepsilon \sin(\omega x), \qquad \varepsilon > 0, \quad \omega = 1\cdot 2\pi, 2\cdot 2\pi, 3\cdot 2\pi, \ldots. \tag{3.7}$$

  The corresponding perturbation in the data is

  $$\delta d(x) = \varepsilon \int_0^1 k(x-x')\sin(\omega x')\,dx'. \tag{3.8}$$

  By the Riemann-Lebesgue lemma (Lemma 3.2.1), $\|\delta d\|_{L^2} \to 0$ as $\omega \to \infty$, while $\|\delta p\|_{L^2} = \varepsilon/\sqrt{2}$ remains fixed. Therefore,

  $$\frac{\|\delta p\|_{L^2}}{\|\delta d\|_{L^2}} \to \infty \qquad \text{as } \omega \to \infty, \tag{3.9}$$

  and no finite constant $C$ can satisfy the stability condition. Arbitrarily small perturbations in the data thus correspond to arbitrarily large errors in the recovered parameter.

For numerical purposes, we discretize the problem $d = Kp+n$ to $\mathbf{d} = \mathbf{K}\mathbf{p}+\mathbf{n}$, where $\mathbf{d}, \mathbf{p}, \mathbf{n} \in \mathbb{R}^N$, $\mathbf{K} \in \mathbb{R}^{N\times N}$, and

$$K_{ij} = hC \exp\left(-\frac{(i-j)^2 h^2}{2\gamma^2}\right), \qquad 1 \le i, j \le N. \tag{3.10}$$

Here, $h = 1/N$ is the mesh spacing, obtained by partitioning $[0, 1]$ into $N$ uniform subintervals and applying the midpoint quadrature rule to the integral (3.3). Figure 3.2 illustrates the action of the Gaussian convolution operator $K$.

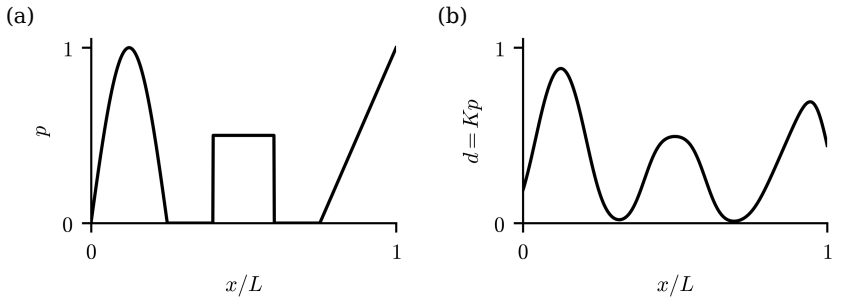



Figure 3.2: Gaussian blur with $C = 1/(\gamma\sqrt{2\pi})$, $\gamma = 0.04$. (a) Original image. (b) Blurred image.

Note that the discretized problem is stable; however, the stability constant grows as $N$ increases. Equivalently, $\mathbf{K}$ becomes increasingly *ill-conditioned* as $N \to \infty$, so that a small noise vector $\mathbf{n}$ can result in large errors in the recovered $\mathbf{p}$. In this sense, the ill-posedness of the original infinite-dimensional problem is reflected in the ill-conditioning of its finite-dimensional discretization, with the *condition number* of $\mathbf{K}$ growing without bound as $N \to \infty$.

**Definition 3.2.3 (Condition number)** *Let $\mathbf{K} \in \mathbb{R}^{N\times N}$ be a nonsingular matrix. The* condition number *of $\mathbf{K}$ is defined as*

$$\kappa(\mathbf{K}) \coloneqq \|\mathbf{K}\|\|\mathbf{K}^{-1}\|, \tag{3.11}$$

*where $\|\cdot\|$ denotes a matrix norm. When the spectral norm $\|\mathbf{K}\|_2 \coloneqq \sigma_{\max}(\mathbf{K})$ is used, the condition number takes the form*

$$\kappa_2(\mathbf{K}) = \frac{\sigma_{\max}(\mathbf{K})}{\sigma_{\min}(\mathbf{K})}, \tag{3.12}$$

*where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values of $\mathbf{K}$, respectively. A matrix with large $\kappa(\mathbf{K})$ is called* ill-conditioned.

This instability has a natural interpretation: the Gaussian kernel acts as a low-pass filter, attenuating high-frequency components of $p$ at a rate controlled by $\gamma$ in (3.26). Inversion requires amplifying these attenuated components, which amplifies noise catastrophically.

**Definition 3.2.4 (Singular Value Decomposition)** *Let $\mathbb{F}$ denote either $\mathbb{R}$ or $\mathbb{C}$, and let $\mathbf{K} \in \mathbb{F}^{m\times n}$ with $p \coloneqq \min\{m,n\}$. There exist unitary matrices*

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{F}^{m\times m}, \qquad \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{F}^{n\times n}, \tag{3.13}$$

*and a rectangular diagonal matrix $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, such that*

$$\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger = \sum_{i=1}^{p} \sigma_i \, \mathbf{u}_i \mathbf{v}_i^\dagger. \tag{3.14}$$

*The $\sigma_i$ are the* singular values *of $\mathbf{K}$; the columns $\mathbf{u}_i$ and $\mathbf{v}_i$ are the* left *and* right singular vectors*, respectively. When $\mathbb{F} = \mathbb{R}$, unitarity reduces to orthogonality and $(\cdot)^\dagger$ reduces to $(\cdot)^T$.*

Unlike the eigendecomposition, which requires $\mathbf{K}$ to be square and may yield complex eigenvalues even for real matrices, the SVD exists for any $\mathbf{K} \in \mathbb{F}^{m\times n}$ and its singular values are always real and nonnegative. For a symmetric positive definite (SPD) matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, the eigenvalues are real and positive, the SVD coincides with the eigendecomposition, and $\sigma_i = \lambda_i$, $\mathbf{U} = \mathbf{V}$. The discrete convolution matrix (3.10) is SPD, so these simplifications apply throughout.

The SVD also reveals the four fundamental subspaces of $\mathbf{K}$. Let $r \leq p$ denote the rank of $\mathbf{K}$, and partition

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_r & \mathbf{U}_0\end{bmatrix}, \qquad \mathbf{V} = \begin{bmatrix}\mathbf{V}_r & \mathbf{V}_0\end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \tag{3.15}$$

where $\mathbf{U}_r \coloneqq [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $\mathbf{U}_0 \coloneqq [\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m]$, $\mathbf{V}_r \coloneqq [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, $\mathbf{V}_0 \coloneqq [\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n]$, and $\mathbf{\Sigma}_r \coloneqq \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$. Then $\mathbf{K} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\dagger$, and the four fundamental subspaces are

$$\begin{aligned} \mathrm{R}(\mathbf{K}) &= \operatorname{col}(\mathbf{U}_r) \subset \mathbb{F}^m, & \mathrm{N}(\mathbf{K}) &= \operatorname{col}(\mathbf{V}_0) \subset \mathbb{F}^n, \\ \mathrm{R}(\mathbf{K}^\dagger) &= \operatorname{col}(\mathbf{V}_r) \subset \mathbb{F}^n, & \mathrm{N}(\mathbf{K}^\dagger) &= \operatorname{col}(\mathbf{U}_0) \subset \mathbb{F}^m. \end{aligned} \tag{3.16}$$

The columns of $\mathbf{U}$ form an orthonormal basis for $\mathbb{F}^m$ and the columns of $\mathbf{V}$ form an orthonormal basis for $\mathbb{F}^n$. The action of $\mathbf{K}$ is completely characterized: it maps $\mathbf{v}_i \mapsto \sigma_i \mathbf{u}_i$ for $i \leq r$ and annihilates $\mathbf{v}_i$ for $i > r$.

In the context of the deblurring problem, $\mathbf{K} \in \mathbb{R}^{N\times N}$ is SPD and hence invertible, so $r = N$, $\mathbf{U} = \mathbf{V}$, and the null space is trivial; however, the singular vectors $\mathbf{v}_i$ associated with small $\sigma_i$ correspond to high-frequency modes that are nearly annihilated by $\mathbf{K}$, making their recovery from noisy data unreliable.

## 3.3 SVD-based approach

Since the discrete convolution matrix $\mathbf{K}$ is SPD, the SVD coincides with the eigendecomposition and $\mathbf{U} = \mathbf{V}$. The naive solution to the inverse problem is

$$\begin{aligned}\mathbf{p} &= \mathbf{K}^{-1}\mathbf{d} \\ &= \mathbf{U}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T\mathbf{d} \\ &= \sum_{i=1}^{N} \sigma_i^{-1} \left(\mathbf{u}_i^T\mathbf{d}\right) \mathbf{u}_i \\ &= \mathbf{p}_{\text{true}} + \sum_{i=1}^{N} \sigma_i^{-1} \left(\mathbf{u}_i^T\mathbf{n}\right) \mathbf{u}_i. \end{aligned} \tag{3.17}$$

The last equation shows that the noise contribution $\mathbf{u}_i^T\mathbf{n}$ is amplified by $\sigma_i^{-1}$. Since the singular values decay rapidly (Figure 3.3), the high-frequency components of $\mathbf{n}$, which correspond to small $\sigma_i$, are amplified catastrophically, rendering the naive inversion unstable.

The remedy is to suppress the contributions from small singular values. The *truncated singular value decomposition* (TSVD) replaces the exact inverse by the filtered expansion

$$\mathbf{p} = \sum_{i=1}^{N} \Omega_{\text{TSVD}}\left(\sigma_i^2\right) \sigma_i^{-1} \left(\mathbf{u}_i^T\mathbf{d}\right) \mathbf{u}_i, \tag{3.18}$$

where the filter function $\Omega_{\text{TSVD}}$ is defined by

$$\Omega_{\text{TSVD}}\left(\sigma^2\right) \coloneqq \begin{cases} 1, & \sigma^2 \geq \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{3.19}$$

and $\alpha > 0$ is the regularization parameter controlling the truncation threshold. Terms corresponding to $\sigma_i^2 < \alpha$ are discarded, eliminating the catastrophic amplification of high-frequency noise at the cost of losing the corresponding signal components.

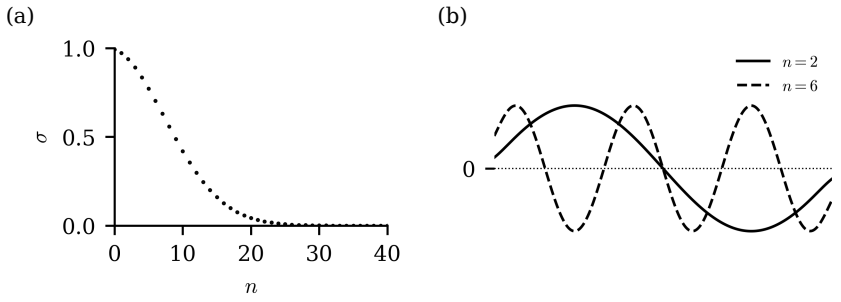



Figure 3.3: Singular value decomposition of $\mathbf{K}$. (a) Singular values. (b) Selected singular vectors.

## 3.4 Tikhonov regularization

Alternatively, the inverse problem can be formulated as an optimization problem. Given the noisy data $d^{\text{obs}} \coloneqq Kp_{\text{true}} + n$, find $p$ such that

$$\min_p \; J(p) \coloneqq \frac{1}{2} \left\| Kp - d^{\text{obs}} \right\|^2 + \frac{\alpha}{2} \|p\|^2 , \tag{3.20}$$

where $\alpha > 0$ is the regularization parameter and $\|\cdot\|$ denotes the $L^2(0,1)$ norm,

$$\|a\|^2 = (a,a) \coloneqq \int_0^1 a(x)^2 \, dx. \tag{3.21}$$

The first term penalizes the data misfit. The second term is called *Tikhonov regularization*, which penalizes the $L^2$ norm of $p$, with $\alpha$ controlling the trade-off between fidelity to the data and regularity of the reconstruction.

The minimizer is characterized by the stationary point condition $DJ(p)[\tilde{p}] = 0$ for all $\tilde{p} \in L^2(0,1)$. Since $\tilde{p}$ is arbitrary, the Euler equation reads

$$\left(K^\dagger K + \alpha \mathcal{I}\right) p = K^\dagger d^{\text{obs}}, \tag{3.22}$$

where $K^\dagger$ is the $L^2$-adjoint of $K$, defined by $(Kp, q) = (p, K^\dagger q)$ for all $p, q \in L^2(0,1)$, and $\mathcal{I}$ is the identity operator. Since $K$ is compact and $\mathcal{I}$ is not, the operator $K^\dagger K + \alpha \mathcal{I}$ is boundedly invertible for any $\alpha > 0$; this is precisely the regularizing effect of the Tikhonov term. For the convolution operator (3.3) with the symmetric Gaussian kernel $k(x) = k(-x)$, $K$ is self-adjoint, i.e., $K^\dagger = K$.

Discretizing the Euler equation and solving for $\mathbf{p}$ gives

$$\left(\mathbf{K}^T\mathbf{K} + \alpha \mathbf{I}\right) \mathbf{p} = \mathbf{K}^T \mathbf{d}^{\text{obs}}, \quad \text{i.e.,} \quad \mathbf{p} = \left(\mathbf{K}^T\mathbf{K} + \alpha \mathbf{I}\right)^{-1} \mathbf{K}^T \mathbf{d}^{\text{obs}}. \tag{3.23}$$

Substituting the SVD $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ (valid since $\mathbf{K}$ is SPD) and using $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, we obtain

$$\begin{aligned} \mathbf{p} &= \left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T + \alpha \mathbf{U}\mathbf{U}^T\right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T \mathbf{d}^{\text{obs}} \\ &= \mathbf{U}\left(\mathbf{\Sigma}^2 + \alpha \mathbf{I}\right)^{-1} \mathbf{\Sigma}\, \mathbf{U}^T \mathbf{d}^{\text{obs}} \\ &= \sum_{i=1}^{N} \frac{\sigma_i}{\sigma_i^2 + \alpha} \left(\mathbf{u}_i^T \mathbf{d}^{\text{obs}}\right) \mathbf{u}_i. \end{aligned} \tag{3.24}$$

This is equivalent to the filtered expansion (3.18) with the *Tikhonov filter function*

$$\Omega_{\text{TIK}}\left(\sigma^2\right) \coloneqq \frac{\sigma^2}{\sigma^2 + \alpha}, \tag{3.25}$$

which satisfies $\Omega_{\text{TIK}}(\sigma^2) \to 1$ as $\sigma^2 \gg \alpha$ and $\Omega_{\text{TIK}}(\sigma^2) \to 0$ as $\sigma^2 \ll \alpha$, and thus smoothly attenuates the contributions from small singular values. An important practical advantage of Tikhonov regularization over TSVD is that the solution $\mathbf{p}$ can be computed without explicit knowledge of the SVD of $\mathbf{K}$, since the discrete Euler equation is a linear system that can be solved by standard methods.

**Exercise 3.4.1** *Show that the Gaussian blur*

$$Kp(x) = \int_0^1 k(x-x')p(x')dx', \quad \textit{where}$$
$$k(x) = C\exp\left(-\frac{x^2}{2\gamma^2}\right), \quad C,\, \gamma > 0 \tag{3.26}$$

*is self-adjoint.*

## 3.5 Regularization parameter choice

The choice of $\alpha$ is critical: too small a value yields an unstable reconstruction, while too large a value suppresses signal and degrades accuracy. We discuss *a posteriori* strategies for selecting $\alpha$, both of which require solving the regularized problem over a sequence of parameter values.

Let $\mathbf{p}_\alpha$ denote the filtered solution, either by TSVD or Tikhonov regularization, with regularization parameter $\alpha$, and let $\mathbf{K}_\alpha$ denote the corresponding filtered blur operator such that

$$\mathbf{K}_\alpha^{-1}\mathbf{d} \coloneqq \mathbf{p}_\alpha = \sum_{i=1}^{N} \Omega_\alpha\left(\sigma_i^2\right)\sigma_i^{-1}\left(\mathbf{u}_i^T\mathbf{d}\right)\mathbf{u}_i. \tag{3.27}$$

The reconstruction error is defined as

$$\begin{aligned}\mathbf{e}_\alpha &\coloneqq \mathbf{p}_\alpha - \mathbf{p}_{\text{true}} \\ &= \mathbf{K}_\alpha^{-1}(\mathbf{K}\mathbf{p}_{\text{true}} + \mathbf{n}) - \mathbf{p}_{\text{true}},\end{aligned} \tag{3.28}$$

and is decomposed into a truncation error $\mathbf{e}_\alpha^{\text{trunc}}$ and a noise amplification error $\mathbf{e}_\alpha^{\text{noise}}$:

$$\mathbf{e}_\alpha = \mathbf{e}_\alpha^{\text{trunc}} + \mathbf{e}_\alpha^{\text{noise}}, \tag{3.29}$$

where

$$\begin{aligned}\mathbf{e}_\alpha^{\text{trunc}} &\coloneqq \left(\mathbf{K}_\alpha^{-1}\mathbf{K} - \mathbf{I}\right)\mathbf{p}_{\text{true}} \\ &= \sum_{i=1}^{N}\left(\Omega_\alpha\left(\sigma_i^2\right) - 1\right)\left(\mathbf{u}_i^T\mathbf{p}_{\text{true}}\right)\mathbf{u}_i,\end{aligned} \tag{3.30}$$

$$\begin{aligned}\mathbf{e}_\alpha^{\text{noise}} &\coloneqq \mathbf{K}_\alpha^{-1}\mathbf{n} \\ &= \sum_{i=1}^{N}\Omega_\alpha\left(\sigma_i^2\right)\sigma_i^{-1}\left(\mathbf{u}_i^T\mathbf{n}\right)\mathbf{u}_i.\end{aligned} \tag{3.31}$$

We show that both errors converge to zero as the noise level

$$\delta \coloneqq \|\mathbf{n}\| \tag{3.32}$$

goes to zero, for an appropriate choice of $\alpha$.

For both TSVD and Tikhonov regularization,

$$\Omega_\alpha\left(\sigma^2\right) \to 1 \quad \text{as} \quad \alpha \to 0, \tag{3.33}$$

which immediately implies

$$\left\|\mathbf{e}_\alpha^{\text{trunc}}\right\| \to 0 \quad \text{as} \quad \alpha \to 0. \tag{3.34}$$

For the noise amplification error, define the noise amplification factor $A_\alpha(\sigma) \coloneqq \Omega_\alpha(\sigma^2)\,\sigma^{-1}$. We claim that

$$A_\alpha(\sigma) \leq \frac{1}{\sqrt{\alpha}} \tag{3.35}$$

for all $\sigma > 0$ and $\alpha > 0$.

- *TSVD.* If $\sigma^2 < \alpha$, then $\Omega_\alpha(\sigma^2) = 0$ and the bound holds trivially. If $\sigma^2 \geq \alpha$, then $\Omega_\alpha(\sigma^2) = 1$ and $\sigma \geq \sqrt{\alpha}$, so

$$A_\alpha(\sigma) = \sigma^{-1} \leq \frac{1}{\sqrt{\alpha}}. \tag{3.36}$$

- *Tikhonov.* We seek the supremum of $A_\alpha(\sigma) = \sigma/(\sigma^2+\alpha)$ over $\sigma > 0$. The stationarity condition reads

$$\frac{dA_\alpha}{d\sigma} = \frac{(\sigma^2+\alpha) - 2\sigma^2}{(\sigma^2+\alpha)^2} = \frac{\alpha - \sigma^2}{(\sigma^2+\alpha)^2} = 0, \tag{3.37}$$

  yielding the unique critical point $\sigma^* = \sqrt{\alpha}$, which is a global maximum. Evaluating,

$$A_\alpha(\sigma^*) = \frac{\sqrt{\alpha}}{2\alpha} = \frac{1}{2\sqrt{\alpha}} \leq \frac{1}{\sqrt{\alpha}}. \tag{3.38}$$

Applying this bound and the orthonormality of $\mathbf{U}$, we obtain

$$\left\|\mathbf{e}_\alpha^{\text{noise}}\right\| \leq \frac{1}{\sqrt{\alpha}} \left\| \sum_{i=1}^{N} \left(\mathbf{u}_i^T \mathbf{n}\right) \mathbf{u}_i \right\| = \frac{\delta}{\sqrt{\alpha}}. \tag{3.39}$$

Thus, for $\alpha = \delta^p$ with $0 < p < 2$,

$$\left\|\mathbf{e}_\alpha^{\text{noise}}\right\| \to 0 \quad \text{as} \quad \delta \to 0. \tag{3.40}$$

Combining both estimates, the choice $\alpha = \delta^p$ with $0 < p < 2$ guarantees $\|\mathbf{e}_\alpha\| \to 0$ as $\delta \to 0$.

### 3.5.1 Discrepancy (Morozov) principle

The discrepancy principle selects the largest $\alpha$ such that the data misfit does not exceed the noise level $\delta \coloneqq \|\mathbf{n}\|$. Defining $M(\alpha) \coloneqq \|\mathbf{K}\mathbf{p}_\alpha - \mathbf{d}\|$, the criterion requires

$$M(\alpha) \leq \delta. \tag{3.41}$$

Thus, the above criterion avoids overfitting, i.e., fitting the noise rather than the signal.

We can show that for Tikhonov regularization such an $\alpha$ always exists, provided $\delta < \|\mathbf{d}\|$. Using the eigendecomposition $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, one obtains

$$\mathbf{K}\mathbf{p}_\alpha - \mathbf{d} = \sum_{i=1}^{N} \left( \frac{\sigma_i^2}{\sigma_i^2 + \alpha} - 1 \right) \left( \mathbf{u}_i^T \mathbf{d} \right) \mathbf{u}_i, \tag{3.42}$$

and by orthonormality of $\mathbf{U}$,

$$M^2(\alpha) = \sum_{i=1}^{N} \left( \frac{\sigma_i^2}{\sigma_i^2 + \alpha} - 1 \right)^2 \left( \mathbf{u}_i^T \mathbf{d} \right)^2. \tag{3.43}$$

This expression shows that $M$ is continuous and monotonically increasing in $\alpha$, with $M(0) = 0$ and $M(\alpha) \to \|\mathbf{d}\|$ as $\alpha \to \infty$. Hence, provided $\delta \|\mathbf{d}\|$, the intermediate value theorem guarantees the existence of $\alpha$ such that $M(\alpha) = \delta$.

### 3.5.2 L-curve criterion

The L-curve criterion requires no knowledge of the noise level $\delta$. For each $\alpha$ in a prescribed sequence, one solves for $\mathbf{p}_\alpha$ and plots $\|\mathbf{K}\mathbf{p}_\alpha - \mathbf{d}\|$ against $\|\mathbf{p}_\alpha\|$ on a log-log scale. The resulting curve is typically L-shaped, and the selected $\alpha$ corresponds to the point of maximum curvature. This choice balances data fidelity against solution stability: for $\alpha$ below the optimal value, the data misfit decreases only marginally while $\|\mathbf{p}_\alpha\|$ grows substantially; for $\alpha$ above it, $\|\mathbf{p}_\alpha\|$ decreases only marginally while the misfit increases substantially.

Figure 3.4 shows a deblurring example using Tikhonov regularization with varying $\alpha$. Both the discrepancy principle and the L-curve criterion suggest $\alpha \approx 10^{-2}$, which yields a reasonable reconstruction (panel (g)). For $\alpha = 10^{-3}$, the regularization is too weak and the reconstruction is dominated by noise amplification (panel (e)). Regularization with $\alpha = 10^{0}$ over-smooths the reconstruction (panel (h)).

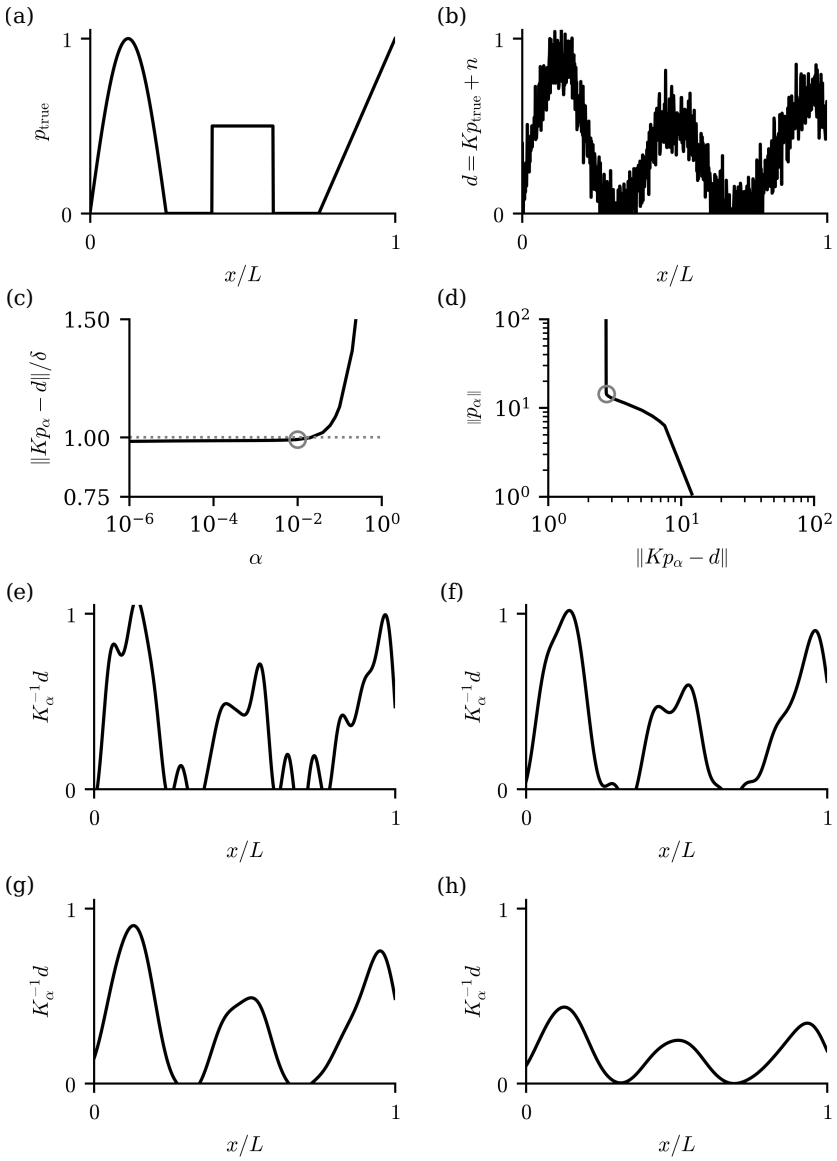



Figure 3.4: Deblurring example. (a) Original image. (b) Blurred image with noise. (c) Discrepancy plot. (d) L-curve plot. (e) Deblurred image with $\alpha = 1.0 \times 10^{-3}$. (f) Deblurred image with $\alpha = 1.0 \times 10^{-2}$. (g) Deblurred image with $\alpha = 1.0 \times 10^{-1}$. (h) Deblurred image with $\alpha = 1.0 \times 10^{0}$.

## 3.6 Image deblurring in two dimension

Consider a 2D image of standard definition, i.e., $N_x \times N_y = 720 \times 480$ pixels, giving $N = N_x N_y = 345{,}600$ unknowns. The 2D blurring operator $\mathbf{K}$ is an $N \times N$ matrix with $N^2 \approx 1.19 \times 10^{11}$ entries. At 8 bytes per entry (64-bit

floating point), explicit storage requires

$$8 \times N^2 \approx 8 \times 1.19 \times 10^{11} \approx 9.5 \times 10^{11} \text{ bytes} \approx 953 \text{ GB}, \tag{3.44}$$

which is infeasible on standard hardware. Moreover, solving the Tikhonov normal equations

$$\left(\mathbf{K}^T\mathbf{K} + \alpha\mathbf{I}\right)\mathbf{p} = \mathbf{K}^T\mathbf{d} \tag{3.45}$$

via Gaussian elimination requires $\mathcal{O}(N^3)$ floating point operations. For $N =$ 345,600, this amounts to

$$N^3 \approx 4.13 \times 10^{16} \text{ flops}, \tag{3.46}$$

which is computationally intractable. Both difficulties are resolved by exploiting the separable structure of the 2D Gaussian kernel: the action of $\mathbf{K}$ on a vector can be computed as successive 1D convolutions in the $x$- and $y$-directions. This enables the use of iterative solvers such as the conjugate gradient method, which requires only matrix-vector operation.

The conjugate gradient method solves the symmetric positive definite system

$$\mathbf{A}\mathbf{p} = \mathbf{b}, \qquad \mathbf{A} \coloneqq \mathbf{K}^T\mathbf{K} + \alpha\mathbf{I}, \qquad \mathbf{b} \coloneqq \mathbf{K}^T\mathbf{d}, \tag{3.47}$$

iteratively without forming $\mathbf{A}$ explicitly. Starting from an initial guess $\mathbf{p}^{(0)}$, each iteration requires one matrix-vector product with $\mathbf{A}$. The algorithm is as follows [Quarteroni et al., 2006]:

**Algorithm 1** Conjugate gradient method for $\mathbf{A}\mathbf{x} = \mathbf{b}$

1: Set initial value $\mathbf{x}^{(0)}$
2: $\mathbf{r}^{(0)} \coloneqq \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ ▷ initial residual
3: $\mathbf{p}^{(0)} \coloneqq \mathbf{r}^{(0)}$ ▷ initial search direction
4: **for** $k = 0, 1, 2, \ldots$ **do**
5: $\quad \alpha^{(k)} \coloneqq \dfrac{\|\mathbf{r}^{(k)}\|^2}{\left(\mathbf{p}^{(k)}\right)^T \mathbf{A}\,\mathbf{p}^{(k)}}$ ▷ step length
6: $\quad \mathbf{x}^{(k+1)} \coloneqq \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$ ▷ update iterate
7: $\quad \mathbf{r}^{(k+1)} \coloneqq \mathbf{r}^{(k)} - \alpha^{(k)}\mathbf{A}\mathbf{p}^{(k)}$ ▷ update residual
8: $\quad$ **if** $\|\mathbf{r}^{(k+1)}\| \leq \varepsilon \|\mathbf{r}^{(0)}\|$ **then**
9: $\quad\quad$ **break** ▷ convergence criterion
10: $\quad$ **end if**
11: $\quad \beta^{(k)} \coloneqq -\dfrac{\|\mathbf{r}^{(k+1)}\|^2}{\|\mathbf{r}^{(k)}\|^2}$ ▷ direction update coefficient
12: $\quad \mathbf{p}^{(k+1)} \coloneqq \mathbf{r}^{(k+1)} - \beta^{(k)}\mathbf{p}^{(k)}$ ▷ update search direction
13: **end for**
14: **return** $\mathbf{x}^{(k+1)}$

Figures 3.5 illustrates two-dimensional image deblurring. The original image is convolved with a Gaussian kernel and contaminated with additive noise. Tikhonov regularization recovers the essential features of the original image.

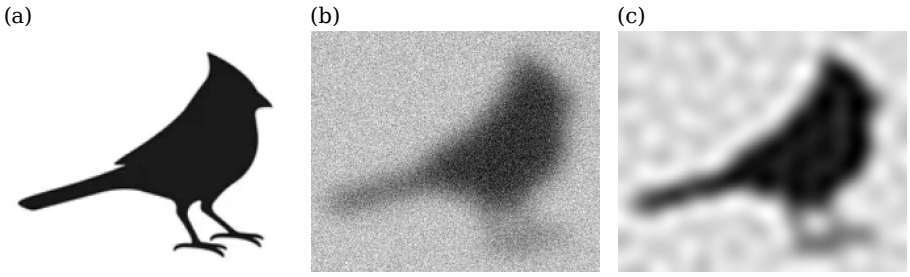



Figure 3.5: Two-dimensional image deblurring with a binary image. (a) Original image (Northern cardinal silhouette). (b) Blurred and noised image. (c) Reconstructed image via Tikhonov regularization.

# Chapter 4

# Parameter Field Inversion

## 4.1 Synopsis

We are interested in recovering system parameter fields from given input and output data, i.e., a *parameter field inversion*. As a representative example, we consider an *inverse medium problem*, in which we seek the spatial distribution of material properties, such as the Lamé parameters, mass density, thermal conductivity, or dielectric permittivity, of a physical system from boundary or interior measurements of its response.

Problems of this type arise across a wide range of disciplines. In geophysics, seismic waveform inversion recovers the elastic moduli and density of subsurface formations from surface recordings of reflected and refracted waves. In medical imaging, electrical impedance tomography seeks the conductivity distribution of biological tissue from boundary voltage measurements. In structural engineering, non-destructive evaluation identifies damage or material degradation in load-bearing components by inverting vibration or guided-wave measurements for the local stiffness field.

Despite the diversity of the underlying physics, each of these problems shares the similar mathematical structure. The inverse problem is formulated as a minimization of a *misfit functional*, which quantifies the discrepancy between the measured data and the response predicted from a trial parameter field. The state field must satisfy a governing partial differential equation (PDE), which enters the formulation as a constraint enforced through a Lagrange multiplier. The resulting problem is therefore classified as a *PDE-constrained optimization* problem.

As in the previous example of image deblurring, the inverse medium problem requires the solution of a nonlinear optimization problem that, at best, exploits gradient information. Computing the derivative of the misfit functional with respect to the parameter field is therefore central to any practical inversion algorithm. Four broad strategies are available.

- **Direct sensitivity method.** One differentiates the governing equation with respect to the parameter field to obtain a *sensitivity equation* whose solution yields the required gradient analytically.
- **Finite differencing.** Each component of the gradient is approximated by perturbing the corresponding parameter entry and re-solving the forward

problem. The implementation is straightforward, but the cost again scales with the parameter dimension and the approximation is susceptible to truncation and round-off errors.

- **Adjoint method.** An *adjoint equation*, derived from the optimality conditions of the Lagrangian, is solved once to obtain the gradient with respect to the entire parameter field. The computational cost is therefore independent of the parameter dimension, making the adjoint method the workhorse of large-scale PDE-constrained optimization.
- **Automatic differentiation.** The chain rule is applied systematically to the discrete computational graph of the forward solver. In reverse mode, automatic differentiation produces gradients at a cost comparable to that of the adjoint method, although with larger memory overhead; it may be viewed as an algorithmic implementation of the adjoint principle applied at the discrete level.

In the deblurring problem, the forward operator was a linear convolution and the misfit functional was quadratic in the unknown image; the gradient was therefore available in closed form by direct differentiation. For the PDE-governed problems treated in the remainder of this chapter, the parameter dimension is large and the forward map is defined only implicitly through the solution of a differential equation. We accordingly develop the adjoint method in detail.

## 4.2 Elliptic problem

### 4.2.1 Problem definition

We begin the discussion with a simple model elliptic problem. Given measured data $u_d$ and a source term $f$, find the parameter field $p$ by solving

$$\min_p J[u,p], \quad J[u,p] \coloneqq \underbrace{\frac{1}{2}(u-u_d, u-u_d)}_{=M[u]} + \underbrace{\frac{\alpha}{2}(\operatorname{grad} p, \operatorname{grad} p)}_{=R[p]}, \tag{4.1}$$

subject to the governing equation and the boundary condition

$$\begin{cases} \operatorname{div}(p \operatorname{grad} u) + f = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ p \operatorname{grad} u \cdot n = 0 & \text{on } \Gamma_N \end{cases} \tag{4.2}$$

where $\Gamma_D \cup \Gamma_N = \partial\Omega$ and the parameter field $p$ belongs to the admissible set

$$W \coloneqq \left\{ p \in H^1(\Omega) \ : \ p \geq p_o > 0 \text{ in } \Omega \right\}. \tag{4.3}$$

Recall

$$(a,b) = (a,b)_\Omega = \int_\Omega ab \, d\Omega,$$

$$(a,b)_\Gamma = \int_\Gamma ab \, d\Gamma,$$

where $\Gamma = \partial\Omega$.

Here, the functional $M[u]$ is called the misfit functional and $R[p]$ is the regularization on the parameter field $p$. Specifically, we use the *first-order Tikhonov regularization*:

$$R[p] = \frac{\alpha}{2}(\operatorname{grad} p, \operatorname{grad} p) = \frac{\alpha}{2}\int_\Omega |\operatorname{grad} p|^2 \, d\Omega, \tag{4.4}$$

which penalizes oscillations in $p$ while leaving the mean value unpenalized: the null space of grad consists of constant functions, so spatially uniform shifts incur no cost.

### 4.2.2 Lagrangian and necessary condition

We enforce the governing equation as a constraint using a Lagrange multiplier (also called *adjoint variable*) $v$, which yields the Lagrangian functional

$$L[u, v, p] = M[u] + R[p] + E[u, v, p]. \tag{4.5}$$

The constraint $E[u, v, p]$ is the weak form of (4.2) tested against $v$,

$$E[u, v, p] = (p \operatorname{grad} u, \operatorname{grad} v) - (f, v). \tag{4.6}$$

Here, we have a symmetric functional setting for $u \in U$ and $v \in V$ such that

$$U = V = \left\{ u \in H^1(\Omega) \ : \ u = 0 \text{ on } \Gamma_D \right\}. \tag{4.7}$$

The first-order necessary conditions (KKT conditions) require that the variations of the Lagrangian with respect to all three fields vanish:

$$d_u L[u, v, p](\tilde{u}) = 0, \quad \forall\, \tilde{u} \in U, \tag{4.8a}$$
$$d_p L[u, v, p](\tilde{p}) = 0, \quad \forall\, \tilde{p} \in W. \tag{4.8b}$$
$$d_v L[u, v, p](\tilde{v}) = 0, \quad \forall\, \tilde{v} \in V, \tag{4.8c}$$

The third necessary condition (4.8c) leads to the constraint, or the *state problem*:

$$0 = (p \operatorname{grad} u, \operatorname{grad} \tilde{v}) - (f, \tilde{v}). \tag{4.9}$$

The first necessary condition (4.8a) gives the *adjoint problem*:

$$0 = (\tilde{u}, u - u_d) + (p \operatorname{grad} \tilde{u}, \operatorname{grad} v). \tag{4.10}$$

The bilinear form $(p \operatorname{grad} u, \operatorname{grad} v)$ is symmetric in $u$ and $v$, so the state and adjoint operators coincide. Notice that the adjoint problem is driven by the misfit $u - u_d$: if the state field were to match the data exactly, the right-hand side would vanish and the adjoint variable would be trivially zero. The corresponding strong form, i.e., the Euler equation, reads

$$\begin{cases} \operatorname{div}(p \operatorname{grad} v) + u - u_d = 0 & \text{in } \Omega, \\ v = 0 & \text{on } \Gamma_D, \\ p \operatorname{grad} v \cdot n = 0 & \text{on } \Gamma_N. \end{cases} \tag{4.11}$$

Lastly, the second necessary condition (4.8b) gives the *control problem*:

$$0 = \alpha (\operatorname{grad} \tilde{p}, \operatorname{grad} p) + (\tilde{p} \operatorname{grad} u, \operatorname{grad} v). \tag{4.12}$$

### 4.2.3 Gradient-based optimization and the adjoint method

The three necessary conditions are coupled and, once discretized, form a system that is too large to store and factorize directly. We therefore adopt an iterative algorithm to find the optimal $p$ starting from an initial guess $p^{(0)}$. At each iteration, the current estimate $p^{(k)}$ is updated by

$$p^{(k+1)} = p^{(k)} + s d^{(k)}, \tag{4.13}$$

where $d^{(k)}$ is the *search direction* and $s$ is the step length. Several choices for the search direction are available: setting $d^{(k)}$ to the negative gradient yields the *steepest descent method*; incorporating information from previous gradient gives the *nonlinear conjugate gradient method*; and retaining a limited history of gradients leads to the *L-BFGS method*. All of these methods require the gradient of the objective functional $J$ with respect to the parameter field $p$.

The gradient can be approximated numerically by finite differencing, i.e.,

$$d_p J[\ldots, p](\tilde{p}) \approx \frac{J[\ldots, p + \epsilon\tilde{p}] - J[\ldots, p]}{\epsilon}. \tag{4.14}$$

In addition to the choice of a suitably small $\epsilon$, this computation is problematic because it requires $N + 1$ functional evaluations for $N$ degrees of freedom in $p$. Alternatively, the adjoint method obtains the gradient by a sequential evaluation of the KKT conditions. Given a current estimate $p^{(k)}$, the gradient $g^{(k)}$ is obtained as follows.

1. Solve the state problem for $u^{(k)}$:

$$0 = \left(p^{(k)} \operatorname{grad} u^{(k)}, \operatorname{grad} \tilde{v}\right) - \left(f, \tilde{v}\right), \quad \forall \tilde{v} \in V. \tag{4.15}$$

2. Solve the adjoint problem for $v^{(k)}$:

$$0 = \left(p^{(k)} \operatorname{grad} \tilde{u}, \operatorname{grad} v^{(k)}\right) + \left(\tilde{u}, u^{(k)} - u_d\right), \quad \forall \tilde{u} \in U. \tag{4.16}$$

3. Evaluate the gradient equation for $g^{(k)}$:

$$\left(g^{(k)}, \tilde{p}\right) = \alpha \left(\operatorname{grad} \tilde{p}, \operatorname{grad} p^{(k)}\right) + \left(\tilde{p} \operatorname{grad} u^{(k)}, \operatorname{grad} v^{(k)}\right), \quad \forall \tilde{p} \in W. \tag{4.17}$$

Notice that the right-hand side of (4.17) is in general nonzero away from the optimum; it vanishes precisely when the KKT conditions (4.8) are satisfied. Strictly speaking, the right-hand side defines a linear functional on $W$, i.e., an element of the dual space $W^*$. The function $g^{(k)} \in W$ is its Riesz representation with respect to the $L^2$ inner product, i.e.,

$$\begin{cases} g^{(k)} = -\alpha \operatorname{div} \operatorname{grad} p^{(k)} + \operatorname{grad} u^{(k)} \cdot \operatorname{grad} v^{(k)}, & x \in \Omega \\ \tilde{p} = 0 \text{ or } \operatorname{grad} p^{(k)} = 0, & x \in \Gamma. \end{cases} \tag{4.18}$$

Upon finite element discretization, this amounts to solving a mass-matrix system $\mathbf{M}\mathbf{g}^{(k)} = \mathbf{r}^{(k)}$, where $\mathbf{r}^{(k)}$ is the assembled right-hand-side vector.

Overall, the adjoint method computes the gradient at a cost independent of the parameter dimension, and the resulting expressions are exact at the continuous level; approximation enters only through the finite element discretization of the function spaces.

**Algorithm 2** Steepest descent method

Choose initial guess $p^{(0)} \in W$ and tolerance $\varepsilon > 0$
Compute gradient direction $g^{(0)}$ via (4.17)
**for** $k = 0, 1, 2, \ldots$ **do**
    **if** $\left\| g^{(k)} \right\| \leq \varepsilon$ **then**
        **break** ▷ convergence
    **end if**
    $d^{(k)} := -g^{(k)}$ ▷ search direction
    Determine step length $s^{(k)} > 0$
    $p^{(k+1)} := p^{(k)} + s^{(k)}\, d^{(k)}$ ▷ update parameter field
    Compute gradient direction $g^{(k+1)}$ via (4.17)
**end for**
**return** $p^{(k)}$

**Algorithm 3** Nonlinear conjugate gradient method

Choose initial guess $p^{(0)} \in W$ and tolerance $\varepsilon > 0$
Compute gradient direction $g^{(0)}$ via (4.17)
$d^{(0)} := -g^{(0)}$ ▷ initial search direction
**for** $k = 0, 1, 2, \ldots$ **do**
    **if** $\left\| g^{(k)} \right\| \leq \varepsilon$ **then**
        **break** ▷ convergence
    **end if**
    Determine step length $s^{(k)} > 0$
    $p^{(k+1)} := p^{(k)} + s^{(k)}\, d^{(k)}$ ▷ update parameter field
    Compute gradient direction $g^{(k+1)}$ via (4.17)
    $\beta^{(k+1)} := \dfrac{\left(g^{(k+1)},\, g^{(k+1)}\right)}{\left(g^{(k)},\, g^{(k)}\right)}$ ▷ Fletcher–Reeves
    $d^{(k+1)} := -g^{(k+1)} + \beta^{(k+1)}\, d^{(k)}$ ▷ update search direction
**end for**
**return** $p^{(k)}$

### 4.2.4 Step length criteria

The gradient determines the search direction, but the step length remains to be chosen. Given a descent direction $d^{(k)}$, a line search seeks a step length $s > 0$ such that the update $p^{(k+1)} = p^{(k)} + sd^{(k)}$ produces a sufficient decrease in the objective. A naive choice such as exact minimization,

$$s = \arg\min_{s>0} J[p^{(k)} + sd^{(k)}], \tag{4.19}$$

is generally impractical for PDE-constrained problems because each evaluation of $J$ requires the solution of the state equation. Instead, one enforces *inexact* line search conditions that guarantee global convergence while requiring only a modest number of functional evaluations per iteration.

Figure 4.1 illustrates the effect of the step length on convergence. A near-optimal step length yields monotone decrease toward the minimizer, whereas an excessively large step length overshoots, causing the iterates to oscillate about the solution.

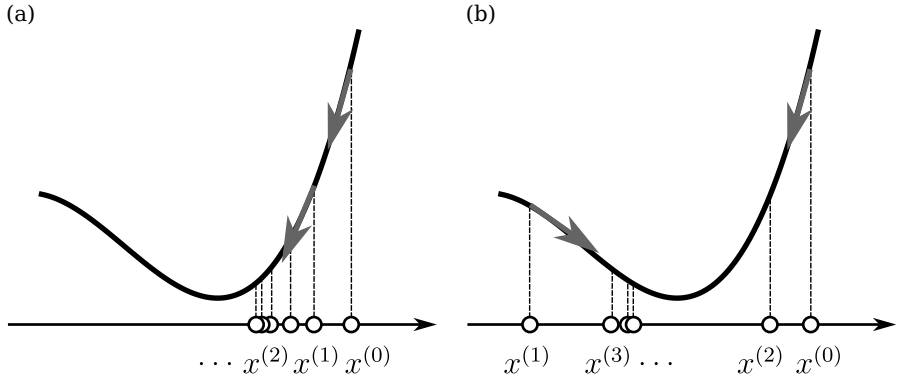



Figure 4.1: Effect of step length on steepest descent. (a) Near-optimal step length: the iterates converge monotonically. (b) Excessively large step length: the iterates overshoot and oscillate.

#### Armijo condition (sufficient decrease of objective)

The step length $s$ satisfies the *Armijo condition* if

$$J[p^{(k)} + sd^{(k)}] \leq J[p^{(k)}] + c_1 s \left(g^{(k)}, d^{(k)}\right), \tag{4.20}$$

where $c_1 \in (0, 1)$ is a tolerance parameter (typically $c_1 = 10^{-4}$). The right-hand side is the first-order Taylor expansion of $J$ along $d^{(k)}$, scaled by $c_1$; the condition requires the actual reduction to be at least a fraction $c_1$ of the predicted linear decrease.

The Armijo condition alone does not prevent $s$ from being arbitrarily small. In practice, one combines it with a *backtracking* strategy [Quarteroni et al., 2006]: starting from an initial trial step $\bar{s}$, the step is repeatedly contracted by a factor $\rho \in (0, 1)$,

$$s \leftarrow \rho s, \tag{4.21}$$

until (4.20) is satisfied (See Algorithm 4). Each backtracking step requires one evaluation of $J$, and hence one solve of the state equation. The backtracking Armijo method is simple, inexpensive per trial step, and widely used in PDE-constrained optimization.

**Algorithm 4** Backtracking line search (Armijo)

**Require:** Descent direction $d^{(k)}$, current objective value $J^{(k)} \coloneqq J[p^{(k)}]$, directional derivative $\delta \coloneqq \left(g^{(k)}, d^{(k)}\right) < 0$

**Require:** Parameters $\bar{s} > 0$ (initial step), $\rho \in (0,1)$ (contraction factor), $c_1 \in (0,1)$ (sufficient decrease parameter)

1: $s \coloneqq \bar{s}$
2: **while** $J[p^{(k)} + sd^{(k)}] > J^{(k)} + c_1 s\delta$ **do**
3: $\quad s \coloneqq \rho s$ ▷ contract step length
4: **end while**
5: **return** $s$

**Wolfe conditions (sufficient decrease of objective and its slope)**

The *Wolfe conditions* augment (4.20) with a *curvature condition* that prevents the accepted step from being too short:

$$J[p^{(k)} + sd^{(k)}] \leq J[p^{(k)}] + c_1 s \left(g^{(k)}, d^{(k)}\right), \tag{4.22}$$

$$\left(g^{(k+1)}, d^{(k)}\right) \geq c_2 \left(g^{(k)}, d^{(k)}\right), \tag{4.23}$$

where $0 < c_1 < c_2 < 1$ (typical choices: $c_1 = 10^{-4}$, $c_2 = 0.9$ for Newton and quasi-Newton methods, $c_2 = 0.1$ for nonlinear conjugate gradient). The curvature condition (4.23) requires the directional derivative at the new iterate to be less negative than at the current iterate, ensuring that the step is not accepted in a region where the objective is still steeply decreasing. A more restrictive variant, the *strong Wolfe condition*, replaces (4.23) with

$$\left|\left(g^{(k+1)}, d^{(k)}\right)\right| \leq c_2 \left|\left(g^{(k)}, d^{(k)}\right)\right|, \tag{4.24}$$

which additionally excludes points where the directional derivative is large and positive.

### 4.2.5 Newton method

Gradient-based algorithms, such as descent methods, approximate a functional by a linear model at each iteration around the current iterate. The resulting linear model determines a descent direction but not a step length, since a linear functional is unbounded below along any descent direction. Consequently, a separate line-search or step-length selection procedure is required.

In contrast, Newton's method approximates the functional by a quadratic model at each iteration (Figure 4.2). Under appropriate conditions, this quadratic model admits a unique minimizer that is analytically computable, yielding both the search direction and the step length simultaneously.

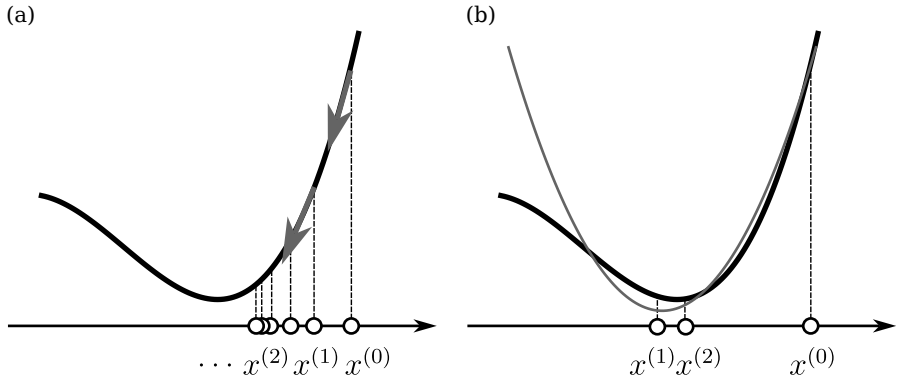



Figure 4.2: Comparison of steepest descent and Newton's method. (a) Steepest descent provides only a search direction; the step length must be determined by a separate line search. (b) Newton's method determines both the search direction and a natural step length from the local quadratic model of the objective.

Constructing such a quadratic model requires both the first and second derivatives of the functional. Suppose the stationary condition for a minimization problem is

$$d_p J[p](\tilde{p}) = 0, \quad \forall\, \tilde{p}. \tag{4.25}$$

Given a current iterate $p^{(k)}$, we seek an update $\hat{p}$ such that $p^{(k+1)} = p^{(k)} + \hat{p}$. Linearizing the stationary condition about $p^{(k)}$ via a first-order Taylor expansion yields

$$\begin{aligned} 0 &= d_p J[p^{(k)} + \hat{p}](\tilde{p}) \\ &\approx d_p J[p^{(k)}](\tilde{p}) + d_p^2 J[p^{(k)}](\hat{p}, \tilde{p}), \quad \forall \tilde{p}. \end{aligned} \tag{4.26}$$

Thus, the update $\hat{p}$, which encodes both the search direction and the step length, is obtained by solving the above linear problem.

We now apply the approach in the KKT condition (4.8). Letting $[\ldots]$ denote $[u, v, p]$, we have

$$0 = d_u L\,[\ldots]\,(\tilde{u}) + d_u d_u L\,[\ldots]\,(\tilde{u}, \hat{u}) + d_p d_u L\,[\ldots]\,(\tilde{u}, \hat{p}) + d_v d_u L\,[\ldots]\,(\tilde{u}, \hat{v})\,, \tag{4.27a}$$

$$0 = d_p L\,[\ldots]\,(\tilde{p}) + d_u d_p L\,[\ldots]\,(\tilde{p}, \hat{u}) + d_p d_p L\,[\ldots]\,(\tilde{p}, \hat{p}) + d_v d_p L\,[\ldots]\,(\tilde{p}, \hat{v})\,, \tag{4.27b}$$

$$0 = d_v L\,[\ldots]\,(\tilde{v}) + d_u d_v L\,[\ldots]\,(\tilde{v}, \hat{u}) + d_p d_v L\,[\ldots]\,(\tilde{v}, \hat{p}) + d_v d_v L\,[\ldots]\,(\tilde{v}, \hat{v})\,. \tag{4.27c}$$

For the model problem, the above equations read ($\nabla = \text{grad}\,$)

$$\begin{aligned} \left(\tilde{u}, \hat{u}^{(k)}\right) &+ \left(\hat{p}^{(k)} \nabla \tilde{u}, \nabla v^{(k)}\right) + \left(p^{(k)} \nabla \tilde{u}, \nabla \hat{v}^{(k)}\right) &&= -\left(p^{(k)} \nabla \tilde{u}, \nabla v^{(k)}\right) - \left(\tilde{u}, u^{(k)} - u_d\right), \\ \left(\tilde{p}\, \nabla \hat{u}^{(k)}, \nabla v^{(k)}\right) &+ \alpha \left(\nabla \tilde{p}, \nabla \hat{p}^{(k)}\right) + \left(\tilde{p}\, \nabla u^{(k)}, \nabla \hat{v}^{(k)}\right) &&= -\alpha \left(\nabla \tilde{p}, \nabla p^{(k)}\right) - \left(\tilde{p}\, \nabla u^{(k)}, \nabla v^{(k)}\right), \\ \left(p^{(k)} \nabla \hat{u}^{(k)}, \nabla \tilde{v}\right) &+ \left(\hat{p}^{(k)} \nabla u^{(k)}, \nabla \tilde{v}\right) &&= -\left(p^{(k)} \nabla u^{(k)}, \nabla \tilde{v}\right) + (f, \tilde{v})\,. \end{aligned} \tag{4.28}$$

Removing the dependence on the test functions $\tilde{u}$, $\tilde{p}$, $\tilde{v}$, the above system can be written as

$$\begin{bmatrix} F & A^\dagger & B^\dagger \\ A & D & C^\dagger \\ B & C & 0 \end{bmatrix} \begin{pmatrix} \hat{u}^{(k)} \\ \hat{p}^{(k)} \\ \hat{v}^{(k)} \end{pmatrix} = - \begin{pmatrix} g_u \\ g_p \\ g_v \end{pmatrix}. \tag{4.29}$$

We would never attempt to store and factorize the matrix in (4.29). Instead, we solve the state and adjoint problems sequentially so that, for a given $p^{(k)}$, we find $u^{(k)}$ and $v^{(k)}$ satisfying $g_u = g_v = 0$. The first and last block rows can then be condensed out, since:

$$B\hat{u}^{(k)} = -C\hat{p}^{(k)}, \tag{4.30}$$

$$B^\dagger \hat{v}^{(k)} = -F\hat{u}^{(k)} - A^\dagger \hat{p}^{(k)}. \tag{4.31}$$

Substituting into the second block row yields the reduced Newton system

$$H\hat{p}^{(k)} = -g_p, \tag{4.32}$$

where

$$H = D + C^\dagger B^{-\dagger}\left(FB^{-1}C - A^\dagger\right) - AB^{-1}C. \tag{4.33}$$

The action of the reduced Hessian on a trial direction $\hat{p}^{(k)}$ is therefore computed sequentially as follows.

1. Solve the incremental state problem for $\hat{u}^{(k)}$:

$$B\hat{u}^{(k)} = -C\hat{p}^{(k)}. \tag{4.34}$$

2. Solve the incremental adjoint problem for $\hat{v}^{(k)}$:

$$B^\dagger \hat{v}^{(k)} = -F\hat{u}^{(k)} - A^\dagger \hat{p}^{(k)}. \tag{4.35}$$

3. Evaluate the reduced Hessian action:

$$H\hat{p}^{(k)} = A\hat{u}^{(k)} + D\hat{p}^{(k)} + C^\dagger \hat{v}^{(k)}. \tag{4.36}$$

The Newton method requires the reduced Hessian $H$ to be positive definite, which is not guaranteed away from the solution. Two strategies are commonly used to address this difficulty.

The first is the *Gauss-Newton method*, in which the second-order adjoint terms involving $A$ and $A^\dagger$ are dropped from $H$. The resulting approximation

$$H^{\mathrm{GN}} = D + C^\dagger B^{-\dagger} F\, B^{-1}C \tag{4.37}$$

is symmetric positive definite by construction, so the conjugate gradient (CG) method can always be applied. The trade-off is that quadratic convergence is lost; Gauss–Newton converges superlinearly or linearly, depending on the size of the residual at the solution.

The second is the *truncated Newton method*, in which one solves the full Newton system with CG but terminates early if a direction of non-positive curvature is encountered, i.e., if $d^\top H d \leq 0$ for some CG iterate $d$. The truncated step is then used within a trust-region framework to ensure global convergence. This approach retains the full Hessian information when it is beneficial and effortlessly degrades to a gradient or Gauss-Newton-like step when $H$ is indefinite.

Algorithm 5 summarizes the truncated Newton-CG procedure for solving the discretized reduced Hessian system $\mathbf{H}\hat{\mathbf{p}} = -\mathbf{g}_p$.

In practice, both the Gauss–Newton and truncated Newton methods are combined with a backtracking line search using an initial step length of 1, since neither method solves the Newton system exactly.

**Algorithm 5** Truncated Newton-CG method for $\mathbf{H}\hat{\mathbf{p}} = -\mathbf{g}_p$

Set initial value $\hat{\mathbf{p}}^{(0)}$
$\mathbf{r}^{(0)} \coloneqq -\mathbf{g}_p - \mathbf{H}\hat{\mathbf{p}}^{(0)}$ ▷ initial residual
$\mathbf{d}^{(0)} \coloneqq \mathbf{r}^{(0)}$ ▷ initial search direction
**for** $k = 0, 1, 2, \ldots$ **do**
  $\kappa^{(k)} \coloneqq \left(\mathbf{d}^{(k)}\right)^\dagger \mathbf{H}\,\mathbf{d}^{(k)}$ ▷ curvature along search direction
  **if** $\kappa^{(k)} \leq 0$ **then**
    **return** $\hat{\mathbf{p}}^{(k)}$ ▷ negative curvature detected; truncate
  **end if**
  $\alpha^{(k)} \coloneqq \dfrac{\|\mathbf{r}^{(k)}\|^2}{\kappa^{(k)}}$ ▷ step length
  $\hat{\mathbf{p}}^{(k+1)} \coloneqq \hat{\mathbf{p}}^{(k)} + \alpha^{(k)}\mathbf{d}^{(k)}$ ▷ update iterate
  $\mathbf{r}^{(k+1)} \coloneqq \mathbf{r}^{(k)} - \alpha^{(k)}\mathbf{H}\,\mathbf{d}^{(k)}$ ▷ update residual
  **if** $\|\mathbf{r}^{(k+1)}\| \leq \varepsilon\,\|\mathbf{r}^{(0)}\|$ **then**
    **return** $\hat{\mathbf{p}}^{(k+1)}$ ▷ convergence criterion
  **end if**
  $\beta^{(k)} \coloneqq -\dfrac{\|\mathbf{r}^{(k+1)}\|^2}{\|\mathbf{r}^{(k)}\|^2}$ ▷ direction update coefficient
  $\mathbf{d}^{(k+1)} \coloneqq \mathbf{r}^{(k+1)} - \beta^{(k)}\mathbf{d}^{(k)}$ ▷ update search direction
**end for**
**return** $\hat{\mathbf{p}}^{(k+1)}$

**Exercise 4.2.1** *Let $B$ be invertible, $F = F^\dagger \succeq 0$, $D = D^\dagger \succ 0$, and let $A$, $C$ be arbitrary matrices of compatible dimensions. Define*

$$H \coloneqq D + C^\dagger B^{-\dagger}(FB^{-1}C - A^\dagger) - AB^{-1}C, \tag{4.38}$$

$$H^{\mathrm{GN}} \coloneqq D + C^\dagger B^{-\dagger} F\, B^{-1}C. \tag{4.39}$$

*Show that $H$ is symmetric and $H^{\mathrm{GN}}$ is symmetric positive definite.*

## 4.3 Hyperbolic problem

In seismic source reconstruction or subsurface imaging, the governing PDE is in general a wave equation. For example, the governing PDE of interest is

$$\begin{cases} \operatorname{div}\left[\mu \operatorname{grad} u\right] - \rho\dfrac{\partial^2 u}{\partial t^2} + f = 0, & (x,t) \in \Omega \times J, \\ u = 0, & (x,t) \in \Omega \times \{0\}, \\ \dfrac{\partial u}{\partial t} = 0, & (x,t) \in \Omega \times \{0\}, \\ u = 0, & (x,t) \in \Gamma_D \times J, \\ \mu \operatorname{grad} u \cdot n = 0, & (x,t) \in \Gamma_N \times J, \end{cases} \tag{4.40}$$

where $J \coloneqq (0, T]$ is the time interval. The group variable $\mathtt{p} = (\mu, \rho)$ denotes the material parameters, and $v$ is the adjoint variable. The Lagrangian reads

$$L\left[u, v, \mathtt{p}\right] = M\left[u\right] + R\left[\mathtt{p}\right] + E\left[u, v, \mathtt{p}\right], \tag{4.41}$$

where the individual terms are

$$M[u] = \frac{1}{2}\int_0^T (u-u_d, u-u_d)\,dt, \tag{4.42}$$

$$R[\mathtt{p}] = \frac{\alpha_\mu}{2}(\operatorname{grad}\mu, \operatorname{grad}\mu) + \frac{\alpha_\rho}{2}(\operatorname{grad}\rho, \operatorname{grad}\rho), \tag{4.43}$$

$$E[u,v,\mathtt{p}] = \int_0^T \left[\left(\rho\frac{\partial^2 u}{\partial t^2}, v\right) + b(u,v) - l(v)\right] dt. \tag{4.44}$$

Here,

$$b(u,v) = \int_\Omega \mu \operatorname{grad} u \cdot \operatorname{grad} v d\Omega \quad \text{and} \tag{4.45}$$

$$l(v) = \int_\Omega f v d\Omega. \tag{4.46}$$

Then, the adjoint method of finding the gradients are

1. Given $\mathtt{p}$, solve the state problem for $u$:

$$\begin{cases} u|_{t \text{ fixed}} \in U,\ u|_{t=0} = 0,\ \left.\dfrac{\partial u}{\partial t}\right|_{t=0} = 0, \\ \left(\rho\dfrac{\partial^2 u}{\partial t^2}, \tilde{v}\right) + b(u,\tilde{v}) = l(\tilde{v}), \quad \forall\ \tilde{v}|_{t \text{ fixed}} \in V,\ \forall t \in J, \end{cases} \tag{4.47}$$

   where $U = V \coloneqq \left\{w \in H^1(\Omega) : w = 0 \text{ on } \Gamma_D\right\}$.

2. Given $\mathtt{p}$ and $u$, solve the adjoint problem for $v$:

$$\begin{cases} v|_{t \text{ fixed}} \in U,\ v|_{t=T} = 0,\ \left.\dfrac{\partial v}{\partial t}\right|_{t=T} = 0, \\ \left(\rho\dfrac{\partial^2 v}{\partial t^2}, \tilde{u}\right) + b(v,\tilde{u}) = -(u-u_d, \tilde{u}), \quad \forall\ \tilde{u}|_{t \text{ fixed}} \in V,\ \forall t \in J^\dagger, \end{cases} \tag{4.48}$$

   where $b(\cdot,\cdot)$ is symmetric and $J^\dagger = [0,T)$.

3. Given $\mathtt{p}$, $u$, and $v$, evaluate the gradient equations for $g_\mu$ and $g_\rho$:

$$\begin{aligned}(\tilde{\mu}, g_\mu) = {} & \alpha_\mu(\operatorname{grad}\tilde{\mu}, \operatorname{grad}\mu) \\ & + \int_0^T (\tilde{\mu}\operatorname{grad} u, \operatorname{grad} v)\,dt, \quad \forall \tilde{\mu} \in W, \end{aligned} \tag{4.49}$$

$$(\tilde{\rho}, g_\rho) = \alpha_\rho(\operatorname{grad}\tilde{\rho}, \operatorname{grad}\rho) + \int_0^T \left(\tilde{\rho}\frac{\partial^2 u}{\partial t^2}, v\right) dt, \quad \forall \tilde{\rho} \in W, \tag{4.50}$$

**Exercise 4.3.1** *Derive the Newton system for the Lagrangian* (4.41)*, and state the corresponding incremental problems whose solutions define the Hessian action.*

## 4.4 Eigenvalue problem

We consider an optimization problem governed by a generalized eigenvalue problem. The state variable is the eigenpair $(u, \lambda)$, where $u$ is the eigenfunction and $\lambda$ the eigenvalue, satisfying

$$\begin{cases} u \in U \setminus \{0\},\ \lambda \in \mathbb{R} \\ a_0(u, v) + \lambda\, a_1(u, v) = 0, \quad \forall v \in U. \end{cases} \tag{4.51}$$

We assume that the sesquilinear forms $a_0$ and $a_1$ are Hermitian, i.e., $a_i(v, u) = \overline{a_i(u, v)}$ for $i = 0, 1$, and depend on a design parameter $p$.

Introduce the grouped state variable $\mathtt{u} \coloneqq (u, \lambda)$ and the grouped adjoint variable $\mathtt{v} \coloneqq (v, \xi)$, where $v \in U$ is the adjoint eigenfunction and $\xi$ is the adjoint eigenvalue. The Lagrangian reads

$$L[\mathtt{u}, \mathtt{v}, p] = M[\mathtt{u}] + R[p] + E[\mathtt{u}, \mathtt{v}, p], \tag{4.52}$$

where $M$ is the objective functional, $R$ is the regularization, and the constraint functional $E$ is given by

$$E[\mathtt{u}, \mathtt{v}, p] = \mathrm{Re}\left\{a_0(u, v) + \lambda\, a_1(u, v)\right\} + \frac{\xi}{2}\left[a_1(u, u) - 1\right]. \tag{4.53}$$

The first two terms enforce the eigenvalue equation through the adjoint variable $v$, while the last term enforces the normalization constraint through the multiplier $\xi$. The normalization $a_1(u, u) = 1$ fixes the magnitude of $u$ but leaves a residual phase (or sign) ambiguity, which is inconsequential when the objective functional is invariant under $u \mapsto e^{i\theta} u$.

The adjoint problem reads: Given $p$ and $(u, \lambda)$, find $(v, \xi) \in U \times \mathbb{R}$ such that

$$\begin{cases} a_0(\tilde{u}, v) + \lambda\, a_1(\tilde{u}, v) + \xi\, a_1(\tilde{u}, u) = -d_u M[\mathtt{u}](\tilde{u}), & \forall \tilde{u} \in U, \\ \tilde{\lambda}\, a_1(u, v) = -d_\lambda M[\mathtt{u}]\left(\tilde{\lambda}\right), & \forall \tilde{\lambda} \in \mathbb{R}. \end{cases} \tag{4.54}$$

Since $\tilde{u}$ and $\tilde{\lambda}$ are arbitrary, the adjoint system (4.54) admits the operator representation

$$\begin{pmatrix} A_0 + \lambda A_1 & A_1 u \\ u^\dagger A_1 & 0 \end{pmatrix} \begin{pmatrix} v \\ \xi \end{pmatrix} = \begin{pmatrix} -\overline{\nabla_u M} \\ -\overline{\partial_\lambda M} \end{pmatrix}, \tag{4.55}$$

where $A_0$ and $A_1$ are the operators induced by $a_0$ and $a_1$, respectively, and $\nabla_u M$ denotes the gradient of $M$ with respect to $u$. The coefficient operator is self-adjoint, reflecting the Hermitian symmetry of the underlying sesquilinear forms.

The formulation extends to general polynomial eigenvalue problems:

$$\begin{cases} u \in U \setminus \{0\},\ \lambda \in \mathbb{C} \\ \displaystyle\sum_{k=0}^{N} \lambda^k a_k(u, v) = 0, \quad \forall v \in U, \end{cases} \tag{4.56}$$

where the sesquilinear forms $a_k$ are assumed Hermitian. A normalization that preserves the Hermitian structure of the adjoint problem is

$$\sum_{k=1}^{N} k \lambda^{k-1} a_k(u, u) = 1. \tag{4.57}$$

The constraint functional $E$ is then given by

$$E[\mathtt{u}, \mathtt{v}, p] = \mathrm{Re}\left\{\sum_{k=0}^{N} \lambda^k a_k(u, v)\right\} + \frac{\xi}{2}\left[\sum_{k=1}^{N} k\lambda^{k-1} a_k(u, u) - 1\right]. \tag{4.58}$$

Define the polynomial operator and its derivative with respect to $\lambda$:

$$P(\lambda) \coloneqq \sum_{k=0}^{N} \lambda^k A_k, \qquad P'(\lambda) \coloneqq \sum_{k=1}^{N} k\lambda^{k-1} A_k. \tag{4.59}$$

Taking variations of the Lagrangian with respect to $\tilde{u}$ and $\tilde{\lambda}$ and invoking the Hermitian symmetry of each $a_k$, the adjoint system admits the operator representation

$$\begin{pmatrix} P(\lambda) & P'(\lambda)\, u \\ u^{\dagger} P'(\lambda) & 0 \end{pmatrix} \begin{pmatrix} v \\ \xi \end{pmatrix} = \begin{pmatrix} -\overline{\nabla_u M} \\ -\overline{\partial_\lambda M} \end{pmatrix}. \tag{4.60}$$

The coefficient operator is self-adjoint, since $P(\lambda)$ and $P'(\lambda)$ are both self-adjoint and the $(1,2)$ block $P'(\lambda)\,u$ is the conjugate transpose of the $(2,1)$ block $u^{\dagger}P'(\lambda)$.

# Chapter 5

# Refinements

## 5.1 Partial observations and data filtering

### 5.1.1 Partial observations

In most physical applications, measurements are accessible only on a portion of the domain boundary; non-destructive and non-invasive testing are representative examples in which direct internal access to the medium is either impractical or inadmissible. This situation arises in seismic imaging, where receivers are confined to the free surface or to sparse boreholes; in electrical impedance tomography, where currents and voltages are measured only through surface electrodes; and in inverse heat conduction, where temperature is sensed on the accessible face of a body while an inaccessible boundary condition is reconstructed. The restriction may equally apply in time, when only a portion of the recorded history is reliable or relevant, as with the early arrivals of a seismic record. We model these restrictions by introducing a weight $w(x,t)$ into the misfit functional. For the model hyperbolic problem (4.41) this yields

$$M[u] \coloneqq \frac{1}{2}\int_0^T \int_\Omega w(u-u_d)^2 d\Omega dt. \tag{5.1}$$

When data are available only on an accessible subregion $\Omega_{\text{obs}} \subseteq \Omega$ (or, for boundary measurements, on $\Gamma_{\text{obs}} \subseteq \partial\Omega$), the weight is one on that region and zero elsewhere, independent of time,

$$w(x,t) = \begin{cases} 1, & x \in \Omega_{\text{obs}}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

For pointwise receiver data at locations $x_j$, $j = 1, \dots, N_{\text{obs}}$, the weight reduces to a sum of Dirac delta functions,

$$w(x,t) = \sum_{j=1}^{N_{\text{obs}}} \delta(x - x_j). \tag{5.3}$$

A restriction in time enters through the dependence on $t$. To retain only a time interval $[t_1, t_2]$, for instance the early arrivals,

$$w(x,t) = \begin{cases} 1, & t_1 \le t \le t_2, \\ 0, & \text{otherwise}, \end{cases} \qquad 0 \le t_1 < t_2 \le T. \tag{5.4}$$

The corresponding adjoint problem reads: Given $\mathtt{p}$ and $u$, solve the adjoint problem for $v$:

$$\begin{cases} v|_{t \text{ fixed}} \in U, \ v|_{t=T} = 0, \ \left.\dfrac{\partial v}{\partial t}\right|_{t=T} = 0, \\ \left(\rho \dfrac{\partial^2 v}{\partial t^2}, \tilde{u}\right) + b\left(v, \tilde{u}\right) = -\left((u - u_d)w, \tilde{u}\right), \quad \forall\ \tilde{u}|_{t \text{ fixed}} \in V, \ \forall t \in J^\dagger. \end{cases} \tag{5.5}$$

### 5.1.2 Data filtering

The weights of the previous subsection act pointwise on the residual. A second class of filters acts non-locally in time, through a convolution, and cannot be expressed as a pointwise weight. Applied identically to the measured data $u_d$ and the reconstructed data $u$, hence to the residual $r \coloneqq u - u_d$, such filters serve for adaptive noise reduction or to emphasize a particular frequency component for sequential use in multiscale inversion, such as frequency continuation or band-pass selection. We write the filtered misfit in terms of a linear filter operator $\mathcal{F}$,

$$M[u] \coloneqq \frac{1}{2} \int_0^T \int_\Omega (\mathcal{F}r)^2 d\Omega dt. \tag{5.6}$$

For example, a Gaussian low-pass filter with bandwidth parameter $\omega_k$ can be used for frequency continuation:

$$(\mathcal{F}r)(x,t) = \int_0^T g(t-s) r(x,s) ds, \tag{5.7}$$

where

$$g(t) = \frac{\omega_k}{\sqrt{2\pi}} e^{-\omega_k^2 t^2/2}. \tag{5.8}$$

Here, the kernel normalized so that $\int_{-\infty}^{\infty} g(t) = 1$. A narrow band $\omega_k$ corresponds to a wide kernel and strong temporal smoothing. A continuation strategy begins with a small $\omega_1$, admitting only the low-frequency content of the residual, and increases the bandwidth $\omega_1 < \omega_2 < \cdots$ across stages, each initialized from the previous reconstruction, recovering the smooth background before the high-frequency detail. As $\omega_k \to \infty$ the kernel tends to a Dirac delta function and the unfiltered misfit is recovered.

The corresponding adjoint problem differs from (5.5) only in the source, which is now filtered by $\mathcal{F}^\dagger\mathcal{F}$: Given $\mathtt{p}$ and $u$, solve the adjoint problem for $v$ such that

$$\begin{cases} v|_{t \text{ fixed}} \in U, \ v|_{t=T} = 0, \ \left.\dfrac{\partial v}{\partial t}\right|_{t=T} = 0, \\ \left(\rho \dfrac{\partial^2 v}{\partial t^2}, \tilde{u}\right) + b\left(v, \tilde{u}\right) = -\left(\mathcal{F}^\dagger\mathcal{F}(u - u_d), \tilde{u}\right), \ \forall\ \tilde{u}|_{t \text{ fixed}} \in V, \ \forall t \in J^\dagger. \end{cases} \tag{5.9}$$

Here $\mathcal{F}^\dagger$ is the adjoint of $\mathcal{F}$ in time, defined at each $x$ by

$$\int_0^T (\mathcal{F}r)(x,t)q(x,t)dt = \int_0^T r(x,t)(\mathcal{F}^\dagger q)(x,t)dt. \tag{5.10}$$

Interchanging the order of integration gives the adjoint explicitly as a convolution with the reflected kernel,

$$(\mathcal{F}^\dagger q)(x,t) = \int_0^T g(s-t)q(x,s)ds. \tag{5.11}$$

The Gaussian kernel is even, $g(s-t) = g(t-s)$, so $\mathcal{F}^\dagger = \mathcal{F}$ and the filter is self-adjoint. Consequently $\mathcal{F}^\dagger\mathcal{F} = \mathcal{F}^2$ amounts to applying the filter twice; away from the record ends its kernel is the convolution $g * g$,

$$(g*g)(t) = \int_{-\infty}^{\infty} g(\tau)g(t-\tau)d\tau = \frac{\omega}{2\sqrt{\pi}}e^{-\omega_k^2 t^2/4}, \tag{5.12}$$

again a Gaussian kernel, with temporal width $\sqrt{2}/\omega_k$, wider than $g$ by a factor $\sqrt{2}$. Unlike the pointwise weight, the source $\mathcal{F}^\dagger\mathcal{F}(u-u_d)$ is non-local in time and must be assembled from the entire residual history before the backward solve.

The above multiscale inversion can be similarly achievevd using regularization factor continuation scheme [Kucukcoban et al., 2019].

## 5.2 Discontinuous parameter fields

Recall the first-order Tikhonov regularization functional

$$R_{\mathrm{TIK}}[p] \coloneqq \frac{\alpha}{2}\int_\Omega \operatorname{grad} p(x)\cdot\operatorname{grad} p(x)\,dx, \tag{5.13}$$

where $\alpha > 0$ is a constant regularization parameter. First-order Tikhonov regularization requires $p \in H^1(\Omega)$ and penalizes large gradients and so suppresses sharp transitions. To see this explicitly, let

$$p(x) = \Delta p\,\theta(x-x_0), \tag{5.14}$$

where $\theta$ denotes the Heaviside step function and $\Delta p > 0$ is the jump size at $x_0 \in \Omega$. The distributional derivative is $p' = \Delta p\,\delta(x-x_0)$. Since $\delta(x-x_0) \notin L^2(\Omega)$, one has $R_{\mathrm{TIK}}[p] = +\infty$. The optimizer is therefore driven to replace any sharp transition by a smooth ramp, producing a blurred reconstruction.

The parameter $\alpha$ need not be a global constant; it may be taken as a spatially varying function $\alpha(x)$, such that

$$R_{\mathrm{TIK}}[p] \coloneqq \frac{1}{2}\int_\Omega \alpha(x)\nabla p(x)\cdot\nabla p(x)\,dx,$$

in direct analogy with a spatially varying diffusion coefficient in an elliptic PDE.

To recover jump-type solutions, one may instead penalize the total variation of $p$, defined for smooth $p$ as

$$R_{\mathrm{TV}}[p] \coloneqq \alpha\int_\Omega \sqrt{\operatorname{grad} p(x)\cdot\operatorname{grad} p(x)}\,dx. \tag{5.15}$$

For smooth $p$, one has $\int_\Omega |\operatorname{grad} p|\,dx = \|\operatorname{grad} p\|_{L^1(\Omega)}$. The $L^1$ norm admits the dual representation

$$\|f\|_{L^1(\Omega)} = \sup\left\{\int_\Omega f(x)\,\phi(x)\,dx \;:\; \|\phi\|_{L^\infty(\Omega)} \le 1\right\}. \tag{5.16}$$

The dual representation of the $L^p$ norm is

$$\|f\|_{L^p(\Omega)} = \sup\left\{\int_\Omega f\phi dx : \|\phi\|_{L^q(\Omega)} \le 1\right\},$$

where $1/p + 1/q = 1$.

This representation extends to non-smooth $p$ by transferring the derivative from $p$ onto a compact-supported test function $\phi$ via integration by parts.

For the step function $p(x) = \Delta p\theta(x - x_0)$ on $\Omega = (0, 1)$ with $x_0 \in (0, 1)$, integration by parts recovers its distributional derivative:

$$\begin{aligned}\int_0^1 p'(x)\phi(x)dx &= [p(x)\phi(x)]_0^1 - \int_0^1 p(x)\phi'(x)dx \\ &= -\Delta p \int_{x_0}^1 \phi'(x)dx = -\Delta p\,[\phi(1) - \phi(x_0)] \\ &= \Delta p\phi(x_0), \end{aligned} \tag{5.17}$$

where the boundary term $[p\phi]_0^1$ and the value $\phi(1)$ vanish because $\phi$ has compact support in $(0, 1)$. Taking the supremum over $\|\phi\|_{L^\infty} \leq 1$ gives

$$R_{\mathrm{TV}}[p] = \alpha \sup_{\|\phi\|_{L^\infty}\leq 1} \Delta p\,\phi(x_0) = \alpha\,\Delta p, \tag{5.18}$$

which is finite. Thus, $R_{\mathrm{TV}}$ admits jumps.

**Example 5.2.1 (Tikhonov vs. Total Variation Regularization)** *Consider the family of functions $u_\epsilon \in C^1([-1, 1])$ parameterized by $\epsilon > 0$, representing a smoothed jump:*

$$u_\epsilon(x) = \begin{cases} 0 & -1 \leq x < -\epsilon, \\ \dfrac{x + \epsilon}{2\epsilon} & -\epsilon \leq x \leq \epsilon, \\ 1 & \epsilon < x \leq 1. \end{cases} \tag{5.19}$$

*Evaluate the behavior of the Tikhonov ($H^1$ semi-norm) regularization penalty $R_{TIK}[u_\epsilon]$ and the Total Variation penalty $R_{TV}[u_\epsilon]$ as $\epsilon \to 0$.*

The classical derivative $u'_\epsilon(x)$ exists almost everywhere and is given by the bounded function

$$u'_\epsilon(x) = \begin{cases} 0 & x \in [-1, -\epsilon) \cup (\epsilon, 1], \\ \dfrac{1}{2\epsilon} & x \in (-\epsilon, \epsilon). \end{cases} \tag{5.20}$$

First, we compute the Tikhonov regularization penalty, which squares the derivative:

$$\begin{aligned} R_{\mathrm{TIK}}[u_\epsilon] &= \int_{-1}^1 |u'_\epsilon(x)|^2\,dx \\ &= \int_{-\epsilon}^{\epsilon} \left(\frac{1}{2\epsilon}\right)^2 dx = \left(\frac{1}{4\epsilon^2}\right)(2\epsilon) = \frac{1}{2\epsilon}. \end{aligned} \tag{5.21}$$

Taking the limit as the transition layer vanishes yields an infinite cost:

$$\lim_{\epsilon\to 0} R_{\mathrm{TIK}}[u_\epsilon] = \lim_{\epsilon\to 0} \frac{1}{2\epsilon} = \infty. \tag{5.22}$$

Next, we compute the Total Variation regularization penalty, which integrates the absolute value of the derivative:

$$\begin{aligned} R_{\mathrm{TV}}[u_\epsilon] &= \int_{-1}^{1} |u'_\epsilon(x)|\, dx \\ &= \int_{-\epsilon}^{\epsilon} \left| \frac{1}{2\epsilon} \right| dx = \left( \frac{1}{2\epsilon} \right) (2\epsilon) = 1. \end{aligned} \tag{5.23}$$

Taking the limit as $\epsilon \to 0$ yields a uniform bound independent of the edge sharpness:

$$\lim_{\epsilon \to 0} R_{\mathrm{TV}}[u_\epsilon] = 1. \tag{5.24}$$

Consequently, Tikhonov regularization penalizes sharp gradients excessively and blows up for step-like discontinuities, whereas Total Variation remains uniformly bounded and preserves sharp edges cleanly.

Because $|\operatorname{grad} p|$ is non-differentiable wherever $\operatorname{grad} p = 0$, a smoothed surrogate is introduced by replacing $|\operatorname{grad} p|$ with $\sqrt{|\operatorname{grad} p|^2 + \epsilon^2}$ for a small parameter $\epsilon > 0$. This yields the smoothed total variation functional

$$R^\epsilon_{\mathrm{TV}}[p] := \alpha \int_\Omega \sqrt{\operatorname{grad} p(x) \cdot \operatorname{grad} p(x) + \epsilon^2}\, dx. \tag{5.25}$$

As $\epsilon \to 0^+$, $R^\epsilon_{\mathrm{TV}}[p]$ converges to $R_{\mathrm{TV}}[p]$, while for any $\epsilon > 0$ the functional is Fréchet differentiable, which renders gradient-based optimization methods applicable.

The directional derivative of the smoothed total variation functional in a direction $\tilde{p}$ is

$$d_p R^\epsilon_{\mathrm{TV}}[p](\tilde{p}) = \alpha \int_\Omega \frac{\operatorname{grad} p \cdot \operatorname{grad} \tilde{p}}{\sqrt{\operatorname{grad} p \cdot \operatorname{grad} p + \epsilon^2}}\, dx. \tag{5.26}$$

Integration by parts gives

$$\begin{aligned} d_p R^\epsilon_{\mathrm{TV}}[p](\tilde{p}) = &- \alpha \int_\Omega \operatorname{grad} \cdot \left( \frac{\operatorname{grad} p}{\sqrt{\operatorname{grad} p \cdot \operatorname{grad} p + \epsilon^2}} \right) \tilde{p}\, dx \\ &+ \alpha \int_{\partial\Omega} \frac{\operatorname{grad} p \cdot n}{\sqrt{\operatorname{grad} p \cdot \operatorname{grad} p + \epsilon^2}}\, \tilde{p}\, ds, \end{aligned} \tag{5.27}$$

so the $L^2$ gradient is

$$g_p = -\alpha \operatorname{grad} \cdot \left( \frac{\operatorname{grad} p}{\sqrt{\operatorname{grad} p \cdot \operatorname{grad} p + \epsilon^2}} \right), \tag{5.28}$$

with the natural boundary condition $\operatorname{grad} p \cdot n = 0$ on $\partial\Omega$.

The second variation in directions $\tilde{p}$ and $\hat{p}$ follows by differentiating the first variation once more. Writing $s[p] \coloneqq \sqrt{\operatorname{grad} p \cdot \operatorname{grad} p + \epsilon^2}$,

$$\begin{aligned} d_p^2 R_{\mathrm{TV}}^{\epsilon}[p](\tilde{p}, \hat{p}) &= \alpha \int_\Omega \left[ \frac{\operatorname{grad} \tilde{p} \cdot \operatorname{grad} \hat{p}}{s[p]} - \frac{(\operatorname{grad} p \cdot \operatorname{grad} \tilde{p})(\operatorname{grad} p \cdot \operatorname{grad} \hat{p})}{s[p]^3} \right] dx \\ &= \alpha \int_\Omega \operatorname{grad} \tilde{p} \cdot A[p] \operatorname{grad} \hat{p} \, dx, \end{aligned} \tag{5.29}$$

where the symmetric tensor field is

$$A[p] \coloneqq \frac{1}{s[p]} \left( I - \frac{\operatorname{grad} p \otimes \operatorname{grad} p}{s[p]^2} \right). \tag{5.30}$$

Integration by parts in $\tilde{p}$ gives the strong-form Hessian operator acting on $\hat{p}$,

$$H_p \hat{p} = -\alpha \operatorname{grad} \cdot \big( A[p] \operatorname{grad} \hat{p} \big), \tag{5.31}$$

with the natural boundary condition $A[p] \operatorname{grad} \hat{p} \cdot n = 0$ on $\partial\Omega$.

## 5.3 Bounded and discrete parameter fields

Many physical parameters, such as mass density of elastic moduli, are often strictly positive. We therefore consider the bound-constrained problem

$$\min_p M[u(p)] \quad \text{subject to} \quad p(x) \geq p_0. \tag{5.32}$$

Two strategies for imposing the constraint are presented: a logarithmic barrier, which augments the objective with a penalty that grows without bound at the constraint, and a nonlinear reparametrization, which removes the constraint by construction.

### 5.3.1 Logarithmic barrier functional

The logarithmic barrier method replaces the inequality-constrained problem by

$$\min_p J[p] \coloneqq M[u(p)] + R[p] + \Phi[p]. \tag{5.33}$$

Here the barrier functional $\Phi[p]$ is defined with a barrier parameter $\mu > 0$ by

$$\Phi[p] \coloneqq -\mu \int_\Omega \log(p - p_0) dx. \tag{5.34}$$

This functional is finite only on the strictly feasible set $p > p_0$, and it diverges to $+\infty$ as $p$ approaches the bound, so any minimizer of $J$ is automatically strictly feasible. However, like any penalty term, the barrier may come to dominate the objective and overshadow the misfit functional.

The logarithmic barrier functional is illustrated in Figure 5.1. The scalar objective function $p^2$ has a minimum at $p = 0$. The augmented function $J[p] = p^2 - \mu \log(p - p_0)$ restricts the feasible region to $p > p_0 = -1$, and for $\mu = 3/2$ its minimizer is relocated to $p = 1/2$.

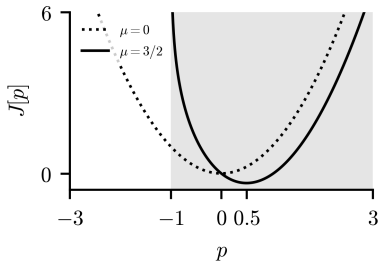



Figure 5.1: Logarithmic barrier functional for $p > p_0$. The minimizer is shifted by the barrier term, and the feasible region is highlighted in light gray.

The barrier enters the stationarity conditions only through the control equation, since it depends on $p$ alone. Its first and second variations are

$$d_p\Phi[p](\tilde{p}) = -\mu \int_\Omega \frac{\tilde{p}}{p - p_0} dx \text{ and } d_p^2\Phi[p](\tilde{p}, \hat{p}) = \mu \int_\Omega \frac{\tilde{p}\hat{p}}{(p - p_0)^2} dx, \tag{5.35}$$

respectively. Forming the Lagrangian for (5.33) with adjoint variable $v$,

$$\mathcal{L}[u, p, v] = M[u] + R[p] + E[u, p, v] + \Phi[p], \tag{5.36}$$

the Newton step at iteration $k$, in terms of the state, parameter, and adjoint increments $\hat{u}^{(k)}$, $\hat{p}^{(k)}$, $\hat{v}^{(k)}$, is the block system

$$\begin{bmatrix} F & A^\dagger & B^\dagger \\ A & D + Z & C^\dagger \\ B & C & 0 \end{bmatrix} \begin{pmatrix} \hat{u}^{(k)} \\ \hat{p}^{(k)} \\ \hat{v}^{(k)} \end{pmatrix} = - \begin{pmatrix} g_u \\ g_p - \mu(p - p_0)^{-1} \\ g_v \end{pmatrix}. \tag{5.37}$$

Here

$$Z = \frac{\mu}{(p - p_0)^2}. \tag{5.38}$$

The additional terms related to the barrier grow without bound where $p$ approaches $p_0$ resulting the Newton system becomes increasingly ill-conditioned near the active part of the constraint.

### 5.3.2 Nonlinear reparametrization

An alternative that avoids the ill-conditioning of the barrier is to choose a parametrization of the field that satisfies the bound by construction, at the cost of increased nonlinearity. Writing the parameter as

$$p = p_0 + \exp(q), \tag{5.39}$$

and inverting for the unconstrained variable $q \in \mathbb{R}$, the constraint $p > p_0$ holds automatically, since the exponential is strictly positive.

Figure 5.2 illustrates the exponential parameterization for the scalar objective function $J[p] = p^2$. The constraint $p > p_0 = -1$ is enforced through a new parameter $q$ via $p = p_0 + e^q$. Since $e^q > 0$ for all $q \in \mathbb{R}$, the feasible region

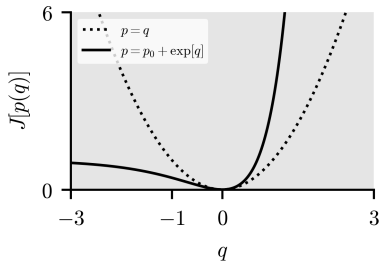



Figure 5.2: Exponential reparameterization for $p > p_0$. The feasible region (highlighted in gray) for the new parameter $q$ is $\mathbb{R}$.

for $q$ is the entire real line, while the bound $p > p_0$ holds automatically. The minimizer remains unchanged.

When the parameter space is binary, as is common in design problems, a quasi-binary parametrization may be used instead,

$$p(q) = \frac{p_H - p_L}{2}\left[1 + \tanh\frac{q}{\gamma}\right] + p_L. \tag{5.40}$$

Here $p_L$ and $p_H$ are the prescribed low and high values, and $\gamma$ controls the sharpness of the transition between them: as $\gamma \to 0$ the map approaches a step, recovering the binary limit.

# Bibliography

[Gelfand et al., 2000] Gelfand, I., Fomin, S., and Silverman, R. (2000). *Calculus of Variations*. Dover Books on Mathematics. Dover Publications.

[Ghattas, 2015] Ghattas, O. (2015). Computational and variational methods for inverse problems. The University of Texas at Austin.

[Kucukcoban et al., 2019] Kucukcoban, S., Goh, H., and Kallivokas, L. F. (2019). On the full-waveform inversion of lame parameters in semi-infinite solids in plane strain. *International Journal of Solids and Structures*, 164:104–119.

[Oden and Demkowicz, 2017] Oden, J. T. and Demkowicz, L. (2017). *Applied functional analysis*. Chapman and Hall/CRC.

[Petra and Stadler, 2011] Petra, N. and Stadler, G. (2011). Model variational inverse problems governed by partial differential equations. Technical report, Oden Institute, The University of Texas at Austin.

[Quarteroni et al., 2006] Quarteroni, A., Sacco, R., and Saleri, F. (2006). *Numerical Mathematics*. Texts in Applied Mathematics. Springer Berlin Heidelberg.

[Stone and Goldbart, 2009] Stone, M. and Goldbart, P. (2009). *Mathematics for physics: a guided tour for graduate students*. Cambridge University Press.

[Vogel, 2002] Vogel, C. (2002). *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics.