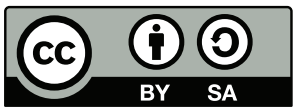
# Inverse Problems

Lecture notes, Michaelmas term 2019
University of Cambridge

**Hanne Kekkonen and Yury Korolev**

May 1, 2020





# Contents

These lecture notes use material from the notes of the courses "Bayesian Inverse Problems" and "Inverse Problems in Imaging" which were held by Hanne Kekkonen in Lent term 2019, and by Yury Korolev in Michaelmas term 2019 at the University of Cambridge.[1] Complementary material can be found in the following books, lecture notes and review papers:

1. Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems.* Springer, 1996.

2. Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier and Frank Lenzen. *Variational Methods in Imaging.* Springer, 2008.

3. Kristian Bredies and Dirk Lorenz. *Mathematische Bildverarbeitung (in German).* Vieweg+Teubner Verlag, Springer, 2011

4. Martin Benning and Martin Burger. *Modern regularization methods for inverse problems.* Acta Numerica, 27, 1-111 (2018)

   `https://www.cambridge.org/core/journals/acta-numerica/article/modern-regularization-methods-for-inverse-problems/1C84F0E91BF20EC36D8E846EF8CCB830`

5. K.Saxe. *Beginning Functional Analysis.* Springer, 2002

6. Masoumeh Dashti and Andrew M. Stuart, *The Bayesian approach to inverse problems*, Handbook of Uncertainty Quantification, 2016.

7. Jari Kaipio and Erkki Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, 2005.

8. O. Kallenberg, *Foundations of modern probability theory*, Springer, 1997.

9. Andrew M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, 2010.

These lecture notes are under constant redevelopment and might contain typos or errors. We very much appreciate if you report any mistakes found (to hnk22@cam.ac.uk or y.korolev@damtp.cam.ac.uk). Thanks!

[1] `http://www.damtp.cam.ac.uk/research/cia/previous-terms`

# Chapter 1

# Introduction to Inverse Problems

Inverse problems arise from the need to gain information about an unknown object of interest from given indirect measurements. Inverse problems have several applications varying from medical imaging and industrial process monitoring to ozone layer tomography and modelling of financial markets. The common feature for inverse problems is the need to understand indirect measurements and to overcome extreme sensitivity to noise and modelling inaccuracies. In this course we employ both deterministic and probabilistic approach to inverse problems to find stable and meaningful solutions that allow us quantify how inaccuracies in the data or model affect the obtained estimate.

## 1.1 Well-posed and ill-posed problems

We start by considering the problem of finding $u \in \mathbb{R}^d$ that satisfies the equation

$$f = Au, \tag{1.1}$$

where $f \in \mathbb{R}^k$ is given. We refer to $f$ as observed data or measurement and $u$ as an unknown. The physical phenomena that relates the unknown and the measurement is modelled by a matrix $A \in \mathbb{R}^{k \times d}$. In real life the perfect data given in (1.1) is perturbed by noise and we observe measurements

$$f_n = Au + n, \tag{1.2}$$

where $n \in \mathbb{R}^k$ represents the observational noise.

We are interested in ill-posed inverse problems, where the inverse problem is more difficult to solve than the direct problem of finding $f_n$ when $u$ is given. To explain this we first need to introduce well-posedness as defined by Jacques Hadamard:

**Definition 1.1.1.** *A problem is called well-posed if*

1. *There exists at least one solution. (Existence)*
2. *There is at most one solution. (Uniqueness)*
3. *The solution depends continuously on data. (Stability)*

The direct or forward problem is assumed to be well-posed. The inverse problems are ill-posed and break at least one of the above conditions.

1. Assume that $d < k$ and $A : \mathbb{R}^d \to \mathcal{R}(A) \subsetneq \mathbb{R}^k$, where the range of $A$ is a proper subset of $\mathbb{R}^k$. Furthermore, we assume that $A$ has a unique inverse $A^{-1} : \mathcal{R}(A) \to \mathbb{R}^k$. Because of the noise in the measurement $f_n \notin \mathcal{R}(A)$ so that simply inverting $A$ with the data given in (1.2) is not possible. Note that usually only the statistical properties of the noise $n$ are known so we cannot just subtract it.

2. Assume next that $d > k$ and $A : \mathbb{R}^d \to \mathbb{R}^k$, in which case the system is underdetermined. We then have more unknowns than equations which means that there are several possible solutions.

3. Consider next case $d = k$ and there exist $A^{-1} : \mathbb{R}^k \to \mathbb{R}^k$ but the condition number $\kappa = \lambda_1/\lambda_k$, where $\lambda_1$ and $\lambda_k$ are the biggest and smallest eigenvalues of $A$, is very large. Such a matrix is said to be ill-conditioned and is almost singular. In this case the problem is sensitive even to smallest errors in the measurement. Hence the naive reconstruction $\widetilde{u} = A^{-1}f_n = u + A^{-1}n$ does not produce a meaningful solution but will be dominated by $A^{-1}n$. Note that $\|A^{-1}n\|_2 \approx \|n\|_2/\lambda_k$ can be arbitrarily large.

The last part illustrates one of the key perspectives of inverse problem theory; How can we stabilise the reconstruction process while maintaining acceptable accuracy?

A deterministic way of achieving a unique and stable solution for the problem (1.2) is to use regularisation theory. In the classical Tikhonov regularisation a solution is attained by solving

$$\min_{u\in\mathbb{R}^d} \left( \|Au - f_n\|^2 + \alpha \|Lu\|^2 \right). \tag{1.3}$$

Above $\alpha$ acts as a tuning parameter balancing the effect of the data fidelity term $\|Au - f_n\|^2$ and the stabilising regularisation term $\|u\|^2$. The first half of the course will concentrate on regularisation theory.

Another way of tackling problems arising from ill-posedness is Bayesian inversion. The idea of statistical inversion methods is to rephrase the inverse problem as a question of statistical inference. We then consider problem

$$F = AU + N, \tag{1.4}$$

where the measurement, unknown and noise are now modelled as random variables. This approach allows us to model the noise through its statistical properties. We can also encode our *a priori* knowledge of the unknown in form of a probability distribution that assigns higher probability to those values of $u$ we expect to see. The solution to (6.1) is so-called *posterior distribution*, which is the conditional probability distribution of $u$ given a measurement $m$. This distribution can then be used to obtain estimates that are most likely in some sense. We will return to the Bayesian approach to inverse problems in the second half of the course

In this course we will concentrate on continuous inverse problems where in (1.1) and (1.2) $A : X \to Y$ is a linear forward operator acting between some spaces $X$ and $Y$, typically Hilbert or Banach spaces, the measured data $f_n \in Y$ is a function and $u \in X$ is the quantity we want to reconstruct from the data. Linear inverse problems include such important applications as computer tomography, magnetic resonance imaging and image deblurring in microscopy or astronomy. There are, however, many other important applications, such as seismic imaging, where the forward operator is non-linear (e.g., parameter identification problems for PDEs). Next we will take a look at some examples of linear inverse problems to see what kind of challenges we face when trying to solve them.

## 1.2 Examples of inverse problems

### 1.2.1 Signal deblurring

The deblurring (or deconvolution) problem of recovering an input signal $u$ form an observed signal

$$f_n(t) = \int_{-\infty}^{\infty} a(t-s)u(s)ds + n(t)$$

occurs in many imaging, and image- and signal processing applications. Here the function $a$ is known as the blurring kernel.

The noiseless data is given by $f(t) = \int_{-\infty}^{\infty} a(t-s)u(s)ds$ and its Fourier transform is $\widehat{f}(\xi) = \int_{-\infty}^{\infty} e^{-i\xi t} f(t)dt$. The convolution theorem implies

$$\widehat{f}(\xi) = \widehat{a}(\xi)\widehat{u}(\xi),$$

and hence by inverse Fourier transform

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\xi}\frac{\widehat{f}(\xi)}{\widehat{a}(\xi)}d\xi.$$

However, we can only observe noisy measurements and hence we have on the frequency domain $\widehat{f_n}(\xi) = \widehat{a}(\xi)\widehat{u}(\xi) + \widehat{n}(\xi)$. The estimate $u_{est}$ based on the convolution theorem is given by

$$u_{est}(t) = u(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\xi}\frac{\widehat{n}(\xi)}{\widehat{a}(\xi)}d\xi,$$

which is often not even well defined, since usually the kernel $a$ decreases exponentially (or has compact support), making the denominator small, whereas the Fourier transform of the noise will be non-zero.

### 1.2.2 Heat equation

Next we study the problem of recovering the initial condition $u$ of the heat equation from a noisy observation $f_n$ of the solution at some time $T > 0$. We consider the heat equation on a torus $\mathbb{T}^d$, with Dirichlet boundary conditions

$$\begin{cases} \frac{dv}{dt} - \Delta v = 0 & \text{on } \mathbb{T}^d \times \mathbb{R}_+ \\ v(x,t) = 0 & \text{on } \partial\mathbb{T}^d \times \mathbb{R}_+ \\ v(x,T) = f(x) & \text{on } \mathbb{T}^d \\ v(x,0) = u(x) & \text{on } \mathbb{T}^d \end{cases}$$

where $\Delta$ denotes the Laplace operator and $\mathcal{D}(\Delta) = H_0^1(\mathbb{T}^d) \cap H^2(\mathbb{T}^d)$. Note that the operator $-\Delta$ is positive and self-adjoint on Hilbert space $\mathcal{H} = L^2(\mathbb{T}^d)$.

Given a function $u \in L^2(\mathbb{T}^d)$ we can decompose it as a Fourier series

$$u(x) = \sum_{n\in\mathbb{Z}^d} u_n e^{2\pi i\langle n,x\rangle},$$

where $u_n = \langle u, e^{2\pi i \langle n,x \rangle} \rangle$ are the Fourier coefficients, and the identity holds for almost every $x \in \mathbb{T}^d$. The $L^2$ norm of $u$ is given by the Parseval's identity $\|u\|_{L^2}^2 = \sum |u_n|^2$. Remember that the Sobolev space $H^s(\mathbb{T}^d)$, $s \in \mathbb{N}$, consist of all $L^2(\mathbb{T}^d)$ integrable functions whose $\alpha^{th}$ order weak derivatives exist and are $L^2(\mathbb{T}^d)$ integrable for all $|\alpha| \leqslant s$. The fractional Sobolev space $H^s(\mathbb{T}^d)$ is given by the subspace of functions $u \in L^2(\mathbb{T}^d)$, such that

$$\|u\|_{H^s}^2 = \sum_{n \in \mathbb{Z}^d} (1 + 4\pi^2 |n|^2)^s |u_n|^2 < \infty. \tag{1.5}$$

Note that for a positive integer $s$, the above definition agrees with the definition given using the weak derivatives. For $s < 0$, we define $H^s(\mathbb{T}^d)$ via duality or as the closure of $L^2(\mathbb{T}^d)$ under the norm (1.5). The resulting spaces are separable for all $s \in \mathbb{R}$.

The eigenvectors of $-\Delta$ in $\mathbb{T}^d$ form the orthonormal basis of $L^2(\mathbb{T}^d)$ and the eigenvalues are given by $4\pi^2|n|^2$, $n \in \mathbb{Z}^d$. We can also work on real-valued functions where the eigenfunctions $\{\varphi_j\}_{j=1}^{\infty}$ comprise sine and cosine functions. The eigenvalues of $-\Delta$, when ordered on a one-dimensional lattice, then satisfy $\lambda_j \asymp j^{\frac{2}{d}}$. The notation $\asymp$ means that there exist constants $C_1, C_2 > 0$, such that $C_1 j^{\frac{2}{d}} \leqslant \lambda_j \leqslant C_2 j^{\frac{2}{d}}$.

The solution to the forward heat equation can be written as

$$v(t) = \sum_{j=1}^{\infty} u_j e^{-\lambda_j t} \varphi_j.$$

We notice that

$$\|v(t)\|_{H^s}^2 \asymp \sum_{j=1}^{\infty} j^{\frac{2s}{d}} e^{-2\lambda_j t} |u_j|^2 = t^{-s} \sum_{j=1}^{\infty} (\lambda_j t)^s e^{-2\lambda_j t} |u_j|^2 \leqslant C t^{-s} \sum_{j=1}^{\infty} |u_j|^2 = C t^{-s} \|u\|_{L^2}$$

which implies that $v(t) \in H^s(\mathbb{T}^d)$ for all $s > 0$.

We now have observation model

$$f_n = Au + n,$$

where $A = e^{T\Delta}$ and $n$ is the observational noise. The noise is not usually smooth (the often assumed white noise is not even an $L^2$ function) and hence measurement $f_n$ is not in the image space $\mathcal{D}(e^{T\Delta}) \subset \cap_{s>0} H^s(\mathbb{T}^d)$.

### 1.2.3 Differentiation

Consider the problems of evaluation the derivative of a function $f \in L^2[0, \pi/2]$. Let

$$Df = f',$$

where $D \colon L^2[0, \pi/2] \to L^2[0, \pi/2]$.

**Proposition 1.2.1.** *The operator $D$ is unbounded from $L^2[0, \pi/2] \to L^2[0, \pi/2]$.*

*Proof.* Take a sequence $f_n(x) = \sin(nx)$, $n = 1, \ldots, \infty$. Clearly, $f_n \in L^2[0, \pi/2]$ for all $n$ and $\|f_n\| = \sqrt{\frac{\pi}{4}}$. However, $Df_n(x) = n\cos(nx)$ and $\|Df_n\| = n \to \infty$ as $n \to \infty$. Therefore, $D$ is unbounded. □

This shows that differentiation is ill-posed from $L^2$ to $L^2$. It does not mean that it can not be well-posed in other spaces. For instance, it is well-posed from $H^1$ (the Sobolev space of $L^2$ functions whose derivatives are also $L^2$) to $L^2$. Indeed, $\forall u \in H^1$ we get

$$\|Df\|_{L^2} = \|f'\|_{L^2} \leqslant \|f\|_{H^1} = \|f\|_{L^2} + \|f'\|_{L^2}.$$

However, since in practice we typically deal with functions corrupted by nonsmooth noise, the $L^2$ setting is practice-relevant, while the $H^1$ setting is not.

Differentiation can be written as an inverse problem for an integral equation. For instance, the derivative $u$ of some function $f \in L^2[0,1]$ with $f(0) = 0$ satisfies

$$f(x) = \int_0^x u(t)\,dt,$$

which can be written as an operator equation $Au = f$ with $(A\cdot)(x) := \int_0^x \cdot(t)\,dt$.

### 1.2.4 Matrix inversion

In finite dimensions, the inverse problem (1.1) is a linear system. Linear systems are formally well-posed in the sense that the error in the solution is bounded by some constant times the error in the right-hand side, however, this constant depends on the condition number of the matrix $A$ and can get arbitrary large for matrices with large condition numbers. In this case, we speak of *ill-conditioned* problems.

Consider the problem (1.1) with $u \in \mathbb{R}^n$ and $f \in \mathbb{R}^n$ being $n$-dimensional vectors with real entries and $A \in \mathbb{R}^{n\times n}$ being a matrix with real entries. Assume further $A$ to be symmetric and positive definite.

We know from the spectral theory of symmetric matrices that there exist eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n > 0$ and corresponding (orthonormal) eigenvectors $a_j \in \mathbb{R}^n$ for $j \in \{1,\ldots,n\}$ such that $A$ can be written as

$$A = \sum_{j=1}^n \lambda_j a_j a_j^\top. \tag{1.6}$$

It is well known from numerical linear algebra that the condition number $\kappa = \lambda_1/\lambda_n$ is a measure of how stable (1.1) can be solved, which we will illustrate what follows.

We assume that we measure $f_\delta$ instead of $f$, with $\|f - f_\delta\|_2 \leqslant \delta\|A\| = \delta\lambda_1$, where $\|\cdot\|_2$ denotes the Euclidean norm of $\mathbb{R}^n$ and $\|A\|$ the operator norm of $A$ (which equals the largest eigenvalue of $A$). Then, if we further denote with $u_\delta$ the solution of $Au_\delta = f_\delta$, the difference between $u_\delta$ and the solution $u$ to (1.1) is

$$u - u_\delta = \sum_{j=1}^n \lambda_j^{-1} a_j a_j^\top (f - f_\delta).$$

Therefore, we can estimate

$$\|u - u_\delta\|_2^2 = \sum_{j=1}^n \lambda_j^{-2} \underbrace{\|a_j\|_2^2}_{=1} |a_j^\top (f - f_\delta)|^2 \leqslant \lambda_n^{-2}\|f - f_\delta\|_2^2,$$

due to the orthonormality of eigenvectors, the Cauchy-Schwarz inequality, and $\lambda_n \leqslant \lambda_j$. Thus, taking square roots on both sides yields the estimate

$$\|u - u_\delta\|_2 \leqslant \lambda_n^{-1}\|f - f_\delta\|_2 \leqslant \kappa\delta.$$

Hence, we observe that in the worst case an error $\delta$ in the data $y$ is amplified by the condition number $\kappa$ of the matrix $A$. A matrix with large $\kappa$ is therefore called *ill-conditioned*. We want to demonstrate the effect of this error amplification with a small example.

**Example 1.2.1.** Let us consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & \frac{1001}{1000} \end{pmatrix},$$

which has eigenvalues $\lambda_j = 1 + \frac{1}{2000} \pm \sqrt{1 + \frac{1}{2000^2}}$, condition number $\kappa \approx 4002 \gg 1$, and operator norm $\|A\| \approx 2$. For given data $f = (1,1)^\top$ the solution to $Au = f$ is $u = (1,0)^\top$.

Now let us instead consider perturbed data $f_\delta = (99/100, 101/100)^\top$. The solution $u_\delta$ to $Au_\delta = f_\delta$ is then $u_\delta = (-19.01, 20)^\top$.

Let us reflect on the amplification of the measurement error. By our initial assumption we find that $\delta = \|f - f_\delta\|/\|A\| \approx \|(0.01, -0.01)^\top\|/2 = \sqrt{2}/200$. Moreover, the norm of the error in the reconstruction is then $\|u - u_\delta\| = \|(20.01, 20)^\top\| \approx 20\sqrt{2}$. As a result, the amplification due to the perturbation is $\|u - u_\delta\|/\delta \approx 4000 \approx \kappa$.

### 1.2.5 Tomography

In almost any tomography application the underlying inverse problem is either the inversion of the Radon transform[1] or of the X-ray transform.

For $u \in C_0^\infty(\mathbb{R}^n)$, $s \in \mathbb{R}$, and $\theta \in S^{n-1}$ the *Radon transform* $R : C_0^\infty(\mathbb{R}^n) \to C^\infty(S^{n-1} \times \mathbb{R})$ can be defined as the integral operator

$$\begin{aligned} f(\theta, s) = (\mathcal{R}u)(\theta, s) &= \int_{x \cdot \theta = s} u(x)\, dx \\ &= \int_{\theta^\perp} u(s\theta + y)\, dy, \end{aligned} \tag{1.7}$$

which, for $n = 2$, coincides with the X-ray transform,

$$f(\theta, s) = (\mathcal{P}u)(\theta, s) = \int_{\mathbb{R}} u(s\theta + t\theta^\perp)\, dt,$$

for $\theta \in S^{n-1}$ and $\theta^\perp$ being the vector orthogonal to $\theta$. Hence, the X-ray transform (and therefore also the Radon transform in two dimensions) integrates the function $u$ over lines in $\mathbb{R}^n$, see Fig. 1.1[2].

**Example 1.2.2.** Let $n = 2$. Then $S^{n-1}$ is simply the unit sphere $S^1 = \{\theta \in \mathbb{R}^2 \mid \|\theta\| = 1\}$. We can choose for instance $\theta = (\cos(\varphi), \sin(\varphi))^\top$, for $\varphi \in [0, 2\pi)$, and parametrise the Radon transform in terms of $\varphi$ and $s$, i.e.

$$f(\varphi, s) = (\mathcal{R}u)(\varphi, s) = \int_{\mathbb{R}} u(s\cos(\varphi) - t\sin(\varphi), s\sin(\varphi) + t\cos(\varphi))\, dt. \tag{1.8}$$

[1] Named after the Austrian mathematician Johann Karl August Radon (16 December 1887 – 25 May 1956).

[2] Figure adapted from Wikipedia `https://commons.wikimedia.org/w/index.php?curid=3001440`, by Begemotv2718, CC BY-SA 3.0.

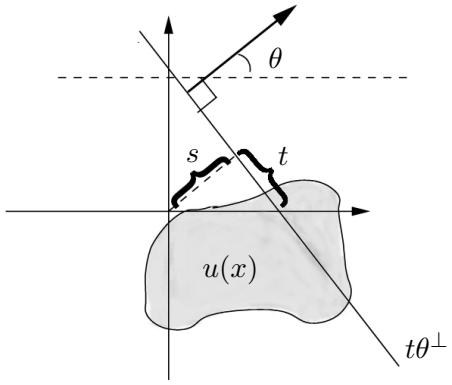



Figure 1.1: Visualization of the Radon transform in two dimensions (which coincides with the X-ray transform). The function $u$ is integrated over the ray parametrized by $\theta$ and $s$.[3]

Note that—with respect to the origin of the reference coordinate system—$\varphi$ determines the angle of the line along one wants to integrate, while $s$ is the offset from that line from the centre of the coordinate system.

It can be shown that the Radon transform is linear and continuous, i.e. $R \in \mathcal{L}(L^2(B), L^2(Z))$, and even compact.

In **X-ray Computed Tomography (CT)**, the unknown quantity $u$ represents a spatially varying density that is exposed to X-radiation from different angles, and that absorbs the radiation according to its material or biological properties.

The basic modelling assumption for the intensity decay of an X-ray beam is that within a small distance $\Delta t$ it is proportional to the intensity itself, the density, and the distance, i.e.

$$\frac{I(x + (t + \Delta t)\theta) - I(x + t\theta)}{\Delta t} = -I(x + t\theta)u(x + t\theta),$$

for $x \in \theta^\perp$. By taking the limit $\Delta t \to 0$ we end up with the ordinary differential equation

$$\frac{d}{dt} I(x + t\theta) = -I(x + t\theta)u(x + t\theta), \tag{1.9}$$

Let $R > 0$ be the radius of the domain of interest centred at the origin. Then, we integrate (1.9) from $t = -\sqrt{R^2 - \|x\|_2^2}$, the position of the emitter, to $t = \sqrt{R^2 - \|x\|_2^2}$, the position of the detector, and obtain

$$\int_{-\sqrt{R^2-\|x\|_2^2}}^{\sqrt{R^2-\|x\|_2^2}} \frac{\frac{d}{dt} I(x + t\theta)}{I(x + t\theta)}\, dt = -\int_{-\sqrt{R^2-\|x\|_2^2}}^{\sqrt{R^2-\|x\|_2^2}} u(x + t\theta)\, dt\,.$$

Note that, due to $d/dx \log(f(x)) = f'(x)/f(x)$, the left hand side in the above equation simplifies to

$$\int_{-\sqrt{R^2-\|x\|_2^2}}^{\sqrt{R^2-\|x\|_2^2}} \frac{\frac{d}{dt} I(x + t\theta)}{I(x + t\theta)}\, dt = \log\left(I\left(x + \sqrt{R^2 - \|x\|_2^2}\theta\right)\right) - \log\left(I\left(x - \sqrt{R^2 - \|x\|_2^2}\theta\right)\right).$$

As we know the radiation intensity at both the emitter and the detector, we therefore know $f(x, \theta) = \log(I(x - \theta\sqrt{R^2 - \|x\|_2^2})) - \log(I(x + \theta\sqrt{R^2 - \|x\|_2^2}))$ and we can write the

estimation of the unknown density $u$ as the inverse problem of the X-ray transform (1.8) (if we further assume that $u$ can be continuously extended to zero outside of the circle of radius $R$).

# Chapter 2

# Generalised Solutions

Functional analysis is the basis of the theory that we will cover in this course. We cannot recall all basic concepts of functional analysis and instead refer to popular textbooks that deal with this subject, e.g., [10, 31, 27]. Nevertheless, we shall recall a few important definitions that will be used in this lecture.

We will focus on inverse problems with *bounded linear operators* $A$, i.e. $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with

$$\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} := \sup_{u \in \mathcal{X}\setminus\{0\}} \frac{\|Au\|_{\mathcal{Y}}}{\|u\|_{\mathcal{X}}} = \sup_{\|u\|_{\mathcal{X}} \leqslant 1} \|Au\|_{\mathcal{Y}} < \infty.$$

For $A\colon \mathcal{X} \to \mathcal{Y}$ we further want to denote by

1. $\mathcal{D}(A) := \mathcal{X}$ the domain,
2. $\mathcal{N}(A) := \{u \in \mathcal{X} \mid Au = 0\}$ the kernel,
3. $\mathcal{R}(A) := \{f \in \mathcal{Y} \mid f = Au, u \in \mathcal{X}\}$ the range

of $A$.

We say that $A$ is continuous at $u \in \mathcal{X}$ if for all $\varepsilon > 0$ there exists $\delta > 0$ with

$$\|Au - Av\|_{\mathcal{Y}} \leqslant \varepsilon \text{ for all } v \in \mathcal{X} \text{ with } \|u - v\|_{\mathcal{X}} \leqslant \delta.$$

For linear $K$ it can be shown that continuity is equivalent to boundedness, i.e. the existence of a constant $C > 0$ such that

$$\|Au\|_{\mathcal{Y}} \leqslant C\|u\|_{\mathcal{X}}$$

for all $u \in \mathcal{X}$. Note that this constant $C$ actually equals the operator norm $\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$.

In this Chapter we only consider $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\mathcal{X}$ and $\mathcal{Y}$ being Hilbert spaces. From functional calculus we know that every Hilbert space $\mathcal{U}$ is equipped with a *scalar product*, which we are going to denote by $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ (or simply $\langle \cdot, \cdot \rangle$, whenever the space is clear from the context). In analogy to the transpose of a matrix, this scalar product structure together with the theorem of Fréchet-Riesz [31, Section 2.10, Theorem 2.E] allows us to define the (unique) *adjoint operator* of $A$, denoted with $A^*$, as follows:

$$\langle Au, v \rangle_{\mathcal{Y}} = \langle u, A^* v \rangle_{\mathcal{X}}, \text{ for all } u \in \mathcal{X}, v \in \mathcal{Y}.$$

In addition to that, a scalar product can be used to define orthogonality. Two elements $u, v \in \mathcal{X}$ are said to be *orthogonal* if $\langle u, v\rangle = 0$. For a subset $\mathcal{X}' \subset \mathcal{X}$ the *orthogonal complement* of $\mathcal{X}'$ in $\mathcal{X}$ is defined as

$$\mathcal{X}'^{\perp} \coloneqq \left\{u \in \mathcal{X} \mid \langle u, v\rangle_{\mathcal{X}} = 0 \text{ for all } v \in \mathcal{X}'\right\}.$$

One can show that $\mathcal{X}'^{\perp}$ is a closed subspace and that $\mathcal{X}^{\perp} = \{0\}$. Moreover, we have that $\mathcal{X}' \subset (\mathcal{X}'^{\perp})^{\perp}$. If $\mathcal{X}'$ is a closed subspace then we even have $\mathcal{X}' = (\mathcal{X}'^{\perp})^{\perp}$. In this case there exists the *orthogonal decomposition*

$$\mathcal{X} = \mathcal{X}' \oplus \mathcal{X}'^{\perp},$$

which means that every element $u \in \mathcal{X}$ can uniquely be represented as

$$u = x + x^{\perp} \text{ with } x \in \mathcal{X}' \text{ and } x^{\perp} \in \mathcal{X}'^{\perp},$$

see for instance [31, Section 2.9, Corollary 1].

The mapping $u \mapsto x$ defines a linear operator $P_{\mathcal{X}'} \in \mathcal{L}(\mathcal{X}, \mathcal{X})$ that is called *orthogonal projection* on $\mathcal{X}'$.

**Lemma 2.0.1** (cf. [24, Section 5.16])**.** *Let $\mathcal{X}' \subset \mathcal{X}$ be a closed subspace. The orthogonal projection onto $\mathcal{X}'$ satisfies the following conditions:*

1. *$P_{\mathcal{X}'}$ is self-adjoint, i.e. $P^*_{\mathcal{X}'} = P_{\mathcal{X}'}$,*
2. *$\|P_{\mathcal{X}'}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} = 1$ (if $\mathcal{X}' \neq \{0\}$),*
3. *$I - P_{\mathcal{X}'} = P_{\mathcal{X}'^{\perp}}$,*
4. *$\|u - P_{\mathcal{X}'}u\|_{\mathcal{X}} \leqslant \|u - v\|_{\mathcal{X}}$ for all $v \in \mathcal{X}'$,*
5. *$x = P_{\mathcal{X}'}u$ if and only if $x \in \mathcal{X}'$ and $u - x \in \mathcal{X}'^{\perp}$.*

**Remark 2.0.2.** Note that for a non-closed subspace $\mathcal{X}'$ we only have $(\mathcal{X}'^{\perp})^{\perp} = \overline{\mathcal{X}'}$. For $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ we therefore have

- $\mathcal{R}(A)^{\perp} = \mathcal{N}(A^*)$ and thus $\mathcal{N}(A^*)^{\perp} = \overline{\mathcal{R}(A)}$,
- $\mathcal{R}(A^*)^{\perp} = \mathcal{N}(A)$ and thus $\mathcal{N}(A)^{\perp} = \overline{\mathcal{R}(A^*)}$.

Hence, we can deduce the following orthogonal decompositions

$$\mathcal{X} = \mathcal{N}(A) \oplus \overline{\mathcal{R}(A^*)} \text{ and } \mathcal{Y} = \mathcal{N}(A^*) \oplus \overline{\mathcal{R}(A)}.$$

We will also need the follwoing relationship between the ranges of $A^*$ and $A^*A$.

**Lemma 2.0.3.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then $\overline{\mathcal{R}(A^*A)} = \overline{\mathcal{R}(A^*)}$.*

*Proof.* It is clear that $\overline{\mathcal{R}(A^*A)} = \overline{\mathcal{R}(A^*|_{\mathcal{R}(A)})} \subseteq \overline{\mathcal{R}(A^*)}$, so we are left to prove that $\overline{\mathcal{R}(A^*)} \subseteq \overline{\mathcal{R}(A^*A)}$.

Let $u \in \overline{\mathcal{R}(A^*)}$ and let $\varepsilon > 0$. Then, there exists $f \in \mathcal{N}(A^*)^{\perp} = \overline{\mathcal{R}(A)}$ with $\|A^*f - u\|_{\mathcal{X}} < \varepsilon/2$ (recall the orthogonal decomposition in Remark 2.0.2). As $\mathcal{N}(A^*)^{\perp} = \overline{\mathcal{R}(A)}$, there exists $x \in \mathcal{X}$ such that $\|Ax - f\|_{\mathcal{Y}} < \varepsilon/(2\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})})$. Putting these together we have

$$\begin{aligned}\|A^*Ax - u\|_{\mathcal{X}} &\leqslant \|A^*Ax - A^*f\|_{\mathcal{X}} + \|A^*f - u\|_{\mathcal{X}} \\ &\leqslant \underbrace{\|A^*\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})}\|Ax - f\|_{\mathcal{Y}}}_{<\varepsilon/2} + \underbrace{\|A^*f - u\|_{\mathcal{X}}}_{<\varepsilon/2} < \varepsilon\end{aligned}$$

which shows that $u \in \overline{\mathcal{R}(A^*A)}$ and thus also $\overline{\mathcal{R}(A^*)} \subseteq \overline{\mathcal{R}(A^*A)}$. □

## 2.1 Generalised Inverses

Recall the inverse problem

$$Au = f, \tag{2.1}$$

where $A\colon \mathcal{X} \to \mathcal{Y}$ is a linear bounded operator and $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces.

**Definition 2.1.1** (Minimal-norm solutions)**.** *An element $u \in \mathcal{X}$ is called*

- *a least-squares solution of* (2.1) *if*

$$\|Au - f\|_{\mathcal{Y}} = \inf\{\|Av - f\|_{\mathcal{Y}}, \quad v \in \mathcal{X}\};$$

- *a minimal-norm solution of* (2.1) *(and is denoted by $u^\dagger$) if*

$$\|u^\dagger\|_{\mathcal{X}} \leqslant \|v\|_{\mathcal{X}} \quad \textit{for all least squares solutions } v.$$

**Remark 2.1.2.** Since $\mathcal{R}(A)$ is not closed in general (it is never closed for a compact operator, unless the range is finite-dimensional), a least-squares solution may not exist. If it exists, then the minimal-norm solution is unique (it is the orthogonal projection of the zero element onto an affine subspace defined by $\|Au - f\|_{\mathcal{Y}} = \min\{\|Av - f\|_{\mathcal{Y}}, \quad v \in \mathcal{X}\}$).

In numerical linear algebra it is a well known fact that the normal equations can be used to compute least-squares solutions. The same holds true in the infinite-dimensional case.

**Theorem 2.1.3.** *Let $f \in \mathcal{Y}$ and $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then, the following three assertions are equivalent.*

1. *$u \in \mathcal{X}$ satisfies $Au = P_{\overline{\mathcal{R}(A)}} f$.*
2. *$u$ is a least squares solution of the inverse problem* (2.1)*.*
3. *$u$ solves the* normal equation

$$A^*Au = A^*f. \tag{2.2}$$

**Remark 2.1.4.** The name normal equation is derived from the fact that for any solution $u$ its residual $Au - f$ is orthogonal (normal) to $\mathcal{R}(A)$. This can be readily seen, as we have for any $v \in \mathcal{X}$ that

$$0 = \langle v, A^*(Au - f)\rangle_{\mathcal{X}} = \langle Av, Au - f\rangle_{\mathcal{Y}}$$

which shows $Au - f \in \mathcal{R}(A)^\perp$.

*Proof of Theorem 2.1.3.* For $1 \Rightarrow 2$: Let $u \in \mathcal{X}$ such that $Au = P_{\overline{\mathcal{R}(A)}} f$ and let $v \in \mathcal{X}$ be arbitrary. With the basic properties of the orthogonal projection, Lemma 2.0.1 4, we have

$$\|Au - f\|_{\mathcal{Y}}^2 = \|(I - P_{\overline{\mathcal{R}(A)}})f\|_{\mathcal{Y}}^2 \leqslant \inf_{g \in \overline{\mathcal{R}(A)}} \|g - f\|_{\mathcal{Y}}^2 \leqslant \inf_{g \in \mathcal{R}(A)} \|g - f\|_{\mathcal{Y}}^2 = \inf_{v \in \mathcal{X}} \|Av - f\|_{\mathcal{Y}}^2,$$

which shows that $u$ is a least squares solution.

For $2 \Rightarrow 3$: Let $u \in \mathcal{X}$ be a least squares solution and let $v \in \mathcal{X}$ an arbitrary element. We define the quadratic polynomial $F\colon \mathbb{R} \to \mathbb{R}$,

$$F(\lambda) := \|A(u + \lambda v) - f\|_{\mathcal{Y}}^2 = \lambda^2 \|Av\|_{\mathcal{Y}}^2 - 2\lambda \langle Av, f - Au\rangle_{\mathcal{Y}} + \|f - Au\|_{\mathcal{Y}}^2 .$$

A necessary condition for $u \in \mathcal{X}$ to be a least squares solution is $F'(0) = 0$, which leads to $\langle v, A^*(f - Au)\rangle_{\mathcal{X}} = 0$. As $v$ was arbitrary, it follows that the normal equation (2.2) must hold.

For $3 \Rightarrow 1$: From the normal equation it follows that $A^*(f - Au) = 0$, which is equivalent to $f - Au \in \mathcal{R}(A)^\perp$, see Remark 2.1.4. Since $\mathcal{R}(A)^\perp = \left(\overline{\mathcal{R}(A)}\right)^\perp$ and $Au \in \mathcal{R}(A) \subset \overline{\mathcal{R}(A)}$, the assertion follows from Lemma 2.0.1 5:

$$Au = P_{\overline{\mathcal{R}(A)}} f \Leftrightarrow Au \in \overline{\mathcal{R}(A)} \text{ and } f - Au \in \left(\overline{\mathcal{R}(A)}\right)^\perp.$$

□

**Lemma 2.1.5.** *Let $f \in \mathcal{Y}$ and let $\mathbb{L}$ be the set of least squares solutions to the inverse problem* (2.1)*. Then, $\mathbb{L}$ is non-empty if and only if $f \in \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$.*

*Proof.* Let $u \in \mathbb{L}$. It is easy to see that $f = Au + (f - Au) \in \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$ as the normal equations are equivalent to $f - Au \in \mathcal{R}(A)^\perp$.

Consider now $f \in \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$. Then there exists $u \in \mathcal{X}$ and $g \in \mathcal{R}(A)^\perp = \left(\overline{\mathcal{R}(A)}\right)^\perp$ such that $f = Au + g$ and thus $P_{\overline{\mathcal{R}(A)}} f = P_{\overline{\mathcal{R}(A)}} Au + P_{\overline{\mathcal{R}(A)}} g = Au$ and the assertion follows from Theorem 2.1.3 1. □

**Remark 2.1.6.** If the dimensions of $\mathcal{X}$ and $\mathcal{R}(A)$ are finite, then $\mathcal{R}(A)$ is closed, i.e. $\overline{\mathcal{R}(A)} = \mathcal{R}(A)$. Thus, in a finite dimensional setting, there always exists a least squares solution.

**Theorem 2.1.7.** *Let $f \in \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$. Then there exists a unique minimal norm solution $u^\dagger$ to the inverse problem* (2.1) *and all least squares solutions are given by $\{u^\dagger\} + \mathcal{N}(A)$.*

*Proof.* From Lemma 2.1.5 we know that there exists a least squares solution. As noted in Remark 2.1.2, in this case the minimal-norm solution is unique. Let $\varphi$ be an arbitrary least-squares solution. Using Theorem 2.1.3 we get

$$A(\varphi - u^\dagger) = A\varphi - Au^\dagger = P_{\overline{\mathcal{R}(A)}} f - P_{\overline{\mathcal{R}(A)}} f = 0, \tag{2.3}$$

which shows that $\varphi - u^\dagger \in \mathcal{N}(A)$, hence the assertion. □

If a least-squares solution exists for a given $f \in \mathcal{Y}$ then the minimal-norm solution can be computed (at least in theory) using the Moore-Penrose generalised inverse.

**Definition 2.1.8.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and let*

$$\widetilde{A} \coloneqq A|_{\mathcal{N}(A)^\perp} : \mathcal{N}(A)^\perp \to \mathcal{R}(A)$$

*denote the restriction of $A$ to $\mathcal{N}(A)^\perp$. The* Moore-Penrose inverse *$A^\dagger$ is defined as the unique linear extension of $\widetilde{A}^{-1}$ to*

$$\mathcal{D}(A^\dagger) = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$$

*with*

$$\mathcal{N}(A^\dagger) = \mathcal{R}(A)^\perp.$$

**Remark 2.1.9.** Due to the restriction to $\mathcal{N}(A)^\perp$ and $\mathcal{R}(A)$ we have that $\widetilde{A}$ is injective and surjective. Hence, $\widetilde{A}^{-1}$ exists and is linear and – as a consequence – $A^\dagger$ is well-defined on $\mathcal{R}(A)$.

Moreover, due to the orthogonal decomposition $\mathcal{D}(A^\dagger) = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$, there exist for arbitrary $f \in \mathcal{D}(A^\dagger)$ elements $f_1 \in \mathcal{R}(A)$ and $f_2 \in \mathcal{R}(A)^\perp$ with $f = f_1 + f_2$. Therefore, we have

$$A^\dagger f = A^\dagger f_1 + A^\dagger f_2 = A^\dagger f_1 = \widetilde{A}^{-1} f_1 = \widetilde{A}^{-1} P_{\overline{\mathcal{R}(A)}} f\,, \tag{2.4}$$

where we used that $f_2 \in \mathcal{R}(A)^\perp = \mathcal{N}(A^\dagger)$. Thus, $A^\dagger$ is well-defined on the entire domain $\mathcal{D}(A^\dagger)$.

**Remark 2.1.10.** As orthogonal complements are always closed we get that

$$\overline{\mathcal{D}(A^\dagger)} = \overline{\mathcal{R}(A)} \oplus \mathcal{R}(A)^\perp = \mathcal{Y},$$

and hence, $\mathcal{D}(A^\dagger)$ is dense in $\mathcal{Y}$. Thus, if $\mathcal{R}(A)$ is closed it follows that $\mathcal{D}(A^\dagger) = \mathcal{Y}$ and on the other hand, $\mathcal{D}(A^\dagger) = \mathcal{Y}$ implies $\mathcal{R}(A)$ is closed. We note that for ill-posed problems $\mathcal{R}(A)$ is usually not closed; for instance, if $A$ is compact then $\mathcal{R}(A)$ is closed if and only if it is finite-dimensional [1, Ex.1 Section 7.1].

If $A$ is bijective we have that $A^\dagger = A^{-1}$. We also highlight that the extension $A^\dagger$ is not necessarily continuous.

**Theorem 2.1.11** ([16, Prop. 2.4])**.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then $A^\dagger$ is continuous, i.e. $A^\dagger \in \mathcal{L}(\mathcal{D}(A^\dagger), \mathcal{X})$, if and only if $\mathcal{R}(A)$ is closed.*

**Example 2.1.12.** To illustrate the definition of the Moore-Penrose inverse we consider a simple example in finite dimensions. Let the linear operator $A\colon \mathbb{R}^3 \to \mathbb{R}^2$ be given by

$$Ax = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 0 \end{pmatrix}.$$

It is easy to see that $\mathcal{R}(A) = \{f \in \mathbb{R}^2 \mid f_2 = 0\}$ and $\mathcal{N}(A) = \{x \in \mathbb{R}^3 \mid x_1 = 0\}$. Thus, $\mathcal{N}(A)^\perp = \{x \in \mathbb{R}^3 \mid x_2, x_3 = 0\}$. Therefore, $\widetilde{A}\colon \mathcal{N}(A)^\perp \to \mathcal{R}(A)$, given by $x \mapsto (2x_1, 0)^\top$, is bijective and its inverse $\widetilde{A}^{-1}\colon \mathcal{R}(A) \to \mathcal{N}(A)^\perp$ is given by $f \mapsto (f_1/2, 0, 0)^\top$.

To get the Moore-Penrose inverse $A^\dagger$, we need to extend $\widetilde{A}^{-1}$ to $\mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$ in such a way that $A^\dagger f = 0$ for all $f \in \mathcal{R}(A)^\perp = \{f \in \mathbb{R}^2 \mid f_1 = 0\}$. It is easy to see that the Moore-Penrose inverse $A^\dagger\colon \mathbb{R}^2 \to \mathbb{R}^3$ is given by the following expression

$$A^\dagger f = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} f_1/2 \\ 0 \\ 0 \end{pmatrix}.$$

Let us consider data $\widetilde{f} = (8, 1)^\top \notin \mathcal{R}(A)$. Then, $A^\dagger \widetilde{f} = A^\dagger (8, 1)^\top = (4, 0, 0)^\top$.

It can be shown that $A^\dagger$ can be characterised by the Moore-Penrose equations.

**Lemma 2.1.13** ([16, Prop. 2.3])**.** *The Moore-Penrose inverse $A^\dagger$ satisfies $\mathcal{R}(A^\dagger) = \mathcal{N}(A)^\perp$ and the Moore-Penrose equations*

1. $AA^\dagger A = A$,
2. $A^\dagger AA^\dagger = A^\dagger$,
3. $A^\dagger A = I - P_{\mathcal{N}(A)}$,
4. $AA^\dagger = P_{\overline{\mathcal{R}(A)}}\Big|_{\mathcal{D}(A^\dagger)}$,

*where $P_{\mathcal{N}(A)}$ and $P_{\overline{\mathcal{R}(A)}}$ denote the orthogonal projections on $\mathcal{N}(A)$ and $\overline{\mathcal{R}(A)}$, respectively.*

The next theorem shows that minimal-norm solutions can indeed be computed using the Moore-Penrose generalised inverse.

**Theorem 2.1.14.** *For each $f \in \mathcal{D}(A^\dagger)$, the minimal norm solution $u^\dagger$ to the inverse problem* (2.1) *is given via*

$$u^\dagger = A^\dagger f.$$

*Proof.* As $f \in \mathcal{D}(A^\dagger)$, we know from Theorem 2.1.7 that the minimal norm solution $u^\dagger$ exists and is unique. With $u^\dagger \in \mathcal{N}(A)^\perp$, Lemma 2.1.13, and Theorem 2.1.3 we conclude that

$$u^\dagger = (I - P_{\mathcal{N}(A)})u^\dagger = A^\dagger Au^\dagger = A^\dagger P_{\overline{\mathcal{R}(A)}} f = A^\dagger AA^\dagger f = A^\dagger f.$$

□

As a consequence of Theorem 2.1.14 and Theorem 2.1.3, we find that the minimum norm solution $u^\dagger$ of $Au = f$ is a minimum norm solution of the normal equation (2.2), i.e.

$$u^\dagger = (A^*A)^\dagger A^* f.$$

Thus, in order to compute $u^\dagger$ we can equivalently consider finding the minimum norm solution of the normal equation.

## 2.2 Compact Operators

**Definition 2.2.1.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then $A$ is said to be* compact *if for any bounded set $B \subset \mathcal{X}$ the closure of its image $\overline{A(B)}$ is compact in $\mathcal{Y}$. We denote the space of compact operators by $\mathcal{K}(\mathcal{X}, \mathcal{Y})$.*

**Remark 2.2.2.** We can equivalently define an operator $A$ to be compact if the image of a bounded sequence $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ contains a convergent subsequence $\{Au_{j_k}\}_{k\in\mathbb{N}} \subset \mathcal{Y}$.

Compact operators are very common in inverse problems. In fact, almost all (linear) inverse problems involve the inversion of a compact operator. As the following result shows, compactness of the forward operator is a major source if ill-posedness.

**Theorem 2.2.3.** *Let $A \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$ with an infinite dimensional range. Then, the Moore-Penrose inverse of $A$ is discontinuous.*

*Proof.* As the range $\mathcal{R}(A)$ is of infinite dimension, we can conclude that $\mathcal{X}$ and $\mathcal{N}(A)^\perp$ are also infinite dimensional. We can therefore find a sequence $\{u_j\}_{j\in\mathbb{N}}$ with $u_j \in \mathcal{N}(A)^\perp$, $\|u_j\|_\mathcal{X} = 1$ and $\langle u_j, u_k\rangle_\mathcal{X} = 0$ for $j \neq k$. Since $A$ is a compact operator the sequence

$f_j = Au_j$ has a convergent subsequence, hence, for all $\delta > 0$ we can find $j, k$ such that $\|f_j - f_k\|_{\mathcal{Y}} < \delta$. However, we also obtain

$$\begin{aligned}\|A^\dagger f_j - A^\dagger f_k\|_{\mathcal{X}}^2 &= \|A^\dagger A u_j - A^\dagger A u_k\|_{\mathcal{X}}^2 \\ &= \|u_j - u_k\|_{\mathcal{X}}^2 = \|u_j\|_{\mathcal{X}}^2 - 2\langle u_j, u_k\rangle_{\mathcal{X}} + \|u_k\|_{\mathcal{X}}^2 = 2,\end{aligned}$$

which shows that $A^\dagger$ is discontinuous. Here, the second identity follows from Lemma 2.1.13 3 and the fact that $u_j, u_k \in \mathcal{N}(A)^\perp$. □

To have a better understanding of when we have $f \in \overline{\mathcal{R}(A)} \setminus \mathcal{R}(A)$ for compact operators $A$, we want to consider the singular value decomposition of compact operators.

### Singular value decomposition of compact operators

**Theorem 2.2.4** ([20, p. 225, Theorem 9.16])**.** *Let $\mathcal{X}$ be a Hilbert space and $A \in \mathcal{K}(\mathcal{X}, \mathcal{X})$ be self-adjoint. Then there exists an orthonormal basis $\{x_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ of $\overline{\mathcal{R}(A)}$ and a sequence of eigenvalues $\{\lambda_j\}_{j\in\mathbb{N}} \subset \mathbb{R}$ with $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots > 0$ such that for all $u \in \mathcal{X}$ we have*

$$Au = \sum_{j=1}^{\infty} \lambda_j \langle u, x_j\rangle_{\mathcal{X}}\, x_j\,.$$

*The sequence $\{\lambda_j\}_{j\in\mathbb{N}}$ is either finite or we have $\lambda_j \to 0$.*

**Remark 2.2.5.** The notation in the theorem above only makes sense if the sequence $\{\lambda_j\}_{j\in\mathbb{N}}$ is infinite. For the case that there are only finitely many $\lambda_j$ the sum has to be interpreted as a finite sum.

Moreover, as the eigenvalues are sorted by absolute value $|\lambda_j|$, we have $\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} = |\lambda_1|$.

If $A$ is not self-adjoint, the decomposition in Theorem 2.2.4 does not hold any more. Instead, we can consider the so-called *singular value decomposition* of a compact linear operator.

**Theorem 2.2.6.** *Let $A \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$. Then there exists*

1. *a not-necessarily infinite null sequence $\{\sigma_j\}_{j\in\mathbb{N}}$ with $\sigma_1 \geqslant \sigma_2 \geqslant \ldots > 0$,*
2. *an orthonormal basis $\{x_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ of $\mathcal{N}(A)^\perp$,*
3. *an orthonormal basis $\{y_j\}_{j\in\mathbb{N}} \subset \mathcal{Y}$ of $\overline{\mathcal{R}(A)}$ with*

$$Ax_j = \sigma_j y_j, \quad A^* y_j = \sigma_j x_j, \quad \text{for all } j \in \mathbb{N}. \tag{2.5}$$

*Moreover, for all $u \in \mathcal{X}$ we have the representation*

$$Au = \sum_{j=1}^{\infty} \sigma_j \langle u, x_j\rangle\, y_j. \tag{2.6}$$

*The sequence $\{(\sigma_j, x_j, y_j)\}$ is called* singular system *or* singular value decomposition (SVD) *of $A$.*

*For the adjoint operator $A^*$ we have the representation*

$$A^* f = \sum_{j=1}^{\infty} \sigma_j \langle f, y_j\rangle\, x_j \quad \forall f \in \mathcal{Y}. \tag{2.7}$$

*Proof.* Consider $B = A^*A$ and $C = AA^*$. Both $B$ and $C$ are compact, self-adjoint and even positive semidefinite, so that by Theorem 2.2.4 both admit a spectral representation and, by positive semidefiniteness, their eigenvalues are positive, i.e.

$$Bu = \sum_{j=1}^{\infty} \sigma_j^2 \langle u, x_j \rangle x_j \quad \forall u \in \mathcal{X}, \quad Cf = \sum_{j=1}^{\infty} \widetilde{\sigma}_j^2 \langle f, y_j \rangle y_j \quad \forall f \in \mathcal{Y},$$

where $\{x_j\}$ and $\{y_j\}$ are orthonormal bases of $\overline{\mathcal{R}(A^*A)}$ and $\overline{\mathcal{R}(AA^*)}$, respectively, and $\sigma_j, \widetilde{\sigma}_j > 0$ for all $j$. As pointed out in Remark 2.0.2 and Lemma 2.0.3, we have $\overline{\mathcal{R}(A^*A)} = \overline{\mathcal{R}(A^*)} = \mathcal{N}(A)^\perp$ and, therefore, $\{x_j\}$ is also a basis of $\mathcal{N}(A)^\perp$. Analogously, $\{y_j\}$ is also a basis of $\overline{\mathcal{R}(A)}$.

Since $\widetilde{\sigma}_j^2$ is an eigenvalue of $C$ for the eigenvector $y_j$, we get that

$$\widetilde{\sigma}_j^2 A^* y_j = A^*(\widetilde{\sigma}_j^2 y_j) = A^* C y_j = A^* A A^* y_j = B A^* y_j$$

and therefore $\widetilde{\sigma}_j^2$ is also an eigenvalue of $B$ (for the eigenvector $A^*y_j$). Hence, with no loss of generality we can assume that $\widetilde{\sigma}_j = \sigma_j$. We further observe that $\left\{ \frac{A^*y_j}{\sigma_j} \right\}$ form an orthonormal basis of $\overline{\mathcal{R}(A^*)} = \mathcal{N}(A)^\perp$, since

$$\left\langle \frac{A^*y_j}{\sigma_j}, \frac{A^*y_k}{\sigma_k} \right\rangle = \frac{1}{\sigma_j \sigma_k} \langle y_j, AA^*y_k \rangle = \frac{1}{\sigma_j \sigma_k} \left\langle y_j, \sigma_k^2 y_k \right\rangle = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, we can choose $\{x_j\}$ to be

$$x_j = \sigma_j^{-1} A^* y_j$$

and we get that

$$A^* y_j = \sigma_j x_j.$$

We also observe that

$$Ax_j = \sigma_j^{-1} AA^* y_j = \sigma_j^{-1} \sigma_j^2 y_j = \sigma_j y_j,$$

which proves (2.5).

Extending the basis $\{x_j\}$ of $\overline{\mathcal{R}(A^*)}$ to a basis of $\mathcal{X}$, we expand an arbitrary $u \in \mathcal{X}$ as $u = \sum_{j=1}^{\infty} \langle u, x_j \rangle x_j$ and, since $\mathcal{X} = \mathcal{N}(A) \oplus \overline{\mathcal{R}(A^*)}$ (Remark 2.0.2), obtain the singular value decompositions (2.6) – (2.7)

$$Au = \sum_{j=1}^{\infty} \sigma_j \langle u, x_j \rangle y_j \quad \forall u \in \mathcal{X}, \quad A^* f = \sum_{j=1}^{\infty} \sigma_j \langle f, y_j \rangle x_j \quad \forall f \in \mathcal{Y}.$$

□

We can now derive a representation of the Moore-Penrose inverse in terms of the singular value decomposition.

**Theorem 2.2.7.** *Let $A \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$ with singular system $\{(\sigma_j, x_j, y_j)\}_{j \in \mathbb{N}}$ and $f \in \mathcal{D}(A^\dagger)$. Then the Moore-Penrose inverse of $A$ can be written as*

$$A^\dagger f = \sum_{j=1}^{\infty} \sigma_j^{-1} \langle f, y_j \rangle x_j \,. \tag{2.8}$$

*Proof.* We know that, since $f \in \mathcal{D}(A^\dagger)$, $u^\dagger = A^\dagger f$ solves the normal equations

$$A^* A u^\dagger = A^* f.$$

From Theorem 2.2.6 we know that

$$A^* A u^\dagger = \sum_{j=1}^{\infty} \sigma_j^2 \left\langle u^\dagger, x_j \right\rangle x_j, \quad A^* f = \sum_{j=1}^{\infty} \sigma_j \langle f, y_j \rangle x_j, \tag{2.9}$$

which implies that

$$\left\langle u^\dagger, x_j \right\rangle = \sigma_j^{-1} \langle f, y_j \rangle$$

Expanding $u^\dagger \in \mathcal{N}(A)^\perp$ in the basis $\{x_j\}$, we get

$$u^\dagger = \sum_{j=1}^{\infty} \left\langle u^\dagger, x_j \right\rangle x_j = \sum_{j=1}^{\infty} \sigma_j^{-1} \langle f, y_j \rangle x_j = A^\dagger f.$$

□

The representation (2.8) makes it clear again that the Moore-Penrose inverse is unbounded if $\mathcal{R}(A)$ is infinite dimensional. Indeed, taking the sequence $y_j$ we note that $\|A^\dagger y_j\| = \sigma_j^{-1} \to \infty$, although $\|y_j\| = 1$.

The unboundedness of the Moore-Penrose inverse is also reflected in the fact that the series in (2.8) may not converge for a given $f$. The convergence criterion for the series is called the *Picard criterion*.

**Definition 2.2.8.** *We say that the data $f$ satisfy the Picard criterion, if*

$$\|A^\dagger f\|^2 = \sum_{j=1}^{\infty} \frac{|\langle f, y_j \rangle|^2}{\sigma_j^2} < \infty. \tag{2.10}$$

**Remark 2.2.9.** The Picard criterion is a condition on the decay of the coefficients $\langle f, y_j \rangle$. As the singular values $\sigma_j$ decay to zero as $j \to \infty$, the Picard criterion is only met if the coefficients $\langle f, y_j \rangle$ decay sufficiently fast.

In case the singular system is given by the Fourier basis, then the coefficients $\langle f, y_j \rangle$ are just the Fourier coefficients of $f$. Therefore, the Picard criterion is a condition on the decay of the Fourier coefficients which is equivalent to the smoothness of $f$.

It turns our that the Picard criterion also can be used to characterise elements in the range of the forward operator.

**Theorem 2.2.10.** *Let $A \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$ with singular system $\{(\sigma_j, x_j, y_j)\}_{j \in \mathbb{N}}$, and $f \in \overline{\mathcal{R}(A)}$. Then $f \in \mathcal{R}(A)$ if and only if the* Picard criterion

$$\sum_{j=1}^{\infty} \frac{\left|\langle f, y_j \rangle_{\mathcal{Y}}\right|^2}{\sigma_j^2} < \infty \tag{2.11}$$

*is met.*

*Proof.* Let $f \in \mathcal{R}(A)$, thus there is a $u \in \mathcal{X}$ such that $Au = f$. It is easy to see that we have

$$\langle f, y_j \rangle_{\mathcal{Y}} = \langle Au, y_j \rangle_{\mathcal{Y}} = \langle u, A^* y_j \rangle_{\mathcal{X}} = \sigma_j \langle u, x_j \rangle_{\mathcal{X}}$$

and therefore

$$\sum_{j=1}^{\infty} \sigma_j^{-2} |\langle f, y_j \rangle_{\mathcal{Y}}|^2 = \sum_{j=1}^{\infty} |\langle u, x_j \rangle_{\mathcal{X}}|^2 \leqslant \|u\|_{\mathcal{X}}^2 < \infty .$$

Now let the Picard criterion (2.11) hold and define $u := \sum_{j=1}^{\infty} \sigma_j^{-1} \langle f, y_j \rangle_{\mathcal{Y}} x_j \in \mathcal{X}$. It is well-defined by the Picard criterion (2.11) and we conclude

$$Au = \sum_{j=1}^{\infty} \sigma_j^{-1} \langle f, y_j \rangle_{\mathcal{Y}} Ax_j = \sum_{j=1}^{\infty} \langle f, y_j \rangle_{\mathcal{Y}} y_j = P_{\overline{\mathcal{R}(A)}} f = f ,$$

which shows $f \in \mathcal{R}(A)$. □

Although all ill-posed problems are not easy to solve, some are worse than others, depending on how fast the singular values decay to zero.

**Definition 2.2.11.** *We say that an ill-posed inverse problem* (2.1) *is* mildly ill-posed *if the singular values decay at most with polynomial speed, i.e. there exist $\gamma, C > 0$ such that $\sigma_j \geqslant C j^{-\gamma}$ for all $j$. We call the ill-posed inverse problem* severely ill-posed *if its singular values decay faster than with polynomial speed, i.e. for all $\gamma, C > 0$ one has that $\sigma_j \leqslant C j^{-\gamma}$ for $j$ sufficiently large.*

**Example 2.2.12.** Let us consider the example of differentiation again, as introduced in Section 1.2.3. The forward operator $A \colon L^2([0,1]) \to L^2([0,1])$ in this problem is given by

$$(Au)(t) = \int_0^t u(s)\, ds = \int_0^1 K(s,t) u(s)\, ds ,$$

with $K \colon [0,1] \times [0,1] \to \mathbb{R}$ defined as

$$K(s,t) := \begin{cases} 1 & s \leqslant t \\ 0 & \text{else} \end{cases} .$$

This is a special case of the integral operators as introduced in Section 1.2.1. Since the kernel $K$ is square integrable, $A$ is compact.

The adjoint operator $A^*$ is given via

$$(A^* f)(s) = \int_0^1 K(t,s) f(t)\, dt = \int_s^1 v(t)\, dt . \tag{2.12}$$

Now we want to compute the eigenvalues and eigenvectors of $A^*A$, i.e. we look for $\sigma^2$ and $x \in L^2([0,1])$ with

$$\sigma^2 x(s) = (A^* A x)(s) = \int_s^1 \int_0^t x(r)\, dr\, dt .$$

We immediately observe $x(1) = 0$ and further

$$\sigma^2 x'(s) = \frac{d}{ds} \int_s^1 \int_0^t x(r)\, dr\, dt = -\int_0^s x(r)\, dr\,,$$

from which we conclude $x'(0) = 0$. Taking the derivative another time thus yields the ordinary differential equation

$$\sigma^2 x''(s) + x(s) = 0\,,$$

for which solutions are of the form

$$x(s) = c_1 \sin(\sigma^{-1} s) + c_2 \cos(\sigma^{-1} s)\,,$$

with some constants $c_1, c_2$. In order to satisfy the boundary conditions $x(1) = c_1 \sin(\sigma^{-1}) + c_2 \cos(\sigma^{-1}) = 0$ and $x'(0) = c_1 = 0$, we chose $c_1 = 0$ and $\sigma$ such that $\cos(\sigma^{-1}) = 0$. Hence, we have

$$\sigma_j = \frac{2}{(2j-1)\pi} \text{ for } j \in \mathbb{N}\,,$$

and by choosing $c_2 = \sqrt{2}$ we obtain the following normalised representation of $x_j$:

$$x_j(s) = \sqrt{2} \cos\left(\left(j - \frac{1}{2}\right)\pi s\right)\,.$$

According to (2.5) we further obtain

$$y_j(s) = \sigma_j^{-1}(Ax_j)(s) = \left(j - \frac{1}{2}\right)\pi \int_0^s \sqrt{2} \cos\left(\left(j - \frac{1}{2}\right)\pi t\right)\, dt = \sqrt{2} \sin\left(\left(j - \frac{1}{2}\right)\pi s\right)\,,$$

and hence, for $f \in L^2([0,1])$ the Picard criterion becomes

$$2 \sum_{j=1}^{\infty} \sigma_j^{-2} \left(\int_0^1 f(s) \sin\left(\sigma_j^{-1} s\right)\, ds\right)^2 < \infty\,.$$

Expanding $f$ in the basis $\{y_j\}$

$$f(t) = 2 \sum_{j=1}^{\infty} \left(\int_0^1 f(s) \sin\left(\sigma_j^{-1} s\right)\, ds\right) \sin\left(\sigma_j^{-1} t\right)$$

and formally differentiating the series, we obtain

$$f'(t) = 2 \sum_{j=1}^{\infty} \sigma_j^{-1} \left(\int_0^1 f(s) \sin\left(\sigma_j^{-1} s\right)\, ds\right) \cos\left(\sigma_j^{-1} t\right)\,.$$

Therefore, the Picard criterion is nothing but the condition for the legitimacy of such differentiation, i.e. for the differentiability of the Fourier series by differentiating its components, and it holds if $f$ is differentiable and $f' \in L^2([0,1])$.

From the decay of the singular values we see that this inverse problem is mildly ill-posed.

# Chapter 3

# Classical Regularisation Theory

## 3.1 What is Regularisation?

We have seen that the Moore-Penrose inverse $A^\dagger$ is unbounded if $\mathcal{R}(A)$ is not closed. Therefore, given noisy data $f_\delta$ such that $\|f_\delta - f\| \leqslant \delta$, we cannot expect convergence $A^\dagger f_\delta \to A^\dagger f$ as $\delta \to 0$. To achieve convergence, we replace $A^\dagger$ with a family of well-posed (bounded) operators $R_\alpha$ with $\alpha = \alpha(\delta, f_\delta)$ and require that $R_{\alpha(\delta,f_\delta)}(f_\delta) \to A^\dagger f$ for all $f \in \mathcal{D}(A^\dagger)$ and all $f_\delta \in \mathcal{Y}$ s.t. $\|f - f_\delta\|_\mathcal{Y} \leqslant \delta$ as $\delta \to 0$.

**Definition 3.1.1.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ be a bounded operator. A family $\{R_\alpha\}_{\alpha>0}$ of continuous operators is called* regularisation *(or* regularisation operator*) of $A^\dagger$ if*

$$R_\alpha f \to A^\dagger f = u^\dagger$$

*for all $f \in \mathcal{D}(A^\dagger)$ as $\alpha \to 0$.*

**Definition 3.1.2.** *If the family $\{R_\alpha\}_{\alpha>0}$ consists of linear operators, then one speaks of* linear regularisation *of $A^\dagger$.*

Hence, a regularisation is a pointwise approximation of the Moore–Penrose inverse with continuous operators. As in the interesting cases the Moore–Penrose inverse may not be continuous we cannot expect that the norm of $R_\alpha$ stays bounded as $\alpha \to 0$. This is confirmed by the following results (in the linear case).

**Theorem 3.1.3** (Banach–Steinhaus e.g. [10, p. 78], [32, p. 173])**.** *Let $\mathcal{X}, \mathcal{Y}$ be Hilbert spaces and $\{A_j\}_{j\in\mathbb{N}} \subset \mathcal{L}(\mathcal{X}, \mathcal{Y})$ a family of point-wise bounded operators, i.e. for all $u \in \mathcal{X}$ there exists a constant $C(u) > 0$ s.t. $\sup_{j\in\mathbb{N}} \|A_j u\|_\mathcal{Y} \leqslant C(u)$. Then*

$$\sup_{j\in\mathbb{N}} \|A_j\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} < \infty\,.$$

**Corollary 3.1.4** ([32, p. 174])**.** Let $\mathcal{X}, \mathcal{Y}$ be Hilbert spaces and $\{A_j\}_{j\in\mathbb{N}} \subset \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then the following two conditions are equivalent:

1. There exists $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ such that

$$Au = \lim_{j\to\infty} A_j u \quad \text{for all } u \in \mathcal{X}.$$

2. There is a dense subset $\mathcal{X}' \subset \mathcal{X}$ such that $\lim_{j\to\infty} A_j u$ exists for all $u \in \mathcal{X}'$ and

$$\sup_{j\in\mathbb{N}} \|A_j\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} < \infty\,.$$

**Theorem 3.1.5.** *Let $\mathcal{X}$, $\mathcal{Y}$ be Hilbert spaces, $A \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ and $\{R_\alpha\}_{\alpha>0}$ a linear regularisation as defined in Definition 3.1.2. If $A^\dagger$ is not continuous, $\{R_\alpha\}_{\alpha>0}$ cannot be uniformly bounded. In particular this implies the existence of an element $f \in \mathcal{Y}$ with $\|R_\alpha f\| \to \infty$ for $\alpha \to 0$.*

*Proof.* We prove the theorem by contradiction and assume that $\{R_\alpha\}_{\alpha>0}$ is uniformly bounded. Hence, there exists a constant $C$ with $\|R_\alpha\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \leqslant C$ for all $\alpha > 0$. Due to Definition 3.1.1, we have $R_\alpha \to A^\dagger$ on $\mathcal{D}(A^\dagger)$. Since $\mathcal{D}(A^\dagger)$ is dense in $\mathcal{Y}$, by Corollary 3.1.4 we get that $A^\dagger \in \mathcal{L}(\mathcal{Y},\mathcal{X})$, which is a contradiction to the assumption that $A^\dagger$ is not continuous.

It remains to show the existence of an element $f \in \mathcal{Y}$ with $\|R_\alpha f\|_{\mathcal{Y}} \to \infty$ for $\alpha \to 0$. If such an element would not exist, then $\{R_\alpha\}_{\alpha>0}$. would be point-wise bounded for all $f \in \mathcal{Y}$. However, Theorem 3.1.3 then implies that $\{R_\alpha\}_{\alpha>0}$ has to be uniformly bounded, which contradicts the first part of the proof. □

With the additional assumption that $\|AR_\alpha\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$ is bounded, we can even show that $R_\alpha f$ diverges for all $f \notin \mathcal{D}(A^\dagger)$.

**Theorem 3.1.6.** *Let $A \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ and $\{R_\alpha\}_{\alpha>0}$ be a linear regularisation of $A^\dagger$. If*

$$\sup_{\alpha>0} \|AR_\alpha\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} < \infty\,,$$

*then $\|R_\alpha f\|_{\mathcal{X}} \to \infty$ for $f \notin \mathcal{D}(A^\dagger)$.*

*Proof.* Define $u_\alpha := R_\alpha f$ for $f \notin \mathcal{D}(A^\dagger)$. Assume that there exists a sequence $\alpha_k \to 0$ such that $\|u_{\alpha_k}\|_{\mathcal{X}}$ is uniformly bounded. Since bounded sets in a Hilbert space are weakly pre-compact, there exists a weakly convergent subsequence $u_{\alpha_{k_l}}$ with some limit $u \in \mathcal{X}$, cf. [18, Section 2.2, Theorem 2.1]. As continuous linear operators are also weakly continuous, we further have $Au_{\alpha_{k_l}} \rightharpoonup Au$. On the other hand, for any $f \in \mathcal{D}(A^\dagger)$ we have that $AR_\alpha f \to AA^\dagger f = P_{\overline{\mathcal{R}(A)}} f$. By Corollary 3.1.4 we then conclude that this also holds for any $f \in \mathcal{Y}$, i.e. also for $f \notin \mathcal{D}(A^\dagger)$. Therefore, we get that $Au = P_{\overline{\mathcal{R}(A)}} f$. Since $\mathcal{Y} = \overline{\mathcal{R}(A)} \oplus \mathcal{R}(A)^\perp$, we get that $f \in \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp = \mathcal{D}(A^\dagger)$ in contradiction to the assumption $f \notin \mathcal{D}(A^\dagger)$. □

## 3.2 Parameter Choice Rules

We have stated in the beginning of this chapter that we would like to obtain a regularisation that would guarantee that $R_\alpha(f_\delta) \to A^\dagger f$ for all $f \in \mathcal{D}(A^\dagger)$ and all $f_\delta \in \mathcal{Y}$ s.t. $\|f - f_\delta\|_{\mathcal{Y}} \leqslant \delta$ as $\delta \to 0$. This means that the parameter $\alpha$, referred to as the *regularisation parameter*, needs to be chosen as a function of $\delta$ (and perhaps also $f_\delta$) so that $\alpha \to 0$ as $\delta \to 0$ (i.e. we need to regularise less as the data get more precise).

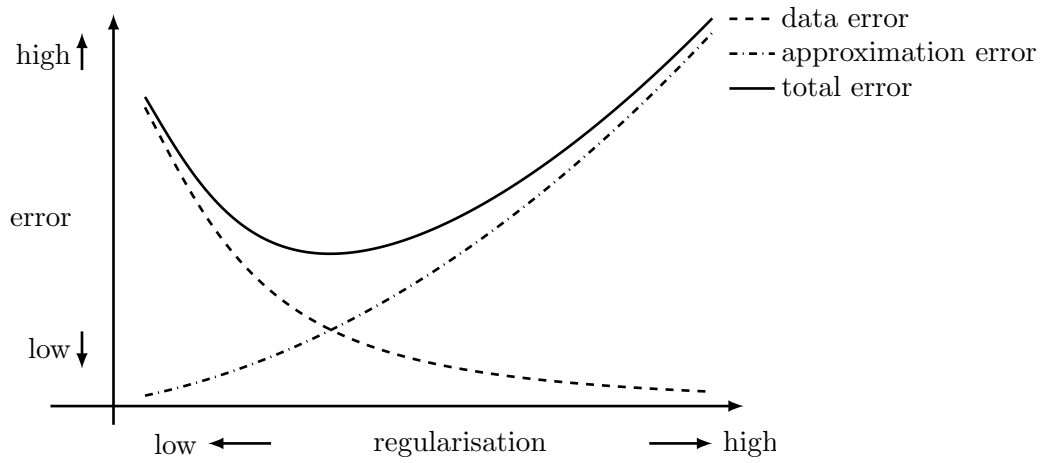



Figure 3.1: The *total error* between a regularised solution and the minimal norm solution decomposes into the *data error* and the *approximation error*. These two errors have opposing trends: For a small regularisation parameter $\alpha$ the error in the data gets amplified through the ill-posedness of the problem and for large $\alpha$ the operator $R_\alpha$ is a poor approximation of the Moore–Penrose inverse.

This can be illustrated with the following observation. For linear regularisations we can split the *total error* between the regularised solution of the noisy problem $R_\alpha f_\delta$ and the minimal norm solution of the noise-free problem $u^\dagger = A^\dagger f$ as

$$\begin{aligned} \|R_\alpha f_\delta - u^\dagger\|_{\mathcal{X}} &\leqslant \|R_\alpha f_\delta - R_\alpha f\|_{\mathcal{X}} + \|R_\alpha f - u^\dagger\|_{\mathcal{X}} \\ &\leqslant \underbrace{\delta \|R_\alpha\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})}}_{\text{data error}} + \underbrace{\|R_\alpha f - A^\dagger f\|_{\mathcal{X}}}_{\text{approximation error}} . \end{aligned} \tag{3.1}$$

The first term of (3.1) is the *data error*; this term unfortunately does not stay bounded for $\alpha \to 0$, which we can conclude from Theorem 3.1.5. The second term, known as the *approximation error*, however vanishes for $\alpha \to 0$, due to the pointwise convergence of $R_\alpha$ to $A^\dagger$. Hence it becomes evident from (3.1) that a good choice of $\alpha$ depends on $\delta$, and needs to be chosen such that the approximation error becomes as small as possible, whilst the data error is being kept at bay. See Figure 3.1 for an illustration.

Parameter choice rules are defined as follows.

**Definition 3.2.1.** *A function* $\alpha\colon \mathbb{R}_{>0} \times \mathcal{Y} \to \mathbb{R}_{>0}$, $(\delta, f_\delta) \mapsto \alpha(\delta, f_\delta)$ *is called a* parameter choice rule. *We distinguish between*

1. *a priori parameter choice rules, which depend on* $\delta$ *only;*

2. *a posteriori parameter choice rules, which depend on both* $\delta$ *and* $f_\delta$*;*

3. *heuristic parameter choice rules, which depend on* $f_\delta$ *only.*

Now we are ready to define a regularisation that ensures the convergence $R_{\alpha(\delta,f_\delta)}(f_\delta) \to A^\dagger f$ as $\delta \to 0$.

**Definition 3.2.2.** *Let $\{R_\alpha\}_{\alpha>0}$ be a regularisation of $A^\dagger$. If for all $f \in \mathcal{D}(A^\dagger)$ there exists a parameter choice rule $\alpha : \mathbb{R}_{>0} \times \mathcal{Y} \to \mathbb{R}_{>0}$ such that*

$$\lim_{\delta\to 0} \sup_{f_\delta :\, \|f-f_\delta\|_\mathcal{Y} \leqslant \delta} \|R_\alpha f_\delta - A^\dagger f\|_\mathcal{X} = 0 \tag{3.2}$$

*and*

$$\lim_{\delta\to 0} \sup_{f_\delta :\, \|f-f_\delta\|_\mathcal{Y} \leqslant \delta} \alpha(\delta, f_\delta) = 0 \tag{3.3}$$

*then the pair $(R_\alpha, \alpha)$ is called a* convergent regularisation.

### 3.2.1 A priori parameter choice rules

First of all we want to discuss a priori parameter choice rules in more detail. Historically, they were the first to be studied. For every regularisation there exists an a priori parameter choice rule and thus a convergent regularisation.

**Theorem 3.2.3** ([16, Prop 3.4])**.** *Let $\{R_\alpha\}_{\alpha>0}$ be a regularisation of $A^\dagger$, for $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Then there exists an a priori parameter choice rule $\alpha = \alpha(\delta)$ such that $(R_\alpha, \alpha)$ is a convergent regularisation.*

For linear regularisations, an important characterisation of a priori parameter choice strategies that lead to convergent regularisation methods is as follows.

**Theorem 3.2.4.** *Let $\{R_\alpha\}_{\alpha>0}$ be a linear regularisation, and $\alpha : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ an a priori parameter choice rule. Then $(R_\alpha, \alpha)$ is a convergent regularisation method if and only if*

*a)* $\lim_{\delta\to 0} \alpha(\delta) = 0$

*b)* $\lim_{\delta\to 0} \delta \|R_{\alpha(\delta)}\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} = 0$

*Proof.* $\Leftarrow$: Let condition a) and b) be fulfilled. From (3.1) we then observe that for any $f \in \mathcal{D}(A^\dagger)$ and $f_\delta \in \mathcal{Y}$ s.t. $\|f - f_\delta\|_\mathcal{Y} \leqslant \delta$

$$\left\| R_{\alpha(\delta)} f_\delta - A^\dagger f \right\|_\mathcal{X} \to 0 \text{ for } \delta \to 0.$$

Hence, $(R_\alpha, \alpha)$ is a convergent regularisation method.
$\Rightarrow$: Now let $(R_\alpha, \alpha)$ be a convergent regularisation method. We prove that conditions 1 and 2 have to follow from this by showing that violation of either one of them leads to a contradiction to $(R_\alpha, \alpha)$ being a convergent regularisation method. If condition a) is violated, (3.3) is violated and hence, $(R_\alpha, \alpha)$ is not a convergent regularisation method. If condition a) is fulfilled but condition b) is violated, there exists a null sequence $\{\delta_k\}_{k\in\mathbb{N}}$ with $\delta_k \|R_{\alpha(\delta_k)}\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \geqslant C > 0$, and hence, we can find a sequence $\{g_k\}_{k\in\mathbb{N}} \subset \mathcal{Y}$ with $\|g_k\|_\mathcal{Y} = 1$ and $\delta_k \|R_{\alpha(\delta_k)} g_k\|_\mathcal{X} \geqslant \widetilde{C}$ for some $\widetilde{C}$. Let $f \in \mathcal{D}(A^\dagger)$ be arbitrary and define $f_k := f + \delta_k g_k$. Then we have on the one hand $\|f - f_k\|_\mathcal{Y} \leqslant \delta_k$, but on the other hand the norm of

$$R_{\alpha(\delta_k)} f_k - A^\dagger f = R_{\alpha(\delta_k)} f - A^\dagger f + \delta_k R_{\alpha(\delta_k)} g_k$$

cannot converge to zero, as the second term $\delta_k R_{\alpha(\delta_k)} g_k$ is bounded from below by a positive constant $C$ by construction. Hence, (3.2) is violated for $f_\delta = f + \delta_k g_k$ and thus, $(R_\alpha, \alpha)$ is not a convergent regularisation method. □

### 3.2.2 A posteriori parameter choice rules

It is easy to convince oneself that if an a priori parameter choice rule $\alpha = \alpha(\delta)$ defines a convergence regularisation then $\widetilde{\alpha} = \alpha(C\delta)$ with any $C > 0$ also defines a convergent regularisation (for linear regularisations, it is a trivial corollary of Theorem 3.2.4). Therefore, from the asymptotic point of view, all these regularisations are equivalent. For a fixed error level $\delta$, however, they can produce very different solutions. Since in practice we have to deal with a typically small, but fixed $\delta$, we would like to have a parameter choice rule that is sensitive to this value. To achieve this, we need to use more information than merely the error level $\delta$ to choose the parameter $\alpha$ and we will obtain this information from the approximate data $f_\delta$.

The basic idea is as follows. Let $f \in \mathcal{D}(A^\dagger)$ and $f_\delta \in \mathcal{Y}$ such that $\|f - f_\delta\| \leqslant \delta$ and consider the *residual* between $f_\delta$ and $u_\alpha := R_\alpha f_\delta$, i.e.

$$\|Au_\alpha - f_\delta\|.$$

Let $u^\dagger$ be the minimal norm solution and define

$$\mu := \inf\{\|Au - f\|,\ u \in \mathcal{X}\} = \|Au^\dagger - f\|.$$

We observe that $u^\dagger$ satisfies the following inequality

$$\|Au^\dagger - f_\delta\| \leqslant \|Au^\dagger - f\| + \|f_\delta - f\| \leqslant \mu + \delta$$

and in some cases this estimate may be sharp. Hence, it appears not to be useful to choose $\alpha(\delta, f_\delta)$ with $\|Au_\alpha - f_\delta\| < \mu + \delta$. In general, it may be not straightforward to estimate $\mu$, but if $\mathcal{R}(A)$ is dense in $\mathcal{Y}$, we get that $\mathcal{R}(A)^\perp = \{0\}$ due to Remark 2.0.2 and $\mu = 0$. Therefore, we ideally ensure that $\mathcal{R}(A)$ is dense.

These observations motivate the Morozov's discrepancy principle, which in the case $\mu = 0$ reads as follows.

**Definition 3.2.5** (Morozov's discrepancy principle)**.** *Let $u_\alpha = R_\alpha f_\delta$ with $\alpha(\delta, f_\delta)$ chosen as follows*

$$\alpha(\delta, f_\delta) = \sup\{\alpha > 0 \mid \|Au_\alpha - f_\delta\| \leqslant \eta\delta\} \tag{3.4}$$

*for given $\delta$, $f_\delta$ and a fixed constant $\eta > 1$. Then $u_{\alpha(\delta,f_\delta)} = R_{\alpha(\delta,f_\delta)} f_\delta$ is said to satisfy Morozov's discrepancy principle.*

It can be shown that the a-posteriori parameter choice rule (3.4) indeed yields a convergent regularization method [16, Chapter 4.3].

### 3.2.3 Heuristic parameter choice rules

As the measurement error $\delta$ is not always easy to obtain in practice, it is tempting to use a parameter choice rule that only depends on the measured data $f_\delta$ and not on their error $\delta$, i.e. to use a heuristic parameter choice rule. Unfortunately, heuristic rules yield convergent regularisations only for well-posed problems, as the following result, known as the Bakushinskii veto [6], demonstrates.

**Theorem 3.2.6** ([16, Thm 3.3])**.** *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $\{R_\alpha\}$ be a regularization for $A^\dagger$. Let $\alpha = \alpha(f_\delta)$ be a parameter choice rule such that $(R_\alpha, \alpha)$ is a convergent regularization. Then $A^\dagger$ is continuous from $\mathcal{Y}$ to $\mathcal{X}$.*

## 3.3 Spectral Regularisation

Recall the spectral representation (2.8) of the Moore-Penrose inverse $A^\dagger$

$$A^\dagger f = \sum_{j=1}^{\infty} \frac{1}{\sigma_j} \langle f, y_j \rangle \, x_j \,,$$

where $\{(\sigma_j, x_j, y_j)\}$ is the singular system of $A$.

The source of ill-posedness of $A^\dagger$ are the eigenvalues $1/\sigma_j$, which explode as $j \to \infty$, since $\sigma_j \to 0$ as $j \to \infty$. Let us construct a regularisation by modifying these eigenvalues as follows

$$R_\alpha f := \sum_{j=1}^{\infty} g_\alpha(\sigma_j) \langle f, y_j \rangle \, x_j \,, \;\; f \in \mathcal{Y}, \tag{3.5}$$

with an appropriate function $g_\alpha \colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $g_\alpha(\sigma) \to \frac{1}{\sigma}$ as $\alpha \to 0$ for all $\sigma > 0$ and

$$g_\alpha(\sigma) \leqslant C_\alpha \text{ for all } \sigma \in \mathbb{R}_+. \tag{3.6}$$

**Theorem 3.3.1.** *Let $g_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be a piecewise continuous function satisfying* (3.6)*, $\lim_{\alpha \to 0} g_\alpha(\sigma) = \frac{1}{\sigma}$ and*

$$\sup_{\alpha, \sigma} \sigma g_\alpha(\sigma) \leqslant \gamma \tag{3.7}$$

*for some constant $\gamma > 0$. If $R_\alpha$ is defined as in* (3.5)*, we have*

$$R_\alpha f \to A^\dagger f \;\; \text{as } \alpha \to 0$$

*for all $f \in \mathcal{D}(A^\dagger)$.*

*Proof.* From the singular value decomposition of $A^\dagger$ and the definition of $R_\alpha$ we obtain

$$R_\alpha f - A^\dagger f = \sum_{j=1}^{\infty} \left( g_\alpha(\sigma_j) - \frac{1}{\sigma_j} \right) \langle f, y_j \rangle_{\mathcal{Y}} \, x_j = \sum_{j=1}^{\infty} (\sigma_j g_\alpha(\sigma_j) - 1) \, \langle u^\dagger, x_j \rangle_{\mathcal{X}} \, x_j \,.$$

Consider

$$\| R_\alpha f - A^\dagger f \|_{\mathcal{X}}^2 = \sum_{j=1}^{\infty} (\sigma_j g_\alpha(\sigma_j) - 1)^2 \left| \langle u^\dagger, x_j \rangle_{\mathcal{X}} \right|^2 .$$

From (3.7) we can conclude

$$(\sigma_j g_\alpha(\sigma_j) - 1)^2 \leqslant (1 + \gamma^2) \,,$$

whilst

$$\sum_{j=1}^{\infty} (1 + \gamma^2) \left| \langle u^\dagger, x_j \rangle_{\mathcal{X}} \right|^2 = (1 + \gamma^2) \| u^\dagger \|^2 < +\infty.$$

Therefore, by the reverse Fatou lemma we get the following estimate

$$\begin{aligned}\limsup_{\alpha\to 0}\left\|R_\alpha f - A^\dagger f\right\|_{\mathcal{X}}^2 &= \limsup_{\alpha\to 0}\sum_{j=1}^{\infty}\left(\sigma_j g_\alpha(\sigma_j)-1\right)^2\left(\langle u^\dagger, x_j\rangle_{\mathcal{X}}\right)^2 \\ &\leqslant \sum_{j=1}^{\infty}\left(\limsup_{\alpha\to 0}\sigma_j g_\alpha(\sigma_j)-1\right)^2\left|\langle u^\dagger, x_j\rangle_{\mathcal{X}}\right|^2 = 0\,,\end{aligned}$$

where the last equality is due to the pointwise convergence of $g_\alpha(\sigma_j)$ to $1/\sigma_j$. Hence, we have $\left\|R_\alpha f - A^\dagger f\right\|_{\mathcal{X}} \to 0$ for $\alpha \to 0$ for all $f \in \mathcal{D}(A^\dagger)$. □

**Theorem 3.3.2.** *Let the assumptions of Theorem 3.3.1 hold and let $\alpha = \alpha(\delta)$ be an a-priori parameter choice rule. Then $(R_{\alpha(\delta)}, \alpha(\delta))$ with $R_\alpha$ as defined in (3.5) is a convergent regularisation method if*

$$\lim_{\delta\to 0}\delta C_{\alpha(\delta)} = 0.$$

*Proof.* The result follows immediately from $\|R_{\alpha(\delta)}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \leqslant C_{\alpha(\delta)}$ and Theorem 3.2.4. □

### 3.3.1 Truncated singular value decomposition

As a first example for a spectral regularisation of the form (3.5) we want to consider the so-called *truncated singular value decomposition*. The idea is to discard all singular values below a certain threshold $\alpha$, which is achieved using the following function $g_\alpha$

$$g_\alpha(\sigma) = \begin{cases}\frac{1}{\sigma} & \sigma \geqslant \alpha \\ 0 & \sigma < \alpha\end{cases}. \tag{3.8}$$

Note that for all $\sigma > 0$ we naturally obtain $\lim_{\alpha\to 0} g_\alpha(\sigma) = 1/\sigma$. Condition (3.7) is obviously satisfied with $\gamma = 1$ and condition (3.6) with $C_\alpha = \frac{1}{\alpha}$. Therefore, truncated SVD is a convergent regularisation if

$$\lim_{\delta\to 0}\frac{\delta}{\alpha} = 0. \tag{3.9}$$

Equation (3.5) then reads as follows

$$R_\alpha f = \sum_{\sigma_j \geqslant \alpha}\frac{1}{\sigma_j}\langle f, y_j\rangle_{\mathcal{Y}}\, x_j\,, \tag{3.10}$$

for all $f \in \mathcal{Y}$. Note that the sum in (3.10) is always well-defined (i.e. finite) for any $\alpha > 0$ as zero is the only accumulation point of singular vectors of compact operators.

Let $A \in \mathcal{K}(\mathcal{X},\mathcal{Y})$ with singular system $\{(\sigma_j, x_j, y_j)\}_{j\in\mathbb{N}}$, and choose for $\delta > 0$ an index function $j^* : \mathbb{R}_+ \to \mathbb{N}$ with $j^*(\delta) \to \infty$ for $\delta \to 0$ and $\lim_{\delta\to 0}\delta/\sigma_{j^*(\delta)} = 0$. We can then choose $\alpha(\delta) = \sigma_{j^*(\delta)}$ as an a-priori parameter choice rule to obtain a convergent regularisation.

Note that in practice a larger $\delta$ implies that more and more singular values have to be cut off in order to guarantee a stable recovery that successfully suppresses the data error.

A disadvantage of this approach is that it requires the knowledge of the singular vectors of $A$ (only finitely many, but the number can still be large).

### 3.3.2 Tikhonov regularisation

The main idea behind Tikhonov regularisation[1] is to consider the normal equations and shift the eigenvalues of $A^*A$ by a constant factor, which will be associated with the regularisation parameter $\alpha$. This shift can be realised via the function

$$g_\alpha(\sigma) = \frac{\sigma}{\sigma^2 + \alpha} \tag{3.11}$$

and the corresponding Tikhonov regularisation (3.5) reads as follows

$$R_\alpha f = \sum_{j=1}^{\infty} \frac{\sigma_j}{\sigma_j^2 + \alpha} \langle f, y_j \rangle_{\mathcal{Y}} \, x_j \,. \tag{3.12}$$

Again, we immediately observe that for all $\sigma > 0$ we have $\lim_{\alpha \to 0} g_\alpha(\sigma) = 1/\sigma$. Condition (3.7) is satisfied with $\gamma = 1$. Since $0 \leqslant (\sigma - \sqrt{\alpha})^2 = \sigma^2 - 2\sigma\sqrt{\alpha} + \alpha$, we get that $\sigma^2 + \alpha \geqslant 2\sigma\sqrt{\alpha}$ and

$$\frac{\sigma}{\sigma^2 + \alpha} \leqslant \frac{1}{2\sqrt{\alpha}}.$$

This estimate implies that (3.6) holds with $C_\alpha = \frac{1}{2\sqrt{\alpha}}$. Therefore, Tikhonov regularisation is a convergent regularisation if

$$\lim_{\delta \to 0} \frac{\delta}{\sqrt{\alpha}} = 0. \tag{3.13}$$

The formula (3.12) suggests that we need all singular vectors of $A$ in order to compute the regularisation. However, we note that $\sigma_j^2$ are the eigenvalues of $A^*A$ and, hence, $\sigma_j^2 + \alpha$ are the eigenvectors of $A^*A + \alpha I$ (where $I$ is the identity operator). Applying this operator to the regularised solution $u_\alpha = R_\alpha f$, we get

$$(A^*A + \alpha I)u_\alpha = \sum_{j=1}^{\infty} (\sigma_j^2 + \alpha) \langle u_\alpha, x_j \rangle_{\mathcal{X}} \, x_j = \sum_{j=1}^{\infty} (\sigma_j^2 + \alpha) \frac{\sigma_j}{\sigma_j^2 + \alpha} \langle f, y_j \rangle_{\mathcal{Y}} \, x_j = A^* f.$$

Therefore, the regularised solution $u_\alpha$ can be computed without knowing the singular system of $A$ by solving the following well-posed linear equation

$$(A^*A + \alpha I)u_\alpha = A^* f. \tag{3.14}$$

**Remark 3.3.3.** Rewriting equation (3.14) as

$$A^*(Au_\alpha - f) + \alpha u_\alpha = 0,$$

we note that it looks like a condition for the minimum of some quadratic form. Indeed, it can be easily checked that (3.14) is the first order optimality condition for the following optimisation problem

$$\min_{u \in \mathcal{X}} \frac{1}{2} \|Au - f\|^2 + \alpha \|u\|^2. \tag{3.15}$$

The condition (3.14) is necessary (and, by convexity, sufficient) for the minimum of the functional in (3.15). Therefore, the regularised solution $u_\alpha$ can also be computed by solving (numerically) the variational problem (3.15). This is the starting point for modern variational regularisation methods, which we will consider in the next chapter.

[1] Named after the Russian mathematician Andrey Nikolayevich Tikhonov (30 October 1906 - 7 October 1993)

# Chapter 4

# Variational Regularisation

Recall the variation formulation of Tikhonov regularisation for some data $f_\delta \in \mathcal{Y}$

$$\min_{u\in\mathcal{X}} \|Au - f_\delta\|^2 + \alpha\|u\|^2.$$

The first term in this expression, $\|Au - f_\delta\|^2$, penalises the misfit between the predictions of the operator $A$ and the measured data $f_\delta$ and is called the *fidelity function* or *fidelity term*. The second term, $\|u\|^2$ penalises some unwanted features of the solution (in this case, a large norm) and is called the *regularistaion term*. The regularisation parameter $\alpha$ in this context balances the influence of these two terms on the functional to be minimised.

More generally, using the notation $\mathcal{J}(u)$ for the regulariser, we can formally write down the variational regularisation problem as follows

$$\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|^2 + \alpha\mathcal{J}(u), \tag{4.1}$$

(the $\frac{1}{2}$ in front of the fidelity term is there to simplify notation later). The regularisation operator $R_\alpha$ is defined as follows

$$R_\alpha f_\delta \in \arg\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|^2 + \alpha\mathcal{J}(u).$$

In general, the minimiser doesn't have to unique, hence the inclusion and not equality. Other fidelity terms (not just $\|Au - f_\delta\|^2$) are possible and useful in many situations. In this course, however, we will use the squared norm for the sake of simplicity.

In this chapter, we will study the properties of (4.1) for different choices of $\mathcal{J}$, but before that we will recall some necessary theoretical concepts.

## 4.1 Background

### 4.1.1 Banach spaces and weak convergence

Banach spaces are complete, normed vector spaces (as Hilbert spaces) but they may not have an inner product. For every Banach space $\mathcal{X}$, we can define the space of linear and continuous functionals which is called the *dual space* $\mathcal{X}^*$ of $\mathcal{X}$, i.e. $\mathcal{X}^* := \mathcal{L}(\mathcal{X}, \mathbb{R})$. Let $u \in \mathcal{X}$ and $p \in \mathcal{X}^*$, then we usually write the *dual product* $\langle p, u\rangle$ instead of $p(u)$. Moreover,

for any $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ there exists a unique operator $A^*\colon \mathcal{Y}^* \to \mathcal{X}^*$, called the *adjoint* of $A$ such that for all $u \in \mathcal{X}$ and $p \in \mathcal{Y}^*$ we have

$$\langle A^* p, u\rangle = \langle p, Au\rangle \ .$$

It is easy to see that either side of the equation are well-defined, e.g. $A^*p \in \mathcal{X}^*$ and $u \in \mathcal{X}$.

The dual space of a Banach space $\mathcal{X}$ can be equipped with the following norm

$$\|p\|_{\mathcal{X}^*} = \sup_{u\in\mathcal{X}, \|u\|_{\mathcal{X}}\leqslant 1} \langle p, u\rangle \, .$$

With this norm the dual space is itself a Banach space. Therefore, it has a dual space as well which we will call the bi-dual space of $\mathcal{X}$ and denote it with $\mathcal{X}^{**} := (\mathcal{X}^*)^*$. As every $u \in \mathcal{X}$ defines a continuous and linear mapping on the dual space $\mathcal{X}^*$ by

$$\langle E(u), p\rangle := \langle p, u\rangle \ ,$$

the mapping $E\colon \mathcal{X} \to \mathcal{X}^{**}$ is well-defined. It can be shown that $E$ is a linear and continuous isometry (and thus injective). In the special case when $E$ is surjective, we call $\mathcal{X}$ *reflexive*. Examples of reflexive Banach spaces include Hilbert spaces and $L^q, \ell^q$ spaces with $1 < q < \infty$. We call the space $\mathcal{X}$ *separable* if there exists a set $\mathcal{X}' \subset \mathcal{X}$ of at most countable cardinality such that $\overline{\mathcal{X}'} = \mathcal{X}$.

A problem in infinite dimensional spaces is that bounded sequences may fail to have convergent subsequences. An example is for instance in $\ell^2$ the sequence $\{u^k\}_{k\in\mathbb{N}} \subset \ell^2, u^k_j = 1$ if $k = j$ and 0 otherwise. It is easy to see that $\|u^k\|_{\ell^2} = 1$ and that there is no $u \in \ell^2$ such that $u^k \to u$. To circumvent this problem, we define a weaker topology on $\mathcal{X}$. We say that $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ *converges weakly* to $u \in \mathcal{X}$ if and only if for all $p \in \mathcal{X}^*$ the sequence of real numbers $\{\langle p, u^k\rangle\}_{k\in\mathbb{N}}$ converges and

$$\langle p, u_j\rangle \to \langle p, u\rangle \ .$$

We will denote weak convergence by $u^k \rightharpoonup u$. On a dual space $\mathcal{X}^*$ we could define another topology (in addition to the strong topology induced by the norm and the weak topology as the dual space is a Banach space as well). We say a sequence $\{p^k\}_{k\in\mathbb{N}} \subset \mathcal{X}^*$ *converges in weak-** to $p \in \mathcal{X}^*$ if and only if

$$\left\langle p^k, u\right\rangle \to \langle p, u\rangle \quad \text{for all } u \in \mathcal{X}$$

and we denote weak-$*$ convergence by $p^k \overset{*}{\to} p$. Similarly, for any topology $\tau$ on $\mathcal{X}$ we denote the convergence in that topology by $u^k \overset{\tau}{\to} u$.

With these two new notions of convergence, we can solve the problem of bounded sequences:

**Theorem 4.1.1** (Sequential Banach-Alaoglu Theorem, e.g. [26, p. 70] or [30, p. 141])**.** *Let $\mathcal{X}$ be a separable normed vector space. Then every bounded sequence $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{X}^*$ has a weak-* convergent subsequence.*

**Theorem 4.1.2** ([32, p. 64])**.** *Each bounded sequence $\{u^k\}_{k\in\mathbb{N}}$ in a reflexive Banach space $\mathcal{X}$ has a weakly convergent subsequence.*

An important property of functionals, which we will need later, is sequential lower semicontinuity. Roughly speaking this means that the functional values for arguments near an argument $u$ are either close to $E(u)$ or greater than $E(u)$.

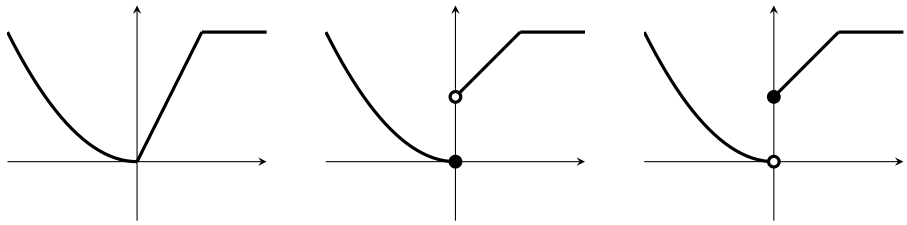


Figure 4.1: Visualisation of lower semi-continuity. The solid dot at a jump indicates the value that the function takes. The function on the left is continuous and thus lower semi-continuous. The functions in the middle and on the right are discontinuous. While the function in the middle is lower semi-continuous, the function on the right is not (due to the limit from the left at the discontinuity).

**Definition 4.1.3.** *Let $\mathcal{X}$ be a Banach space with topology $\tau_{\mathcal{X}}$. The functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ is said to be* sequentially lower semi-continuous with respect to $\tau_{\mathcal{X}}$ *($\tau_{\mathcal{X}}$-l.s.c.) at $u \in \mathcal{X}$ if*

$$E(u) \leqslant \liminf_{j\to\infty} E(u_j)$$

*for all sequences $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ with $u_j \to u$ in the topology $\tau_{\mathcal{X}}$ of $\mathcal{X}$.*

**Remark 4.1.4.** For topologies that are not induced by a metric we have to differ between a topological property and its sequential version, e.g. continuous and sequentially continuous. If the topology is induced by a metric, then these two are the same. However, for instance the weak and weak-$*$ topology are generally not induced by a metric.

**Example 4.1.5.** The functional $\|\cdot\|_1 : \ell^2 \to \bar{\mathbb{R}}$ with

$$\|u\|_1 = \begin{cases} \sum_{j=1}^{\infty} |u_j| & \text{if } u \in \ell^1 \\ \infty & \text{else} \end{cases}$$

is weakly (and, hence, strongly) lower semi-continuous in $\ell^2$.

*Proof.* Let $\{u^j\}_{j\in\mathbb{N}} \subset \ell^2$ be a weakly convergent sequence with $u^j \rightharpoonup u \in \ell^2$. We have with $\delta_k : \ell^2 \to \mathbb{R}, \langle \delta_k, v\rangle = v_k$ that for all $k \in \mathbb{N}$

$$u_k^j = \langle \delta_k, u^j\rangle \to \langle \delta_k, u\rangle = u_k\,.$$

The assertion follows then with Fatou's lemma

$$\|u\|_1 = \sum_{k=1}^{\infty} |u_k| = \sum_{k=1}^{\infty} \lim_{j\to\infty} |u_k^j| \leqslant \liminf_{j\to\infty} \sum_{k=1}^{\infty} |u_k^j| = \liminf_{j\to\infty} \|u^j\|_1\,.$$

Note that it is not clear whether both the left and the right hand side are finite. □

### 4.1.2 Convex analysis

#### Infinity calculus

We will look at functionals $E : \mathcal{X} \to \bar{\mathbb{R}}$ whose range is modelled to be the *extended real line* $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ where the symbol $+\infty$ denotes an element that is not part of the real line that is by definition larger than any other element of the reals, i.e.

$$x < +\infty$$

for all $x \in \mathbb{R}$ (similarly, $x > -\infty$ for all $x \in \mathbb{R}$). This is useful to model constraints: for instance, if we were trying to minimise $E : [-1, \infty) \to \mathbb{R}, x \mapsto x^2$ we could remodel this minimisation problem by $\widetilde{E} : \mathbb{R} \to \bar{\mathbb{R}}$

$$\widetilde{E}(x) = \begin{cases} x^2 & \text{if } x \geqslant -1 \\ \infty & \text{else} \end{cases}.$$

Obviously both functionals have the same minimiser but $\widetilde{E}$ is defined on a vector space and not only on a subset. This has two important consequences: on the on hand, it makes many theoretical arguments easier as we do not need to worry whether $E(x+y)$ is defined or not. On the other hand, it makes practical implementations easier as we are dealing with unconstrained optimisation instead of constrained optimisation. This comes at a cost that some algorithms are not applicable any more, e.g. the function $\widetilde{E}$ is not differentiable everywhere whereas $E$ is (in the interior of its domain).

It is useful to note that one can calculate on the extended real line $\bar{\mathbb{R}}$ as we are used to on the real line $\mathbb{R}$ but the operations with $\pm\infty$ need yet to be defined.

**Definition 4.1.6.** *The extended real line is defined as* $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ *with the following rules that hold for any* $x \in \mathbb{R}$ *and* $\lambda > 0$*:*

$$\begin{aligned} x + \infty := \infty + x := \infty \qquad & \lambda \cdot \infty := \infty \cdot \lambda := \infty \\ x/\infty := 0 \qquad & \infty + \infty := \infty\,. \end{aligned}$$

Some calculations are *not defined*, e.g.,

$$\infty - \infty \text{ and } \quad \infty \cdot \infty\,.$$

Using functions with values on the extended real line, one can easily describe sets $\mathcal{C} \subset \mathcal{X}$.

**Definition 4.1.7** (Characteristic function)**.** *Let* $\mathcal{C} \subset \mathcal{X}$ *be a set. The function* $\chi_{\mathcal{C}} \colon \mathcal{X} \to \bar{\mathbb{R}}$*,*

$$\chi_{\mathcal{C}}(u) = \begin{cases} 0 & u \in \mathcal{C} \\ \infty & u \in \mathcal{X} \setminus \mathcal{C} \end{cases}$$

*is called the characteristic function of the set* $\mathcal{C}$*.*

Using characteristic functions, one can easily write constrained optimisation problems as unconstrained ones:

$$\min_{u \in \mathcal{C}} E(u) \quad \Leftrightarrow \quad \min_{u \in \mathcal{X}} E(u) + \chi_{\mathcal{C}}(u).$$

**Definition 4.1.8.** *Let* $\mathcal{X}$ *be a vector space and* $E \colon \mathcal{X} \to \bar{\mathbb{R}}$ *a functional. Then the* effective domain *of* $E$ *is*

$$\operatorname{dom}(E) := \{u \in \mathcal{X} \mid E(u) < \infty\}\,.$$

**Definition 4.1.9.** *A functional* $E$ *is called* proper *if the effective domain* $\operatorname{dom}(E)$ *is not empty.*

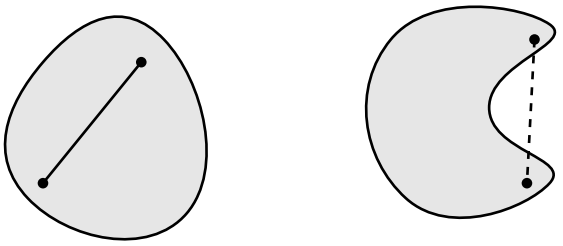


Figure 4.2: Example of a convex set (left) and non-convex set (right).

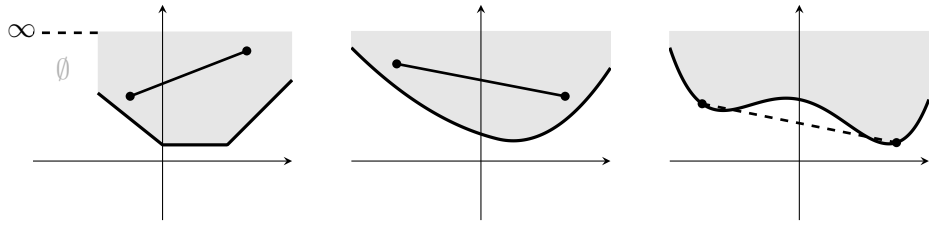



Figure 4.3: Example of a convex function (left), a strictly convex function (middle) and a non-convex function (right).

### Convexity

A property of fundamental importance of sets and functions is convexity.

**Definition 4.1.10.** *Let $\mathcal{X}$ be a vector space. A subset $\mathcal{C} \subset \mathcal{X}$ is called* convex*, if $\lambda u + (1-\lambda)v \in \mathcal{C}$ for all $\lambda \in (0,1)$ and all $u, v \in \mathcal{C}$.*

**Definition 4.1.11.** *A functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ is called* convex*, if*

$$E(\lambda u + (1-\lambda)v) \leqslant \lambda E(u) + (1-\lambda)E(v)$$

*for all $\lambda \in (0,1)$ and all $u, v \in \operatorname{dom}(E)$ with $u \neq v$. It is called* strictly convex *if the inequality is strict. It is called* strongly *convex with constant $\theta$ if $E(u) - \theta\|u\|^2$ is convex.*

Obviously, strong convexity implies strict convexity and strict convexity implies convexity.

**Example 4.1.12.** The absolute value function $\mathbb{R} \to \mathbb{R}, x \mapsto |x|$ is convex but not strictly convex. The quadratic function $x \mapsto x^2$ is strongly (and hence strictly) convex. The function $x \mapsto x^4$ is strictly convex, but not strongly convex. For other examples, see Figure 4.3.

**Example 4.1.13.** The characteristic function $\chi_{\mathcal{C}}(u)$ is convex if and only if $\mathcal{C}$ is a convex set. To see the convexity, let $u, v \in \operatorname{dom}(\chi_{\mathcal{C}}) = \mathcal{C}$. Then by the convexity of $\mathcal{C}$ the convex combination $\lambda u + (1-\lambda)v$ is as well in $\mathcal{C}$ and both the left and the right hand side of the desired inequality are zero.

**Lemma 4.1.14.** *Let $\alpha \geqslant 0$ and $E, F\colon \mathcal{X} \to \bar{\mathbb{R}}$ be two convex functionals. Then $E + \alpha F\colon \mathcal{X} \to \bar{\mathbb{R}}$ is convex. Furthermore, if $\alpha > 0$ and $F$ strictly convex, then $E + \alpha F$ is strictly convex.*

### Fenchel conjugate

In convex optimisation problems (i.e. those involving convex functions) the concept of *Fenchel conjugates* plays a very important role.

**Definition 4.1.15.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a functional. The functional $E^*\colon \mathcal{X}^* \to \bar{\mathbb{R}}$,*

$$E^*(p) = \sup_{u \in \mathcal{X}} [\langle u, p\rangle - E(u)],$$

*is called the Fenchel conjugate of $E$.*

**Theorem 4.1.16** ([15, Prop. 4.1])**.** *For any functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ the following inequality holds:*

$$E^{**} := (E^*)^* \leqslant E.$$

*If $E$ is proper, lower-semicontinuous (see Def. 4.1.3) and convex, then*

$$E^{**} = E.$$

### Subgradients

For convex functions one can generalise the concept of a derivative so that it would also make sense for non-differentiable functions.

**Definition 4.1.17.** *A functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ is called* subdifferentiable *at $u \in \mathcal{X}$, if there exists an element $p \in \mathcal{X}^*$ such that*

$$E(v) \geqslant E(u) + \langle p, v - u\rangle$$

*holds, for all $v \in \mathcal{X}$. Furthermore, we call $p$ a* subgradient *at position $u$. The collection of all subgradients at position $u$, i.e.*

$$\partial E(u) := \{p \in \mathcal{X}^* \mid E(v) \geqslant E(u) + \langle p, v - u\rangle \ , \forall v \in \mathcal{X}\} \ ,$$

*is called* subdifferential *of $E$ at $u$.*

**Remark 4.1.18.** Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a convex functional. Then the subdifferential is non-empty at all $u \in \mathrm{dom}(E)$. If $\mathrm{dom}(E) \neq \emptyset$, then for all $u \notin \mathrm{dom}(E)$ the subdifferential is empty, i.e. $\partial E(u) = \emptyset$.

**Theorem 4.1.19** ([4, Thm. 7.13])**.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a proper convex function and $u \in \mathrm{dom}(E)$. Then $\partial E(u)$ is a weak-$*$ compact convex subset of $\mathcal{X}^*$.*

For differentiable functions the subdifferential consists of just one element – the derivative. For non-differentiable functionals the subdifferential is multivalued; we want to consider the subdifferential of the absolute value function as an illustrative example.

**Example 4.1.20.** Let $E\colon \mathbb{R} \to \mathbb{R}$ be the absolute value function $E(u) = |u|$. Then, the subdifferential of $E$ at $u$ is given by

$$\partial E(u) = \begin{cases} \{1\} & \text{for } u > 0 \\ [-1, 1] & \text{for } u = 0 \\ \{-1\} & \text{for } u < 0 \end{cases},$$

which you will prove as an exercise. A visual explanation is given in Figure 4.4.

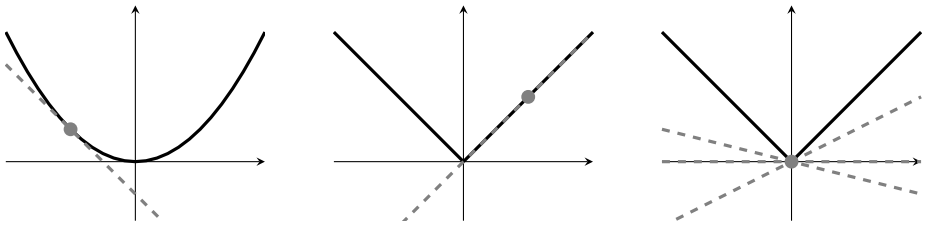


Figure 4.4: Visualisation of the subdifferential. Linear approximations of the functional have to lie completely underneath the function. For points where the function is not differentiable there may be more than one such approximation.

The subdifferential of a sum of two functions can be characterised as follows.

**Theorem 4.1.21** ([15, Prop. 5.6])**.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ and $F\colon \mathcal{X} \to \bar{\mathbb{R}}$ be proper l.s.c. convex functions and suppose $\exists u \in \operatorname{dom}(E) \cup \operatorname{dom}(F)$ such that $E$ is continuous at $u$. Then*

$$\partial(E + F) = \partial E + \partial F.$$

Using the subdifferential, one can characterise minimisers of convex functionals.

**Theorem 4.1.22.** *An element $u \in \mathcal{X}$ is a minimiser of the functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ if and only if $0 \in \partial E(u)$.*

*Proof.* By definition, $0 \in \partial E(u)$ if and only if for all $v \in \mathcal{X}$ it holds

$$E(v) \geqslant E(u) + \langle 0, v - u\rangle = E(u)\,,$$

which is by definition the case if and only if $u$ is a minimiser of $E$. □

### Bregman distances

Convex functions naturally define some distance measure that became known as the Bregman distance.

**Definition 4.1.23.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a convex functional. Moreover, let $u, v \in \mathcal{X}, E(v) < \infty$ and $q \in \partial E(v)$. Then the* (generalised) Bregman distance *of $E$ between $u$ and $v$ is defined as*

$$D_E^q(u, v) := E(u) - E(v) - \langle q, u - v\rangle\ . \tag{4.2}$$

**Remark 4.1.24.** It is easy to check that a Bregman distance somewhat resembles a metric as for all $u, v \in \mathcal{X}, q \in \partial E(v)$ we have that $D_E^q(u, v) \geqslant 0$ and $D_E^q(v, v) = 0$. There are functionals where the Bregman distance (up to a square root) is actually a metric; e.g. $E(u) := \frac{1}{2}\|u\|_{\mathcal{X}}^2$ for Hilbert space $\mathcal{X}$, then $D_E^q(u, v) = \frac{1}{2}\|u - v\|_{\mathcal{X}}^2$. However, in general, Bregman distances are not symmetric and $D_E^q(u, v) = 0$ does not imply $u = v$, as you will see on the example sheets.

To overcome the issue of non-symmetry, one can introduce the so-called *symmetric Bregman distance*.

**Definition 4.1.25.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a convex functional. Moreover, let $u, v \in \mathcal{X}, E(u) < \infty, E(v) < \infty$, $q \in \partial E(v)$ and $p \in \partial E(u)$. Then the* symmetric Bregman distance *of $E$ between $u$ and $v$ is defined as*

$$D_E^{symm}(u, v) := D_E^q(u, v) + D_E^p(v, u) = \langle p - q, u - v\rangle\ . \tag{4.3}$$

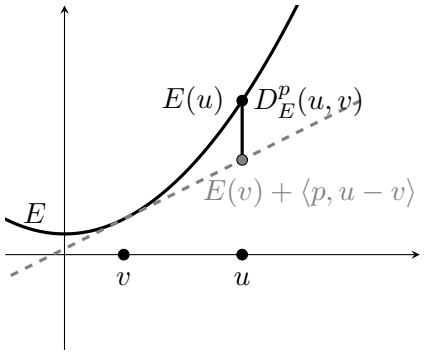



Figure 4.5: Visualization of the Bregman distance.

**Absolutely one-homogeneous functionals**

**Definition 4.1.26.** *A functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ is called absolutely one-homogeneous if*

$$E(\lambda u) = |\lambda| E(u) \quad \forall \lambda \in \mathbb{R},\ \forall u \in \mathcal{X}.$$

Absolutely one-homogeneous convex functionals have some useful properties, for example, it is obvious that $E(0) = 0$. Some further properties are listed below.

**Proposition 4.1.27.** *Let $E(\cdot)$ be a convex absolutely one-homogeneous functional and let $p \in \partial E(u)$. Then the following equality holds:*

$$E(u) = \langle p, u\rangle.$$

*Proof.* Left as exercise. □

**Remark 4.1.28.** The Bregman distance $D_E^p(v, u)$ in this case can be written as follows:

$$D_E^p(v, u) = E(v) - \langle p, v\rangle.$$

**Proposition 4.1.29.** *Let $E(\cdot)$ be a proper, convex, l.s.c. and absolutely one-homogeneous functional. Then the Fenchel conjugate $E^*(\cdot)$ is the characteristic function of the convex set $\partial E(0)$.*

*Proof.* Left as exercise. □

An obvious consequence of the above results is the following

**Proposition 4.1.30.** *For any $u \in \mathcal{X}$, $p \in \partial E(u)$ if and only if $p \in \partial E(0)$ and $E(u) = (p, u)$.*

### 4.1.3 Minimisers

**Definition 4.1.31.** *Let $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a functional. We say that $u^* \in \mathcal{X}$ solves the minimisation problem*

$$\min_{u \in \mathcal{X}} E(u)$$

*if and only if $E(u^*) < \infty$ and $E(u^*) \leqslant E(u)$, for all $u \in \mathcal{X}$. We call $u^*$ a* minimiser *of $E$.*

**Definition 4.1.32.** *A functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ is called* bounded from below *if there exists a constant $C > -\infty$ such that for all $u \in \mathcal{X}$ we have $E(u) \geqslant C$.*

This condition is obviously necessary for the finiteness of the infimum $\inf_{u \in \mathcal{X}} E(u)$.

**Existence**

If all minimising sequences (that converge to the infimum assuming it exists) are unbounded, then there cannot exist a minimiser. A sufficient condition to avoid such a scenario is *coercivity*.

**Definition 4.1.33.** *A functional* $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ *is called* coercive, *if for all* $\{u_j\}_{j\in\mathbb{N}}$ *with* $\|u_j\|_{\mathcal{X}} \to \infty$ *we have* $E(u_j) \to \infty$.

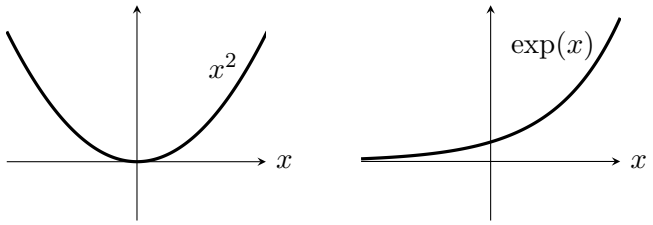



Figure 4.6: While the coercive function on the left has a minimiser, it is easy to see that the non-coercive function on the right does not have a minimiser.

**Remark 4.1.34.** Coercivity is equivalent to its negated statement which is "if the function values $\{E(u_j)\}_{j\in\mathbb{N}} \subset \mathbb{R}$ are bounded, so is the sequence $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$".

Although coercivity is not strictly speaking necessary, it is sufficient that all minimising sequences are bounded.

**Lemma 4.1.35.** *Let* $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ *be a proper, coercive functional and bounded from below. Then the infimum* $\inf_{u\in\mathcal{X}} E(u)$ *exists in* $\mathbb{R}$*, there are minimising sequences, i.e.* $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ *with* $E(u_j) \to \inf_{u\in\mathcal{X}} E(u)$*, and all minimising sequences are bounded.*

*Proof.* As $E$ is proper and bounded from below, there exists a $C_1 > 0$ such that we have $-\infty < -C_1 < \inf_u E(u) < \infty$ which also guarantees the existence of a minimising sequence. Let $\{u_j\}_{j\in\mathbb{N}}$ be any minimising sequence, i.e. $E(u_j) \to \inf_u E(u)$. Then there exists a $j_0 \in \mathbb{N}$ such that for all $j > j_0$ we have

$$E(u_j) \leqslant \underbrace{\inf_u E(u) + 1}_{=:C_2} < \infty\,.$$

With $C := \max\{C_1, C_2\}$ we have that $|E(u_j)| < C$ for all $j > j_0$ and thus from the coercivity it follows that $\{u_j\}_{j>j_0}$ is bounded, see Remark 4.1.34. Including a finite number of elements does not change its boundedness which proves the assertion. □

A positive answer about the existence of minimisers is given by the following Theorem known as the "direct method" or "fundamental theorem of optimisation".

**Theorem 4.1.36** ("Direct method", David Hilbert, around 1900)**.** *Let* $\mathcal{X}$ *be a Banach space and* $\tau_{\mathcal{X}}$ *a topology (not necessarily the one induced by the norm) on* $\mathcal{X}$ *such that bounded sequences have* $\tau_{\mathcal{X}}$*-convergent subsequences. Let* $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ *be proper, bounded from below, coercive and* $\tau_{\mathcal{X}}$*-l.s.c. Then* $E$ *has a minimiser.*

*Proof.* From Lemma 4.1.35 we know that $\inf_{u\in\mathcal{X}} E(u)$ is finite, minimising sequences exist and that they are bounded. Let $\{u_j\}_{j\in\mathbb{N}} \in \mathcal{X}$ be a minimising sequence. Thus, from the assumption on the topology $\tau_{\mathcal{X}}$ there exists a subsequence $\{u_{j_k}\}_{k\in\mathbb{N}}$ and $u^* \in \mathcal{X}$ with $u_{j_k} \overset{\tau_{\mathcal{X}}}{\to} u^*$ for $k \to \infty$. From the sequential lower semi-continuity of $E$ we obtain

$$E(u^*) \leqslant \liminf_{k\to\infty} E(u_{j_k}) = \lim_{j\to\infty} E(u_j) = \inf_{u\in\mathcal{X}} E(u) < \infty\,,$$

which shows that $E(u^*) < \infty$ and $E(u^*) \leqslant E(u)$ for all $u \in \mathcal{X}$; thus $u^*$ minimises $E$. □

The above theorem is very general but its conditions are hard to verify but the situation is a easier in *reflexive* Banach spaces (thus also in Hilbert spaces).

**Corollary 4.1.37.** Let $\mathcal{X}$ be a reflexive Banach space and $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ be a functional which is proper, bounded from below, coercive and l.s.c. with respect to the weak topology. Then there exists a minimiser of $E$.

*Proof.* The statement follows from the direct method, Theorem 4.1.36, as in reflexive Banach spaces bounded sequences have weakly convergent subsequences, see Theorem 4.1.2. □

**Remark 4.1.38.** For convex functionals on reflexive Banach spaces, the situation is even easier. It can be shown that a convex function is l.s.c. with respect to the weak topology if and only if it is l.s.c. with respect to the strong topology (see e.g. [15, Corollary 2.2., p. 11] or [7, p. 149] for Hilbert spaces).

**Remark 4.1.39.** It is easy to see that the key ingredient for the existence of minimisers is that bounded sequences have a convergent subsequence. In variational regularisation this is usually ensured by an appropriate choice of the regularisation functional.

### Uniqueness

**Theorem 4.1.40.** *Assume that the functional $E\colon \mathcal{X} \to \bar{\mathbb{R}}$ has at least one minimiser and is strictly convex. Then the minimiser is unique.*

*Proof.* Let $u, v$ be two minimisers of $E$ and assume that they are different, i.e. $u \neq v$. Then it follows from the minimising properties of $u$ and $v$ as well as the strict convexity of $E$ that

$$E(u) \leqslant E(\tfrac{1}{2}u + \tfrac{1}{2}v) < \frac{1}{2}E(u) + \frac{1}{2}\underbrace{E(v)}_{\leqslant E(u)} \leqslant E(u)$$

which is a contradiction. Thus, $u = v$ and the assertion is proven. □

**Example 4.1.41.** Convex (but not strictly convex) functions may have have more than one minimiser, examples include constant and trapezoidal functions, see Figure 4.7. On the other hand, convex (and even non-convex) functions may have a unique minimiser, see Figure 4.7.

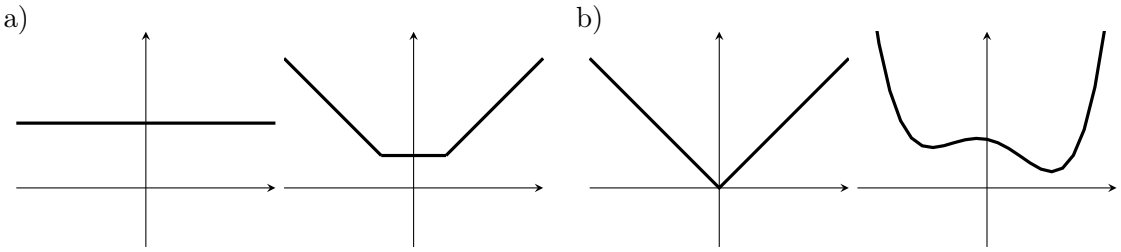



Figure 4.7: a) Convex functions may not have a unique minimiser. b) Neither strict convexity nor convexity is necessary for the uniqueness of a minimiser.

## 4.2 Well-posedness and Regularisation Properties

Our goal is to study the properties of optimisation problem (4.1) as a convergent regularisation for the ill-posed problem

$$Au = f, \tag{4.4}$$

where $A\colon \mathcal{X} \to \mathcal{Y}$ is a linear bounded operator and $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces (and not Hilbert spaces as in Chapter 3). In particular, we will ask questions of existence of minimisers (well-posedness of the regularised problem) and parameter choice rules that guarantee the convergence of the minimisers to an appropriate generalised solution of (4.4) for different choices of the regularisation functional. To this end, we need to extend the definition of a minimal-norm solution (Def. 2.1.1) to an arbitrary regularisation term.

**Definition 4.2.1** ($\mathcal{J}$-minimising solutions)**.** *Let $u^\dagger_{\mathcal{J}}$ be a least squares solution, i.e.*

$$\|Au^\dagger_{\mathcal{J}} - f\|_{\mathcal{Y}} = \inf\{\|Av - f\|_{\mathcal{Y}}, \quad v \in \mathcal{X}\}$$

*and*

$$\mathcal{J}(u^\dagger_{\mathcal{J}}) \leqslant \mathcal{J}(\widetilde{u}) \quad \textit{for all least squares solutions } \widetilde{u}.$$

*Then $u^\dagger_{\mathcal{J}}$ is called a $\mathcal{J}$-minimising solution of* (4.4)*.*

We will assume that equation (4.4) has a solution with a finite value of $\mathcal{J}$, i.e. there exists at least one element $u$ such that $Au = f$ and $\mathcal{J}(u) < +\infty$. Under this assumption, least squares solutions are actually solutions of (4.4).

**Remark 4.2.2.** A $\mathcal{J}$-minimising solution may not exist and if it does, it may be non-unique. We will later see conditions, under which a $\mathcal{J}$-minimising solution exists. Non-uniqueness, however, is common with popular choices of $\mathcal{J}$. In this case we need to define a *selection operator* that will select a single element from all the $\mathcal{J}$-minimising solutions (see [8]). We will not explicitly mention this, stating all results for just *a* $\mathcal{J}$-minimising solution.

First of all we will establish the existence of a $\mathcal{J}$-minimising solution and a regularised solution for a finite $\alpha$.

**Theorem 4.2.3.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{Y}}$ some topologies (not necessarily induced by the norm) in $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that $\|\cdot\|_{\mathcal{Y}}$ is $\tau_{\mathcal{Y}}$-lower semicontinuous. Suppose that problem* (4.4) *has a solution with a finite value of $\mathcal{J}$. Assume also that*

*(i)* $A\colon \mathcal{X} \to \mathcal{Y}$ *is* $\tau_{\mathcal{X}} \to \tau_{\mathcal{Y}}$ *continuous;*

*(ii)* $\mathcal{J}\colon \mathcal{X} \to \bar{\mathbb{R}}_+$ *is proper,* $\tau_{\mathcal{X}}$*-l.s.c. and its non-empty sublevel-sets* $\{u \in \mathcal{X}\colon \mathcal{J}(u) \leqslant C\}$ *are* $\tau_{\mathcal{X}}$*-sequentially compact;*

*Then*

*(i')* *there exists a* $\mathcal{J}$*-minimising solution* $u^\dagger_{\mathcal{J}}$ *of* (4.4)*;*

*(ii')* *for any fixed* $\alpha > 0$ *and* $f_\delta \in \mathcal{Y}$ *there exists a minimiser*

$$u^\alpha_\delta \in \operatorname*{arg\,min}_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u).$$

*Proof.* (i) Let $f \in \mathcal{R}(A)$ be the exact data. Let us denote the set of all solutions of (4.4) by $\mathbb{L}$. It is non-empty (since we assumed the existence of at least one solution) and $\tau_{\mathcal{X}}$-closed. To see this, consider a sequence $\{u_n\} \subset \mathbb{L}$ such that $u_n \overset{\tau_{\mathcal{X}}}{\to} \bar{u}$. Since $A$ is $\tau_{\mathcal{X}} \to \tau_{\mathcal{Y}}$ continuous, $Au_n \overset{\tau_{\mathcal{Y}}}{\to} A\bar{u}$. On the other hand, since $\{u_n\} \subset \mathbb{L}$, $Au_n = f$ for any $n$, hence $A\bar{u} = f$ and $\bar{u} \in \mathbb{L}$.

A $\mathcal{J}$-minimsing solution solves the following problem

$$\min_{u\in\mathbb{L}} \mathcal{J}(u).$$

Since $\mathcal{J}$ is bounded from below (by zero) and proper on $\mathbb{L}$ by assumption, the infimum in this problem is finite and we denote it by $\mathcal{J}_{min}$. Consider any minimising sequence $\{u_k\}$. By Assumption (ii), the sublevel-sets of $\mathcal{J}$ are $\tau_{\mathcal{X}}$-sequentially compact and $u_k$ contains a $\tau_{\mathcal{X}}$-converging subsequence $u_{k_j} \overset{\tau_{\mathcal{X}}}{\to} \widetilde{u}$ as $j \to \infty$. Since $\mathbb{L}$ is $\tau_{\mathcal{X}}$-closed, $\widetilde{u} \in \mathbb{L}$. Since $\mathcal{J}$ is $\tau_{\mathcal{X}}$-l.s.c., we get that

$$\mathcal{J}(\widetilde{u}) \leqslant \liminf_{j\to\infty} \mathcal{J}(u_{k_j}) = \mathcal{J}_{min}.$$

Therefore, $\mathcal{J}(\widetilde{u}) = \mathcal{J}_{min}$ and $\widetilde{u}$ is a $\mathcal{J}$-minimising solution, which we from now on denote by $u^\dagger_{\mathcal{J}}$.

(ii) Let $f_\delta \in \mathcal{Y}$ be noisy data s.t. $\|f - f_\delta\|_{\mathcal{Y}} \leqslant \delta$. For fixed $\alpha > 0$ and $\delta > 0$ consider the following optimisation problem

$$\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u). \tag{4.5}$$

Comparing the value of the objective function at minimising sequence $\{u_n\}$ and the $\mathcal{J}$-minimising solution $u^\dagger_{\mathcal{J}}$, we get that

$$\frac{1}{2}\|Au_n - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u_n) \leqslant \frac{1}{2}\|Au^\dagger_{\mathcal{J}} - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u^\dagger_{\mathcal{J}})$$

and

$$\mathcal{J}(u_n) \leqslant \frac{1}{2\alpha}\|Au^\dagger_{\mathcal{J}} - f_\delta\|^2_{\mathcal{Y}} + \mathcal{J}(u^\dagger_{\mathcal{J}}) \leqslant \frac{\delta^2}{2\alpha} + \mathcal{J}(u^\dagger_{\mathcal{J}}) = \text{const}.$$

By the sequential $\tau_{\mathcal{X}}$-compactness of the sublevel sets of $\mathcal{J}$ we get that $\{u_n\}$ contains a $\tau_{\mathcal{X}}$-converging subsequence $u_{n_j} \overset{\tau_{\mathcal{X}}}{\to} \widehat{u}$. Since $A$ is $\tau_{\mathcal{X}} \to \tau_{\mathcal{Y}}$ continuous, we get that $Au_{n_j} \overset{\tau_{\mathcal{Y}}}{\to} A\widehat{u}$. Since $\|\cdot\|_{\mathcal{Y}}$ is $\tau_{\mathcal{Y}}$-l.s.c and $\mathcal{J}(\cdot)$ is $\tau_{\mathcal{X}}$-l.s.c., we get that

$$\frac{1}{2}\|A\widehat{u} - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(\widehat{u}) \leqslant \liminf_{j\to\infty} \frac{1}{2}\|Au_{n_j} - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u_{n_j}) = \inf_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|^2_{\mathcal{Y}} + \alpha\mathcal{J}(u).$$

Therefore, $\widehat{u}$ is a minimiser in (4.5), which we will denote by $u^\alpha_\delta$. □

Let us study the behaviour of $u_\delta^\alpha$ when $\delta \to 0$ and $\alpha$ is chosen according to an appropriate a priori parameter choice rule.

**Theorem 4.2.4.** *Let the assumptions of Theorem 4.2.3 hold. If $\alpha = \alpha(\delta)$ is chosen s.t. $\frac{\delta^2}{\alpha(\delta)} \to 0$ and $\alpha(\delta) \to 0$ as $\delta \to 0$ then $u_\delta := u_\delta^{\alpha(\delta)} \overset{\tau_\mathcal{X}}{\to} u_\mathcal{J}^\dagger$ as $\delta \to 0$ (possibly, along a subsequence) and $\mathcal{J}(u_\delta) \to \mathcal{J}(u_\mathcal{J}^\dagger)$, where $u_\mathcal{J}^\dagger$ is a $\mathcal{J}$-minimising solution.*

*Proof.* Since $u_\delta$ solves (4.5) with $\alpha = \alpha(\delta)$, we get that

$$\frac{1}{2}\|Au_\delta - f_\delta\|_\mathcal{Y}^2 + \alpha(\delta)\mathcal{J}(u_\delta) \leqslant \frac{1}{2}\|Au_\mathcal{J}^\dagger - f_\delta\|_\mathcal{Y}^2 + \alpha(\delta)\mathcal{J}(u_\mathcal{J}^\dagger) \tag{4.6}$$

and

$$\mathcal{J}(u_\delta) \leqslant \frac{1}{2\alpha(\delta)}\|Au_\mathcal{J}^\dagger - f_\delta\|_\mathcal{Y}^2 + \mathcal{J}(u_\mathcal{J}^\dagger) \leqslant \frac{\delta^2}{2\alpha(\delta)} + \mathcal{J}(u_\mathcal{J}^\dagger) \tag{4.7}$$

The right-hand side is bounded uniformly in $\delta$ since $\lim_{\delta\to 0} \delta^2/\alpha(\delta) = 0$ by assumption and $\mathcal{J}(u_\mathcal{J}^\dagger)$ is a constant independent of $\delta$.

Choosing an arbitrary null sequence $\delta_n \downarrow 0$ and again using the $\tau_\mathcal{X}$-compactness of the sublevel sets of $\mathcal{J}$, we conclude that the sequence $u_{\delta_n}$ contains a $\tau_\mathcal{X}$-convergent subsequence (that we do not relabel to avoid triple subscripts) $u_{\delta_n} \overset{\tau_\mathcal{X}}{\to} u_0$ and $Au_{\delta_n} \overset{\tau_\mathcal{Y}}{\to} Au_0$ due to the $\tau_\mathcal{X} \to \tau_\mathcal{Y}$ continuity of $A$.

Due to the $\tau_\mathcal{Y}$-lower semicontinuity of the norm in $\mathcal{Y}$ we further obtain the following estimate

$$\begin{aligned}\frac{1}{2}\|Au_0 - f\|_\mathcal{Y}^2 &\leqslant \liminf_{n\to\infty} \frac{1}{2}\|Au_{\delta_n} - f_{\delta_n}\|_\mathcal{Y}^2 \\ &\leqslant \liminf_{n\to\infty} \frac{1}{2}\|Au_{\delta_n} - f_{\delta_n}\|_\mathcal{Y}^2 + \alpha(\delta_n)\mathcal{J}(u_{\delta_n}).\end{aligned}$$

Since $u_{\delta_n}$ is a minimiser in (4.5) with $\alpha = \alpha(\delta)$, we obtain

$$\begin{aligned}\frac{1}{2}\|Au_0 - f\|_\mathcal{Y}^2 &\leqslant \liminf_{n\to\infty} \frac{1}{2}\|Au_\mathcal{J}^\dagger - f_{\delta_n}\|_\mathcal{Y}^2 + \alpha(\delta_n)\mathcal{J}(u_\mathcal{J}^\dagger) \\ &\leqslant \liminf_{n\to\infty} \frac{\delta_n^2}{2} + \alpha(\delta_n)\mathcal{J}(u_\mathcal{J}^\dagger) = 0.\end{aligned}$$

Hence, $u_0$ is a solution (4.4).

Now it is left to show that $u_0$ has minimal value of $\mathcal{J}$ among all solutions (4.4). Using the estimate (4.7) and $\tau_\mathcal{X}$-lower semicontinuity of $\mathcal{J}$, we obtain

$$\mathcal{J}(u_0) \leqslant \liminf_{n\to\infty} \mathcal{J}(u_{\delta_n}) \leqslant \limsup_{n\to\infty} \mathcal{J}(u_{\delta_n}) \leqslant \limsup_{n\to\infty} \frac{\delta^2}{2\alpha(\delta)} + \mathcal{J}(u_\mathcal{J}^\dagger) = \mathcal{J}(u_\mathcal{J}^\dagger).$$

Since $u_\mathcal{J}^\dagger$ has by definition the smallest value of $\mathcal{J}$ among all solutions of (4.4), we also get that $\mathcal{J}(u_\mathcal{J}^\dagger) \leqslant \mathcal{J}(u_0)$. Therefore, there exists $\lim_{n\to\infty} \mathcal{J}(u_{\delta_n}) = \mathcal{J}(u_0) = \mathcal{J}(u_\mathcal{J}^\dagger)$ and $u_0$ is a $\mathcal{J}$-minimising solution of (4.4) (which is possibly different from $u_\mathcal{J}^\dagger$).

□

**Remark 4.2.5.** The compactness of the level sets of $\mathcal{J}(u)$ in Assumption (ii) can be replaced by compactness of the level sets of $\Phi_f^\alpha(u) := \frac{1}{2}\|Au - f\|_\mathcal{Y}^2 + \alpha\mathcal{J}(u)$.

**Remark 4.2.6.** The theorem proves convergence of the regularised solutions in $\tau_{\mathcal{X}}$, which may differ from the strong topology. However, if $\mathcal{J}$ satisfies the *Radon-Riesz property* with respect to the topology $\tau_{\mathcal{X}}$, i.e. $u_j \xrightarrow{\tau_{\mathcal{X}}} u$ and $\mathcal{J}(u_j) \to \mathcal{J}(u)$ imply $\|u_j - u\| \to 0$, then we get convergence in the norm topology. An example of a functional satisfying the Radon-Riesz property is the norm in a Hilbert (or reflexive Banach) space with $\tau_{\mathcal{X}}$ being the weak topology.

### Examples of regularisers

**Example 4.2.7.** Let $\mathcal{X}$ be a Hilbert space and $\mathcal{J}(u) = \|u\|^2$. The norm in a Hilbert space is weakly l.s.c. By Theorem 4.1.2 we know that (norm) bounded sequences have weakly convergent subsequences. Therefore, Assumption (ii) of Theorem 4.2.3 is satisfied with $\tau_{\mathcal{X}}$ being the weak topology and we obtain weak convergence of the regularised solutions. However, since the norm in a Hilbert space has the Radon-Riesz property, we also get strong convergence. The same approach works in reflexive Banach spaces.

A classical example is regularisation in Sobolev spaces such as the space $H^1$ of $L^2$ functions whose weak derivatives are also in $L^2$. In the one-dimensional case, the space $H^1$ consists only of continuous functions (in higher dimensions it is true for Sobolev spaces with some other exponents), therefore, the regularised solutions will also be continuous. For this reason, the regulariser $\mathcal{J}(u) = \|u\|_{H^1}$ is sometimes referred to as the *smoothing functional*. Whilst desirable in some applications, in imaging smooth reconstructions are usually not favourable, since images naturally contain edges and therefore are not continuous functions. To overcome this issue, other regularisers have been introduced that we will discuss later.

**Example 4.2.8** ($\ell^1$-regularisation)**.** Let $\mathcal{X} = \ell^2$ be space of all square summable sequences (i.e. such that $\|u\|_{\ell^2}^2 = \sum_{i=1}^{\infty} u_i^2 < +\infty$). For example, $u$ can represent the coefficients of a function in a basis (e.g., a Fourier basis or a wavelet basis). As a regularisation functional, let us use not the $\ell^2$-norm, but the $\ell^1$-norm:

$$\mathcal{J}(u) = \|u\|_{\ell^1} = \sum_{i=1}^{\infty} |u_i|.$$

By Example 4.1.5 $\mathcal{J}(\cdot)$ is weakly l.s.c. in $\ell^2$. It is evident that $\ell^q \subset \ell^p$ and $\|\cdot\|_{\ell^p} \leqslant \|\cdot\|_{\ell^q}$ for $q \leqslant p$. Therefore, $\mathcal{J}(u) \leqslant C$ implies that $\|\cdot\|_{\ell^2} \leqslant C$ and, since $\ell^2$ is a Hilbert space and bounded sequences have weakly convergent subsequences, we conclude that the sublevel sets of $\mathcal{J}(\cdot)$ are weakly sequentially compact in $\ell^2$. Therefore, Assumption (ii) of Theorem 4.2.3 is satisfied with $\tau_{\mathcal{X}}$ being the weak topology in $\ell^2$. Hence, we get weak convergence of regularised solutions in $\ell^2$.

The motivation for using the $\ell^1$-norm as the regulariser instead of the $\ell^2$-norm is as follows. If the forward operator is non-injective, the inverse problem has more than one solution and the solutions form an affine subspace. In the context of sequence spaces representing coefficients of the solution in a basis, it is sometimes beneficial to look for solutions that are *sparse* in the sense that they have finite support, i.e. $|\operatorname{supp}(u)| < \infty$ with $\operatorname{supp}(u) = \{i \in \mathbb{N} \,|\, u_i \neq 0\}$. This allows explaining the signal with a finite (and often relatively small) number of basis functions and has widely ranging applications in, for instance, compressed sensing. A finite dimensional illustration of the sparsity of $\ell^1$-regularised solutions is given in Figure 4.8. The corresponding minimisation problem

$$\min_{u \in \ell^2} \left\{ \frac{1}{2} \|Au - f\|_{\ell^2}^2 + \alpha \|u\|_1 \right\}. \tag{4.8}$$

is also called *lasso* in the statistical literature.

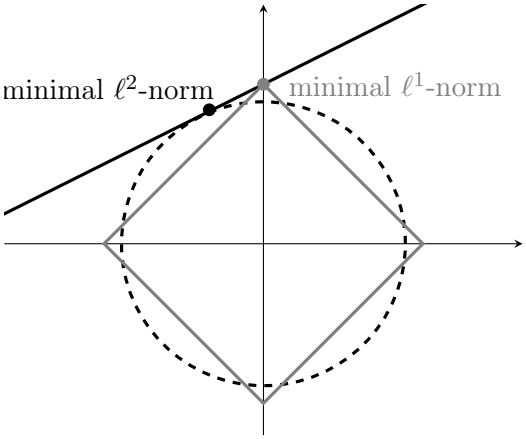



Figure 4.8: Non-injective operators have a non-trivial kernel such that the inverse problem has more than one solution and the solutions form an affine subspace visualised by the solid line. Different regularisation functionals favour different solutions. The circle and the diamond indicate all points with constant $\ell^2$-norm, respectively $\ell^1$-norm, and the minimal $\ell^2$-norm and $\ell^1$-norm solutions are the intersections of the line with the circle, respectively the diamond. As it can be seen, the minimal $\ell^2$-norm solution has two non-zero components while the minimal $\ell^1$-norm solution has only one non-zero component and thus is *sparser*.

## 4.3 Total Variation Regularisation

As pointed out in Example 4.2.7, in imaging we are interested in regularisers that allow for discontinuities while maintaining sufficient regularity of the reconstructions. One popular choice is the so-called *total variation* regulariser.

**Definition 4.3.1.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $u \in L^1(\Omega)$. Let $\mathcal{D}(\Omega, \mathbb{R}^n)$ be the following set of vector-valued* test functions *(i.e. functions that map from $\Omega$ to $\mathbb{R}^n$)*

$$\mathcal{D}(\Omega, \mathbb{R}^n) := \Big\{ \varphi \in C_0^\infty(\Omega; \mathbb{R}^n) \;\Big|\; \operatorname{ess\,sup}_{x\in\Omega} \|\varphi(x)\|_2 \leqslant 1 \Big\}.$$

*Total variation of $u \in L^1(\Omega)$ is defined as follows*

$$\mathrm{TV}(u) = \sup_{\varphi \in \mathcal{D}(\Omega,\mathbb{R}^n)} \int_\Omega u(x) \operatorname{div} \varphi(x)\, dx\,.$$

**Remark 4.3.2.** Definition 4.3.1 may seem a bit strange at the first glance, but we note that for a function $u \in L^1(\Omega)$ whose weak derivative $\nabla u$ exists and is also in $L^1(\Omega, \mathbb{R}^n)$ (i.e. $u$ belongs to the Sobolev space $W^{1,1}(\Omega)$) we obtain, integrating by parts, that

$$\mathrm{TV}(u) = \sup_{\varphi \in \mathcal{D}(\Omega,\mathbb{R}^n)} \int_\Omega - \langle \nabla u(x), \varphi(x) \rangle \; dx.$$

By the Cauchy-Schwartz inequality we get that $|\langle \nabla u(x), \varphi(x)\rangle| \leqslant \|\nabla u(x)\|_2 \|\varphi(x)\|_2 \leqslant \|\nabla u(x)\|_2$ for a.e. $x \in \Omega$. On the other hand, choosing $\varphi$ such that $\varphi(x) = -\frac{\nabla u(x)}{\|\nabla u(x)\|_2}$ (technically, such $\varphi$ is not necessarily in $\mathcal{D}(\Omega, \mathbb{R}^n)$, but we can approximate it with functions from

$\mathcal{D}(\Omega, \mathbb{R}^n)$, since any function in $W^{1,1}(\Omega)$ can be approximated with smooth functions [3, Thm. 3.17]; we omit the technicalities here), we get that $-\langle \nabla u(x), \varphi(x) \rangle = \|\nabla u(x)\|_2$. Therefore, the supremum over $\varphi \in \mathcal{D}(\Omega, \mathbb{R}^n)$ is equal to

$$\mathrm{TV}(u) = \int_\Omega \|\nabla u(x)\|_2 \, dx.$$

This shows that TV just penalises the the $L^1$ norm (of the pointwise 2-norm) of the gradient for any $u \in W^{1,1}(\Omega)$. However, we will see that the space of functions that have finite value of TV is larger than $W^{1,1}(\Omega)$ and contains, for instance, discontinuous functions.

**Proposition 4.3.3.** TV *is a proper and convex functional* $L^1(\Omega) \to \bar{\mathbb{R}}$. *For any constant function* $\mathbf{c}\colon \mathbf{c}(x) \equiv c \in \mathbb{R}$ *for all* $x$ *and any* $u \in L^1(\Omega)$

$$\mathrm{TV}(\mathbf{c}) = 0 \quad \textit{and} \quad \mathrm{TV}(u + \mathbf{c}) = \mathrm{TV}(u).$$

*Proof.* Left as exercise. □

**Definition 4.3.4.** *The functions* $u \in L^1(\Omega)$ *with a finite value of* TV *form a normed space called the space of functions of bounded variation (the* BV*-space) defined as follows*

$$\mathrm{BV}(\Omega) := \left\{ u \in L^1(\Omega) \;\middle|\; \|u\|_{\mathrm{BV}} := \|u\|_{L^1} + \mathrm{TV}(u) < \infty \right\}.$$

It can be shown that BV is a Banach space [5].

**Example 4.3.5** (TV of an indicator function)**.** Suppose $\mathcal{C} \subset \Omega \subset \mathbb{R}^2$ is a bounded domain with smooth boundary and $u(\cdot) = \mathbf{1}_{\mathcal{C}}(\cdot)$ is its indicator function, i.e.

$$\mathbf{1}_{\mathcal{C}}(u) = \begin{cases} 1 & u \in \mathcal{C} \\ 0 & u \in \mathcal{X} \setminus \mathcal{C} \end{cases}.$$

Then, using the divergence theorem, we get that for any test function $\varphi \in \mathcal{D}(\Omega, \mathbb{R}^n)$

$$\int_\Omega u(x) \operatorname{div} \varphi(x) \, dx = \int_{\mathcal{C}} \operatorname{div} \varphi(x) \, dx = \int_{\partial \mathcal{C}} \langle \varphi(x), \mathbf{n}_{\partial \mathcal{C}}(x) \rangle \, dl,$$

where $\partial \mathcal{C}$ is the boundary of $\mathcal{C}$ and $\mathbf{n}_{\partial \mathcal{C}}(x)$ is the unit normal at $x$. We, obviously, have that for every $x$

$$\langle \varphi(x), \mathbf{n}(x) \rangle = \frac{1}{2}(\|\varphi(x)\|^2 + \|\mathbf{n}_{\partial \mathcal{C}}(x)\|^2 - \|\varphi(x) - \mathbf{n}_{\partial \mathcal{C}}(x)\|^2),$$

so we get that

$$\mathrm{TV}(u) = \sup_{\varphi \in \mathcal{D}(\Omega, \mathbb{R}^n)} \int_{\partial \mathcal{C}} \frac{1}{2}(\|\varphi(x)\|^2 + \|\mathbf{n}_{\partial \mathcal{C}}(x)\|^2 - \|\varphi(x) - \mathbf{n}_{\partial \mathcal{C}}(x)\|^2) \, dl.$$

Since $\partial \mathcal{C}$ is smooth and $\|\mathbf{n}_{\partial \mathcal{C}}(x)\| = 1$ for every $x$, $\mathbf{n}_{\partial \mathcal{C}}$ can be extended to feasible vector field on $\Omega$ (i.e. one that is in $D(\Omega, \mathbb{R}^n)$) and the supremum is attained at $\varphi = \mathbf{n}_{\partial \mathcal{C}}$. Therefore, we get that

$$\mathrm{TV}(u) = \int_{\partial \mathcal{C}} \|\mathbf{n}_{\partial \mathcal{C}}(x)\|^2 \, dl = \int_{\partial \mathcal{C}} 1 \cdot \, dl = \mathrm{Per}(\mathcal{C}),$$

where $\mathrm{Per}(\mathcal{C})$ is the perimeter of $\mathcal{C}$.

Therefore, total variation of the characteristic function of a domain with smooth boundary is equal to its perimeter. This can be extended to domains with Lipschitz boundary by constructing a sequence of functions in $D(\Omega, \mathbb{R}^n)$ that converge pointwise to $\mathbf{n}_{\partial \mathcal{C}}$.

To apply Theorem 4.2.3, we need to study the properties of TV as a functional $L^1(\Omega) \to \bar{\mathbb{R}}$. First of all, we note that $\mathrm{BV}(\Omega)$ is compactly embedded in $L^1(\Omega)$. We start with the following classical result.

**Theorem 4.3.6** (Rellich-Kondrachov, [3, Thm. 6.3])**.** *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain (i.e. non-empty, open, connected and bounded with Lipschitz boundary) and either*

$$\begin{aligned} & n > mp \quad \text{and} \quad p^* := np/(n-mp) \\ \text{or} \quad & n \leqslant mp \quad \text{and} \quad p^* := \infty\,. \end{aligned}$$

*Then the embedding $W^{m,p}(\Omega) \to L^q(\Omega)$ is continuous if $1 \leqslant q \leqslant p^*$ and compact if in addition $q < p^*$.*

Since functions from $\mathrm{BV}(\Omega)$ can be approximated by smooth functions [5, Thm. 3.9], the Rellich-Kandrachov Theorem (for $m = 1, p = 1$) gives us compactness for $\mathrm{BV}(\Omega)$.

**Corollary 4.3.7** ([5, Corrollary 3.49])**.** For any bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ the embedding

$$\mathrm{BV}(\Omega) \to L^1(\Omega)$$

is compact.

Therefore, the level sets of $\mathcal{J}(u) = \|u\|_{\mathrm{BV}}$ are strongly sequentially compact in $L^1(\Omega)$. This is one of the ingredients we need to apply Theorem 4.2.3. The other one is lower-semicontinuity, which is guaranteed by the following theorem.

**Theorem 4.3.8.** *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then the total variation is strongly l.s.c. in $L^1(\Omega)$.*

*Proof.* Let $\{u_j\}_{j\in\mathbb{N}} \subset \mathrm{BV}(\Omega)$ be a sequence converging in $L^1(\Omega)$ with $u_j \to u$ in $L^1(\Omega)$. Then for any test function $\varphi \in \mathcal{D}(\Omega, \mathbb{R}^n)$ we have that

$$\int_\Omega [u(x) - u_j(x)] \operatorname{div} \varphi(x) dx \leqslant \underbrace{\int_\Omega |u(x) - u_j(x)| dx}_{= \|u - u_j\|_{L^1} \to 0} \underbrace{\operatorname{ess\,sup}_{x\in\Omega} |\operatorname{div} \varphi(x)|}_{<\infty} \to 0$$

and therefore

$$\begin{aligned} \int_\Omega u(x) \operatorname{div} \varphi(x) dx &= \lim_{j\to\infty} \int_\Omega u_j(x) \operatorname{div} \varphi(x) dx = \liminf_{j\to\infty} \int_\Omega u_j(x) \operatorname{div} \varphi(x) dx \\ &\leqslant \liminf_{j\to\infty} \sup_{\varphi\in\mathcal{D}(\Omega,\mathbb{R}^n)} \int_\Omega u_j(x) \operatorname{div} \varphi(x) dx = \liminf_{j\to\infty} \mathrm{TV}(u_j). \end{aligned}$$

Taking the supremum over all test functions on the left-hand side (and noting that the right-hand side already does not depend on $\varphi$), we get the assertion:

$$\mathrm{TV}(u) = \sup_{\varphi\in\mathcal{D}(\Omega,\mathbb{R}^n)} \int_\Omega u(x) \operatorname{div} \varphi(x) dx \leqslant \liminf_{j\to\infty} \mathrm{TV}(u_j).$$

Note that the left and right hand sides may not be finite. □

**Remark 4.3.9.** Combining these results, we conclude that with a suitable parameter choice rule the regulariser $\mathcal{J}(u) = \mathrm{TV}(u) + \|u\|_1$ ensures strong $L^1$-convergence of the regularised solutions. If the forward operator is such that boundedness of the fidelity term implies boundedness of $\|u\|_1$, then the term $\|u\|_1$ can be dropped and $\mathcal{J}(u) = \mathrm{TV}(u)$ can be used instead, ensuring the same convergence properties. See Remark 4.3.13 for an example of a situation when this is the case.

**Proposition 4.3.10** ([5, Remark 3.50])**.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain. Then there exists a constant $C > 0$ such that for all $u \in \mathrm{BV}(\Omega)$ the* Poincaré–Wirtinger *type inequality is satisfied*

$$\|u - u_\Omega\|_{L^1} \leqslant C\,\mathrm{TV}(u),$$

*where $u_\Omega := \frac{1}{|\Omega|}\int_\Omega u(x)dx$ is the mean-value of $u$ over $\Omega$.*

**Corollary 4.3.11.** It is often useful to consider a subspace $\mathrm{BV}_0(\Omega) \subset \mathrm{BV}(\Omega)$ of functions with zero mean, i.e.

$$\mathrm{BV}_0(\Omega) := \{u \in \mathrm{BV}(\Omega)\colon \int_\Omega u(x)dx = 0\}. \tag{4.9}$$

Then for every function $u \in \mathrm{BV}_0(\Omega)$ we have that

$$\|u\|_{L^1} \leqslant C\,\mathrm{TV}(u).$$

**Remark 4.3.12.** On two-dimensional domains $\Omega \subset \mathbb{R}^2$ one can show similar inequalities for the $L^2$ norm.

**Remark 4.3.13.** Many realistic forward operators satisfy the condition $A\mathbf{1} \neq 0$, where $\mathbf{1}(x) \equiv 1$ for all $x$. In this case, the boundedness of $\mathrm{TV}(u)$ together with the boundedness of the fidelity term $\|Au - f_\delta\|^2$ imply the boundedness of the mean value $u_\Omega \in \mathbb{R}$.

Indeed, suppose that there exists a sequence $u^n$ is such that $u^n_\Omega$ is unbounded. Then, since $A\mathbf{1} \neq 0$, the sequence $Au^n_\Omega$ is also unbounded. Consider $u^n_0 := u^n - u^n_\Omega \in \mathrm{BV}_0(\Omega)$. By Proposition 4.3.3 we have that

$$\mathrm{TV}(u^n_0) = \mathrm{TV}(u^n - u^n_\Omega) = \mathrm{TV}(u^n)$$

and therefore bounded. We also have that

$$\begin{aligned}\|Au^n_\Omega\| &= \|Au^n_\Omega + Au^n_0 - f_\delta - (Au^n_0 - f_\delta)\| \leqslant \|Au^n - f_\delta\| + \|Au^n_0 - f_\delta\| \\ &\leqslant \|Au^n - f_\delta\| + \|A\|\|u^n_0\| + \|f_\delta\|.\end{aligned}$$

The first term on the right-hand side is bounded by assumption; the second one is bounded due to Corollary 4.3.11; the third one is also obviously bounded. Therefore, $\|Au^n_\Omega\|$ is bounded, which is a contradiction.

Total Variation is widely used in imaging applications [28]. For instance, the so-called Rudin–Osher–Fatemi (ROF) model for image denoising [25] consists in minimising the following functional

$$\min_{u \in L^2(\Omega)} \|u - f_\delta\|_2^2 + \alpha\,\mathrm{TV}(u). \tag{4.10}$$

In this case, the forward operator is the identity operator (and $A\mathbf{1} \neq 0$ is satisfied trivially). More generally, one considers the following optimisation problem

$$\min_{u \in L^2(\Omega)} \|Au - f_\delta\|_2^2 + \alpha\,\mathrm{TV}(u), \tag{4.11}$$

where $A\colon L^2(\Omega) \to L^2(\Omega)$ is injective. Injectiveness is equivalent to the condition that $\ker A = \{0\}$ and guarantees that $A\mathbf{1} \neq 0$.

# Chapter 5

# Convex Duality

In Chapter 4 we have established convergence of a regularised solution $u_\delta$ to a $\mathcal{J}$-minimising solution $u^\dagger_{\mathcal{J}}$ as $\delta \to 0$. However, we didn't get any results on the *speed* of this convergence, which is referred to as the *convergence rate*.

In modern regularisation methods, convergence rates are usually studied using *Bregman distances* associated with the (convex) regularisation functional $\mathcal{J}$. Recall that for a convex functional $\mathcal{J}$, $u, v \in \mathcal{X}$ such that $\mathcal{J}(v) < \infty$ and $q \in \partial\mathcal{J}(v)$, the (generalised) Bregman distance is given by the following expression (cf. Def. 4.1.23)

$$D^q_{\mathcal{J}}(u, v) = \mathcal{J}(u) - \mathcal{J}(v) - \langle q, u - v\rangle \ .$$

Also widely used is the *symmetric* Bregman distance (cf. Def. 4.1.25) given by the following expression (here $p \in \partial\mathcal{J}(u)$)

$$D^{symm}_{\mathcal{J}}(u, v) = D^q_{\mathcal{J}}(u, v) + D^p_{\mathcal{J}}(v, u) = \langle p - q, u - v\rangle \ .$$

Bregman distances appear to be a natural distance measure between a regularised solution $u_\delta$ and a $\mathcal{J}$-minimising solution $u^\dagger_{\mathcal{J}}$. For instance, for classical $L^2$- regularisation with $\mathcal{J}(u) = \frac{1}{2}\|u\|^2_{\mathcal{X}}$, the subgradient at $u^\dagger_{\mathcal{J}}$ is $p_{u^\dagger_{\mathcal{J}}} = u^\dagger_{\mathcal{J}}$ (since $\mathcal{J}$ is differentiable) and we get the following expression

$$\begin{aligned} D^{u^\dagger_{\mathcal{J}}}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) &= \frac{1}{2}\|u_\delta\|^2_{\mathcal{X}} - \frac{1}{2}\|u^\dagger_{\mathcal{J}}\|^2_{\mathcal{X}} - \left\langle u^\dagger_{\mathcal{J}}, u_\delta - u^\dagger_{\mathcal{J}}\right\rangle \\ &= \frac{1}{2}(\|u_\delta\|^2_{\mathcal{X}} - 2\left\langle u^\dagger_{\mathcal{J}}, u_\delta\right\rangle + \|u^\dagger_{\mathcal{J}}\|^2_{\mathcal{X}}) = \frac{1}{2}\|u_\delta - u^\dagger_{\mathcal{J}}\|^2_{\mathcal{X}}, \end{aligned}$$

which happens to coincide with the symmetric Bregman distance. Therefore, in the classical $L^2$-case, the Bregman distance just measures the $L^2$-distance between a regularised solution and a $\mathcal{J}$-minimising solution.

We are looking for a convergence rate of the following form

$$D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) \leqslant \psi(\delta),$$

where $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a known function of $\delta$ such that $\psi(\delta) \to 0$ as $\delta \to 0$. To obtain such an estimate, we need to not only understand the convergence of $u_\delta$ (to $u^\dagger_{\mathcal{J}}$), but also that of the subgradient $p_\delta \in \partial\mathcal{J}(u_\delta)$, which should ideally converge to some $p_{\mathcal{J}} \in \partial\mathcal{J}(u^\dagger_{\mathcal{J}})$.

## 5.1 Dual Problem

Recall that $u_\delta$ solves the following problem

$$\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f_\delta\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u). \tag{5.1}$$

with an appropriately chosen $\alpha = \alpha(\delta)$. In this chapter we will assume that $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces and that the regulariser is proper, convex, l.s.c., absolute one-homogeneous and satisfies conditions of Theorem 4.2.3.

We will see that all subgradients $p_\delta \in \partial\mathcal{J}(u_\delta)$ are closely related to solutions of the dual problem of (5.1) in the sense of duality in convex optimisation [15].

**The saddle point problem.** Let us consider the function $\varphi\colon \mathcal{Y} \to \mathbb{R}$, $\varphi(x) := \frac{1}{2}\|x-f\|_{\mathcal{Y}}^2$, where $f \in \mathcal{Y}$ is a parameter. The Fenchel conjugate of $\varphi$ is given by

$$\varphi^*(\nu) = \sup_{x\in\mathcal{Y}} \langle \nu, x\rangle - \varphi(x) = \sup_{x\in\mathcal{Y}} \langle \nu, x\rangle - \frac{1}{2}\|x-f\|_{\mathcal{Y}}^2, \quad \nu \in \mathcal{Y}.$$

The supremum is attained at $x = \nu + f$ and therefore

$$\varphi^*(\nu) = \langle \nu, \nu + f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 = \langle \nu, f\rangle + \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2.$$

By Theorem 4.1.16 we have that $\varphi$ is equal to its biconjugate, i.e.

$$\varphi(x) = \sup_{\nu\in\mathcal{Y}} \langle \nu, x\rangle - \varphi^*(\nu) = \sup_{\nu\in\mathcal{Y}} \langle \nu, x\rangle - \langle \nu, f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 = \sup_{\nu\in\mathcal{Y}} \langle \nu, x-f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2.$$

For $x = Au$, therefore, we get that

$$\varphi(Au) = \frac{1}{2}\|Au-f\|_{\mathcal{Y}}^2 = \sup_{\nu\in\mathcal{Y}} \langle \nu, Au-f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2.$$

The objective function attains its maximum at $\nu = Au - f$, so we can replace the supremum with a maximum. Now we can rewrite (5.1) as follows

$$\min_{u\in\mathcal{X}} \max_{\nu\in\mathcal{Y}} \langle \nu, Au-f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u). \tag{5.2}$$

Problem (5.2) is called the *saddle-point* problem. If it has a solution then we can easily derive optimality conditions by differentiating the objective function in $u$ and $\nu$:

$$\nu = Au - f, \quad A^*\left(\frac{-\nu}{\alpha}\right) \in \partial\mathcal{J}(u). \tag{5.3}$$

**The dual problem.** If there exists a point $x \in \mathcal{X}$ s.t. $\varphi(x) < +\infty$, $\mathcal{J}(x) < +\infty$ and $\varphi(x)$ is continuous at $x$, we can swap the minimum and the maximum in (5.2) [15, Ch.III Thm 4.1 and Rem. 4.2]. These conditions are satisfied, e.g., at $x = 0$. Hence we get

$$\begin{aligned}\min_{u\in\mathcal{X}} \max_{\nu\in\mathcal{Y}} \langle \nu, Au-f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u) &= \max_{\nu\in\mathcal{Y}} \min_{u\in\mathcal{X}} \langle \nu, Au-f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u) \\ &= \max_{\nu\in\mathcal{Y}} \left\{ \left[\min_{u\in\mathcal{X}} \langle \nu, Au\rangle + \alpha\mathcal{J}(u)\right] - \langle \nu, f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2 \right\}. \end{aligned} \tag{5.4}$$

The minimum of the expression in the square brackets is given by

$$\begin{aligned}\min_{u\in\mathcal{X}} \langle \nu, Au\rangle + \alpha\mathcal{J}(u) &= \min_{u\in\mathcal{X}} \langle A^*\nu, u\rangle + \alpha\mathcal{J}(u)\\ &= -\alpha \max_{u\in\mathcal{X}} \left\langle A^*\left(\frac{-\nu}{\alpha}\right), u\right\rangle - \mathcal{J}(u) = -\alpha\mathcal{J}^*\left(A^*\left(\frac{-\nu}{\alpha}\right)\right).\end{aligned}$$

Since $\mathcal{J}$ is absolute one-homogeneous, its Fenchel conjugate is the characteristic function of $\partial\mathcal{J}(0)$ (Prop. 4.1.29) and we get

$$\min_{u\in\mathcal{X}} \langle \nu, Au\rangle + \alpha\mathcal{J}(u) = -\alpha\chi_{\partial\mathcal{J}(0)}\left(A^*\left(\frac{-\nu}{\alpha}\right)\right).$$

Substituting this into (5.4), we get

$$\max_{\nu\in\mathcal{Y}}\left\{\left[\min_{u\in\mathcal{X}} \langle \nu, Au\rangle + \alpha\mathcal{J}(u)\right] - \langle \nu, f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2\right\} = \max_{\nu\in\mathcal{Y}:\ A^*\left(\frac{-\nu}{\alpha}\right)\in\partial\mathcal{J}(0)} - \langle \nu, f\rangle - \frac{1}{2}\|\nu\|_{\mathcal{Y}}^2.$$

Denoting $\mu := -\frac{\nu}{\alpha} \in \mathcal{Y}$, we rewrite this problem as follows

$$\max_{\mu\in\mathcal{Y}:\ A^*\mu\in\partial\mathcal{J}(0)} \alpha\left(\langle \mu, f\rangle - \frac{\alpha}{2}\|\mu\|_{\mathcal{Y}}^2\right). \tag{5.5}$$

Problem (5.5) is called the *dual problem*. With this notation, optimality conditions (5.3) take the following form

$$A^*\mu \in \partial\mathcal{J}(u), \quad \mu = \frac{f - Au}{\alpha}. \tag{5.6}$$

**Weak and strong duality.** It can be shown [15] that for any feasible solution $u_0$ of the primal problem (5.1) and for any feasible solution $\mu_0$ of the dual problem (5.5), the objective value of the dual problem does not exceed that of the primal problem, i.e.

$$\frac{1}{2}\|Au_0 - f\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u_0) \geqslant \alpha\langle \mu_0, f\rangle - \frac{\alpha^2}{2}\|\mu_0\|_{\mathcal{Y}}^2. \tag{5.7}$$

This also holds for the optimal solutions $u_\delta$ and $\mu_\delta$ in the case $f = f_\delta$. The difference

$$\frac{1}{2}\|Au_\delta - f_\delta\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u_\delta) - \left(\alpha\langle \mu_\delta, f_\delta\rangle - \frac{\alpha^2}{2}\|\mu_\delta\|_{\mathcal{Y}}^2\right) \geqslant 0 \tag{5.8}$$

is referred to as the *duality gap*. The fact that it is always non-negative is referred to as *weak duality*. If the duality gap is zero, it is said that *strong duality* holds.

**Existence.** While a solution of the primal problem (5.1) exists by Theorem 4.2.3, existence of a dual solution of (5.5) is not always guaranteed. Sufficient conditions are given by

**Theorem 5.1.1** ([15, Ch.III Thm 4.1 and Rem. 4.2])**.** *Consider the following optimisation problem*

$$\inf_{u\in\mathcal{X}} E(Au) + F(u), \tag{$\mathcal{P}$}$$

*where $E\colon \mathcal{Y} \to \bar{\mathbb{R}}$ and $F\colon \mathcal{X} \to \bar{\mathbb{R}}$. Suppose that*

*(i) the function* $E(Au) + F(u) \colon \mathcal{X} \to \bar{\mathbb{R}}$ *is proper, convex, l.s.c. and coercive;*

*(ii)* $\exists u_0 \in \mathcal{X}$ *s.t.* $F(u_0) < +\infty$, $E(Au_0) < +\infty$ *and* $E(x)$ *is continuous at* $x = Au_0$.

*Then*

*(i) The dual problem of* $(\mathcal{P})$ *has at least one solution* $\widehat{\eta}$*;*

*(ii) There is no duality gap between* $(\mathcal{P})$ *and its dual, i.e. strong duality holds;*

*(iii) If* $(\mathcal{P})$ *has an optimal solution* $\widehat{u}$*, then the following optimality conditions hold*

$$A^*\widehat{\eta} \in \partial F(\widehat{u}), \quad -\widehat{\eta} \in \partial E(A\widehat{u}).$$

In our case, $E(Au) = \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2$ and $F(u) = \alpha\mathcal{J}(u)$. Condition (i) is satisfied by the assumptions of Theorem 4.2.3 (in particular, coercivity is implied by the compactness of the sub-level sets). Condition (ii) is satisfied at $u_0 = 0$. Therefore, for any $\delta > 0$ there exists a solution $\mu_\delta$ of the dual problem (5.5) and by strong duality we have that

$$\frac{1}{2}\|Au_\delta - f_\delta\|_{\mathcal{Y}}^2 + \alpha\mathcal{J}(u_\delta) = \alpha\left\langle \mu_\delta, f_\delta \right\rangle - \frac{\alpha^2}{2}\|\mu_\delta\|_{\mathcal{Y}}^2.$$

Optimality conditions (iii) in this case take the following form (cf. (5.6))

$$A^*\mu_\delta \in \partial\mathcal{J}(u_\delta), \quad \mu_\delta = \frac{f_\delta - Au_\delta}{\alpha(\delta)}. \tag{5.9}$$

## 5.2 Source Condition

Formal limits of problems (5.1) and (5.5) at $\delta = 0$ are

$$\inf_{u\colon Au=f} \mathcal{J}(u) = \inf_{u\in\mathcal{X}} \chi_{\{f\}}(Au) + \mathcal{J}(u) \tag{5.10}$$

and

$$\begin{aligned}\sup_{\mu\colon A^*\mu\in\partial\mathcal{J}(0)} \langle \mu, f\rangle = & \sup_{\mu\colon A^*\mu\in\partial\mathcal{J}(0)} \left\langle \mu, Au_{\mathcal{J}}^\dagger \right\rangle \\ = & \sup_{\mu\colon A^*\mu\in\partial\mathcal{J}(0)} \left\langle A^*\mu, u_{\mathcal{J}}^\dagger \right\rangle = \sup_{v\in\mathcal{R}(A^*)\cap\partial\mathcal{J}(0)} \left\langle v, u_{\mathcal{J}}^\dagger \right\rangle. \end{aligned} \tag{5.11}$$

Since the characteristic function $\chi_{\{f\}}(\cdot)$ is not continuous anywhere in its domain, Theorem 5.1.1 does not apply and we cannot guarantee that a solution of the dual limit problem (5.11) exists. Indeed, since $\mathcal{R}(A^*)$ is not closed (strongly and hence weakly, since it is convex [14, Thm. V.3.13]), a solution may not exist.

Therefore, the behaviour of $\mu_\delta$ as $\delta \to 0$ is unclear. A natural question to ask is whether $\mu_\delta$ remains bounded as $\delta \to 0$. The following condition will play an important role in this.

**Definition 5.2.1** (Source condition [11])**.** *We say that a* $\mathcal{J}$*-minimising solution* $u_{\mathcal{J}}^\dagger$ *satisfies the* source condition *if*

$$\exists \mu^\dagger \in \mathcal{Y} \quad \textit{such that} \quad A^*\mu^\dagger \in \partial\mathcal{J}(u_{\mathcal{J}}^\dagger), \tag{5.12}$$

*i.e. if* $\mathcal{R}(A^*) \cap \partial\mathcal{J}(u_{\mathcal{J}}^\dagger) \neq \emptyset$.

**Theorem 5.2.2** (Necessary conditions, [21]). *Suppose that conditions of Theorem 4.2.3 are satisfied with $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{Y}}$ being weak topologies in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Suppose that $\mu_\delta$ is bounded uniformly in $\delta$. Then there exists $\mu^\dagger \in \mathcal{Y}$ such that $A^*\mu^\dagger \in \partial\mathcal{J}(u^\dagger_{\mathcal{J}})$.*

*Proof.* Consider an arbitrary sequence $\delta_n \downarrow 0$. Since $\|\mu_\delta\|_{\mathcal{Y}} \leqslant C$ for all $\delta$, by weak compactness of a ball in a Hilbert space we get that there exists a weakly convergent subsequence (that we do not relabel), i.e.

$$\mu_{\delta_n} \rightharpoonup \mu_0 \in \mathcal{Y}.$$

By the weak-weak continuity of $A^*$ we get that

$$A^*\mu_{\delta_n} \rightharpoonup A^*\mu_0.$$

Since $\partial\mathcal{J}(0)$ is weakly closed (Theorem 4.1.19) and $A^*\mu_{\delta_n} \in \partial\mathcal{J}(0)$ (see optimality conditions (5.9), we get that

$$A^*\mu_0 \in \partial\mathcal{J}(0).$$

Since $\mathcal{J}$ is absolute one-homogeneous, we get by Proposition 4.1.27 that

$$\langle A^*\mu_{\delta_n}, u_{\delta_n}\rangle = \mathcal{J}(u_{\delta_n}).$$

We also note that

$$\begin{aligned}\langle A^*\mu_{\delta_n}, u_{\delta_n}\rangle &= \left\langle A^*\mu_{\delta_n}, u^\dagger_{\mathcal{J}}\right\rangle + \left\langle A^*\mu_{\delta_n}, u_{\delta_n} - u^\dagger_{\mathcal{J}}\right\rangle \\ &= \left\langle A^*\mu_{\delta_n}, u^\dagger_{\mathcal{J}}\right\rangle + \langle \mu_{\delta_n}, Au_{\delta_n} - f\rangle \leqslant \left\langle A^*\mu_{\delta_n}, u^\dagger_{\mathcal{J}}\right\rangle + \|\mu_{\delta_n}\|_{\mathcal{Y}}\|Au_{\delta_n} - f\|_{\mathcal{Y}}.\end{aligned}$$

Since $\|\mu_{\delta_n}\|_{\mathcal{Y}}$ is bounded and $\|Au_{\delta_n} - f\|_{\mathcal{Y}} \to 0$, we get that

$$\langle A^*\mu_{\delta_n}, u_{\delta_n}\rangle \to \left\langle A^*\mu_0, u^\dagger_{\mathcal{J}}\right\rangle.$$

On the other hand, we know that $\mathcal{J}(u_{\delta_n}) \to \mathcal{J}(u^\dagger_{\mathcal{J}})$. Therefore, we get that

$$\mathcal{J}(u^\dagger_{\mathcal{J}}) = \left\langle A^*\mu_0, u^\dagger_{\mathcal{J}}\right\rangle.$$

Since $A^*\mu_0 \in \partial\mathcal{J}(0)$ and $\mathcal{J}(u^\dagger_{\mathcal{J}}) = \left\langle A^*\mu_0, u^\dagger_{\mathcal{J}}\right\rangle$, we conclude, using Proposition 4.1.30, that $A^*\mu_0 \in \partial\mathcal{J}(u^\dagger_{\mathcal{J}})$ and the assertion of the Theorem holds with $\mu^\dagger = \mu_0$. □

So, the source condition is necessary for the boundedness of $\mu_\delta$. It turns out to be also sufficient.

**Theorem 5.2.3** (Sufficient conditions, [21]). *Suppose that the source condition* (5.12) *is satisfied at a $\mathcal{J}$-minimising solution $u^\dagger_{\mathcal{J}}$ and suppose that $\alpha(\delta)$ is chosen such that $\frac{\delta}{\alpha(\delta)} \to 0$ as $\delta \to 0$. Then $\mu_\delta$ is bounded uniformly in $\delta$. Moreover, $\mu_\delta \to \mu^\dagger$ strongly in $\mathcal{Y}$ as $\delta \to 0$, where $\mu^\dagger$ is the solution of the dual limit problem* (5.11) *with minimal norm.*

*Proof.* The source condition (5.12) guarantees that $\exists \mu_0 \in \mathcal{Y}$ s.t. $A^*\mu_0 \in \partial\mathcal{J}(u^\dagger_{\mathcal{J}})$, i.e. that

$$\begin{cases} A^*\mu_0 \in \partial\mathcal{J}(0), \\ \mathcal{J}(u^\dagger_{\mathcal{J}}) = \left\langle A^*\mu_0, u^\dagger_{\mathcal{J}}\right\rangle = \left\langle \mu_0, Au^\dagger_{\mathcal{J}}\right\rangle = \langle \mu_0, f\rangle. \end{cases}$$

From weak duality between the limit primal (5.10) and dual (5.11) problems we conclude that for any feasible solution $\mu$ of the dual limit problem (5.11)

$$\langle \mu, f \rangle \leqslant \mathcal{J}(u^{\dagger}_{\mathcal{J}}),$$

Therefore, $\mu_0$ solves the dual limit problem (5.11) and

$$\langle \mu_0, f \rangle \geqslant \langle \mu_\delta, f \rangle \quad \forall \delta, \tag{5.13}$$

since $\mu_\delta$ is feasible in (5.11).

Analogously, since $\mu_\delta$ solves the dual problem (5.5) and $\mu_0$ is feasible in (5.5), we get that for all $\delta$

$$\langle \mu_\delta, f_\delta \rangle - \frac{\alpha}{2}\|\mu_\delta\|^2_{\mathcal{Y}} \geqslant \langle \mu_0, f_\delta \rangle - \frac{\alpha}{2}\|\mu_0\|^2_{\mathcal{Y}}. \tag{5.14}$$

Summing up these two estimates and rearranging terms, we get that

$$\frac{\alpha}{2}\|\mu_\delta\|^2_{\mathcal{Y}} - \frac{\alpha}{2}\|\mu_0\|^2_{\mathcal{Y}} \leqslant \langle \mu_0 - \mu_\delta, f - f_\delta \rangle \leqslant \delta\|\mu_0 - \mu_\delta\|.$$

Noting that

$$\frac{\alpha}{2}\|\mu_\delta\|^2_{\mathcal{Y}} - \frac{\alpha}{2}\|\mu_0\|^2_{\mathcal{Y}} = \frac{\alpha}{2}(\|\mu_\delta\|_{\mathcal{Y}} - \|\mu_0\|_{\mathcal{Y}})(\|\mu_\delta\|_{\mathcal{Y}} + \|\mu_0\|_{\mathcal{Y}}),$$

we get that

$$\frac{\alpha}{2}(\|\mu_\delta\|_{\mathcal{Y}} - \|\mu_0\|_{\mathcal{Y}})(\|\mu_0\|_{\mathcal{Y}} + \|\mu_\delta\|_{\mathcal{Y}}) \leqslant \delta\|\mu_0 - \mu_\delta\|_{\mathcal{Y}} \leqslant \delta(\|\mu_0\|_{\mathcal{Y}} + \|\mu_\delta\|_{\mathcal{Y}})$$

and

$$\|\mu_\delta\|_{\mathcal{Y}} \leqslant \|\mu_0\|_{\mathcal{Y}} + \frac{2\delta}{\alpha} \leqslant C, \tag{5.15}$$

since $\frac{\delta}{\alpha}$ is bounded.

By weak compactness of a ball in a Hilbert space, we conclude that for any sequence $\delta_n \downarrow 0$ there exists a subsequence (which we do not relabel) such that

$$\mu_{\delta_n} \rightharpoonup \mu^*.$$

By weak-weak continuity of $A^*$ and weak closedness of $\partial\mathcal{J}(0)$ (Theorem 4.1.19) we get that

$$A^*\mu^* \in \partial\mathcal{J}(0)$$

and $\mu^*$ is feasible in (5.11).

From (5.14) we obtain that

$$\begin{aligned}
\langle \mu_0, f_\delta \rangle &\leqslant \langle \mu_\delta, f_\delta \rangle + \frac{\alpha}{2}(\|\mu_0\|^2_{\mathcal{Y}} - \|\mu_{\delta_n}\|^2_{\mathcal{Y}}) \\
&\leqslant \langle \mu_\delta, f \rangle + \langle \mu_\delta, f_\delta - f \rangle + \frac{\alpha}{2}(\|\mu_0\|^2_{\mathcal{Y}} - \|\mu_{\delta_n}\|^2_{\mathcal{Y}}) \\
&\leqslant \langle \mu_0, f \rangle + \delta\|\mu_\delta\| + \frac{\alpha}{2}(\|\mu_0\|^2_{\mathcal{Y}} - \|\mu_{\delta_n}\|^2_{\mathcal{Y}}).
\end{aligned}$$

Taking the limit $\delta \to 0$, we get that

$$\langle \mu_0, f \rangle \leqslant \langle \mu^*, f \rangle$$

Taking the limit in (5.13) also yields

$$\langle \mu_0, f \rangle \geqslant \langle \mu^*, f \rangle ,$$

hence $\langle \mu_0, f \rangle = \langle \mu^*, f \rangle$ and $\mu^*$ solves the dual limit problem (5.11).

Using weak lower semicontinuity of the norm in a Hilbert space, from (5.15) we get that

$$\|\mu^*\|_{\mathcal{Y}} \leqslant \liminf_{n\to\infty} \|\mu_{\delta_n}\|_{\mathcal{Y}} \leqslant \limsup_{n\to\infty} \|\mu_{\delta_n}\|_{\mathcal{Y}} \leqslant \|\mu_0\|_{\mathcal{Y}} + \limsup_{n\to\infty} \frac{\delta_n}{\alpha(\delta_n)} = \|\mu_0\|_{\mathcal{Y}} \tag{5.16}$$

for any $\mu_0$ solving (5.11). Therefore, $\mu^*$ is the minimum norm solution of (5.11). Since (5.16) holds for any solution $\mu_0$ of the dual limit problem (5.15), including $\mu_0 = \mu^*$, we than also get that

$$\exists \lim_{n\to\infty} \|\mu_{\delta_n}\|_{\mathcal{Y}} = \|\mu^*\|_{\mathcal{Y}}$$

and the convergence $\mu_{\delta_n} \to \mu^*$ is actually strong by the Radon-Riesz property of the norm in a Hilbert space (see Remark 4.2.6). Hence, we get the assertion of the Theorem with $\mu^\dagger = \mu^*$. □

**Example 5.2.4** (Total Variation)**.** Let $\mathcal{X} = \mathcal{Y} = L^2(\Omega)$ with $\Omega \subset \mathbb{R}^2$ bounded and $\mathcal{C} \subset \Omega$ a domain with a $C^\infty$ boundary. Let $\mathcal{J}(\cdot) = \mathrm{TV}(\cdot)$ and $A\colon L^2(\Omega) \to L^2(\Omega)$ be the identity operator (i.e., we consider the problem of denoising). From Example 4.3.5 we know that

$$\mathrm{TV}(\mathbf{1}_{\mathcal{C}}) = \mathrm{Per}(\mathcal{C}),$$

where $\mathbf{1}_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}$. Denoting by $\mathbf{n}_{\partial\mathcal{C}}$ the unit normal, we obtain

$$\mathrm{Per}(\mathcal{C}) = \int_{\partial\mathcal{C}} 1 = \int_{\partial\mathcal{C}} \langle \mathbf{n}_{\partial\mathcal{C}}, \mathbf{n}_{\partial\mathcal{C}} \rangle .$$

Since $\mathbf{n}_{\partial\mathcal{C}} \in C^\infty(\partial\mathcal{C}, \mathbb{R}^2)$ and $\|\mathbf{n}_{\partial\mathcal{C}}(x)\|_2 = 1$ for any $x$, we can extend $\mathbf{n}_{\partial\mathcal{C}}$ to a $C_0^\infty(\Omega, \mathbb{R}^2)$ vector field $\psi$ with $\sup_{x\in\Omega} \|\psi(x)\|_2 \leqslant 1$. Therefore, using the divergence theorem, we obtain that

$$\int_{\partial\mathcal{C}} \langle \mathbf{n}_{\partial\mathcal{C}}, \mathbf{n}_{\partial\mathcal{C}} \rangle = \int_{\partial\mathcal{C}} \langle \psi, \mathbf{n}_{\partial\mathcal{C}} \rangle = \int_{\mathcal{C}} \operatorname{div} \psi = \langle \operatorname{div} \psi, \mathbf{1}_{\mathcal{C}} \rangle .$$

Combining all these equalities, we get that

$$\mathrm{TV}(\mathbf{1}_{\mathcal{C}}) = \langle \operatorname{div} \psi, \mathbf{1}_{\mathcal{C}} \rangle .$$

Note that, since $\psi \in C_0^\infty(\Omega, \mathbb{R}^2)$, $\operatorname{div} \psi \in C_0^\infty(\Omega) \subset L^2(\Omega)$.

Taking an arbitrary $u \in \mathcal{X}$, we note that

$$\mathrm{TV}(u) - \langle \operatorname{div} \psi, u \rangle = \sup_{\substack{\varphi \in C_0^\infty(\Omega, \mathbb{R}^2) \\ \sup_{x\in\Omega} \|\varphi(x)\|_2 \leqslant 1}} \langle u, \operatorname{div} \varphi \rangle - \langle u, \operatorname{div} \psi \rangle \geqslant 0,$$

since $\varphi = \psi$ is feasible. Therefore, $\operatorname{div} \psi \in \partial\, \mathrm{TV}(0)$ and, since $\mathrm{TV}(\mathbf{1}_{\mathcal{C}}) = \langle \operatorname{div} \psi, \mathbf{1}_{\mathcal{C}} \rangle$, we also get that

$$\operatorname{div} \psi \in \partial\, \mathrm{TV}(\mathbf{1}_{\mathcal{C}}).$$

Since $A$ is the identity operator, $\mathcal{R}(A^*) = \mathcal{X}$ and the source condition is satisfied at $u = \mathbf{1}_{\mathcal{C}}$ with $\mu = \operatorname{div} \psi$.

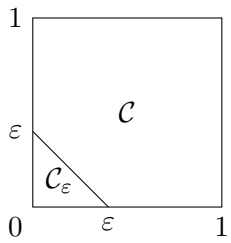



Figure 5.1: Example of a set whose indicator function does not satisfy the source condition.

**Example 5.2.5** (Total Variation)**.** Let $\mathcal{X} = \mathcal{Y} = L^2(\Omega)$ with $\Omega \subset \mathbb{R}^2$ bounded and $\mathcal{C} \subset \Omega$ be a domain with a nonsmooth boundary, e.g., a square $\mathcal{C} = [0,1]^2$. Let $\mathcal{J}(\cdot) = \mathrm{TV}(\cdot)$. We will show in this example that in this case $\partial\,\mathrm{TV}(\mathbf{1}_\mathcal{C}) = \emptyset$.

Assume that there exists $p_0 \in \partial\,\mathrm{TV}(\mathbf{1}_\mathcal{C}) \subset L^2(\Omega)$. Then by the results of Example 4.3.5 we have that

$$\langle p_0, \mathbf{1}_\mathcal{C} \rangle = \mathrm{TV}(\mathbf{1}_\mathcal{C}) = \mathrm{Per}(\mathcal{C}) = 4.$$

Since $p_0$ is a subgradient, we get that for any $u \in L^2(\Omega)$

$$\mathrm{TV}(u) - \langle p_0, u \rangle \geqslant 0.$$

Let us cut a triangle $\mathcal{C}_\varepsilon$ of size $\varepsilon$ from a corner of $\mathcal{C}$ as shown in Figure 5.1. Then for $u = \mathbf{1}_{\mathcal{C}\setminus\mathcal{C}_\varepsilon}$ we get

$$\mathrm{TV}(\mathbf{1}_{\mathcal{C}\setminus\mathcal{C}_\varepsilon}) \geqslant \left\langle p_0, \mathbf{1}_{\mathcal{C}\setminus\mathcal{C}_\varepsilon} \right\rangle = \langle p_0, \mathbf{1}_\mathcal{C} \rangle - \langle p_0, \mathbf{1}_{\mathcal{C}_\varepsilon} \rangle$$

and therefore

$$\langle p_0, \mathbf{1}_{\mathcal{C}_\varepsilon} \rangle \geqslant \mathrm{TV}(\mathbf{1}_\mathcal{C}) - \mathrm{TV}(\mathbf{1}_{\mathcal{C}\setminus\mathcal{C}_\varepsilon}) = \mathrm{Per}(\mathcal{C}) - \mathrm{Per}(\mathcal{C}\setminus\mathcal{C}_\varepsilon) = 4 - (4 - 2\varepsilon + \sqrt{2}\varepsilon) = (2-\sqrt{2})\varepsilon > 0.$$

By Hölder's inequality we get that

$$\langle p_0, \mathbf{1}_{\mathcal{C}_\varepsilon} \rangle = \int_{\mathcal{C}_\varepsilon} p_0 \cdot \mathbf{1} \leqslant \left( \int_{\mathcal{C}_\varepsilon} |p_0|^2 \right)^{1/2} \left( \int_{\mathcal{C}_\varepsilon} 1 \right)^{1/2} = \frac{1}{\sqrt{2}} \varepsilon \left( \int_{\mathcal{C}_\varepsilon} |p_0|^2 \right)^{1/2}.$$

Combining the last two inequalities, we get

$$(2-\sqrt{2})\varepsilon \leqslant \langle p_0, \mathbf{1}_{\mathcal{C}_\varepsilon} \rangle \leqslant \frac{1}{\sqrt{2}} \varepsilon \left( \int_{\mathcal{C}_\varepsilon} |p_0|^2 \right)^{1/2}$$

and therefore

$$\int_{\mathcal{C}_\varepsilon} |p_0|^2 \geqslant 2(2-\sqrt{2})^2 > 0$$

for all $\varepsilon > 0$. However, since $p_0 \in L^2(\Omega)$ by assumption, we must have

$$\int_{\mathcal{C}_\varepsilon} |p_0|^2 \to 0 \quad \text{as } \varepsilon \to 0.$$

This contradiction proves that such $p_0$ does not exist and $\partial\,\mathrm{TV}(\mathbf{1}_\mathcal{C}) = \emptyset$.

## 5.3 Convergence Rates

Now we are ready to answer the question that we asked in the beginning of this Chapter - *how fast* do the regularised solutions converge to a $\mathcal{J}$-minimising solution? The answer is given by the following Theorem.

**Theorem 5.3.1.** *Let the source condition* (5.12) *be satisfied at a $\mathcal{J}$-minimising solution $u^\dagger_{\mathcal{J}}$ and let $u_\delta$ be a regularised solution solving* (5.1)*. Then the following estimate holds*

$$D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) \leqslant \frac{1}{2\alpha}\left(\delta + \alpha\|\mu^\dagger\|_{\mathcal{Y}}\right)^2.$$

*Proof.* Consider the function

$$\varphi(g) = \frac{1}{2}\|g - f_\delta\|^2_{\mathcal{Y}}.$$

It is convex and differentiable and its subdifferential is given by

$$\partial\varphi(g) = \{g - f_\delta\}.$$

Hence the Bregman distance w.r.t. $\varphi$ between $g$ and $f$ corresponding to the subgradient $g - f_\delta \in \partial\varphi(g)$ is given by

$$D^{g-f_\delta}_{\varphi}(f, g) = \frac{1}{2}\|f - f_\delta\|^2_{\mathcal{Y}} - \frac{1}{2}\|g - f_\delta\|^2_{\mathcal{Y}} - \langle g - f_\delta, f - g\rangle \geqslant 0.$$

Hence, taking $g = Au_\delta$, we get that

$$\langle Au_\delta - f_\delta, f - Au_\delta\rangle \leqslant \frac{1}{2}\|f - f_\delta\|^2_{\mathcal{Y}} - \frac{1}{2}\|Au_\delta - f_\delta\|^2_{\mathcal{Y}} \leqslant \frac{\delta^2}{2} - \frac{1}{2}\|Au_\delta - f_\delta\|^2_{\mathcal{Y}}.$$

Consider the symmetric Bregman distance $D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}})$

$$\begin{aligned} D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) &= \left\langle A^*\mu^\dagger - A^*\mu_\delta, u^\dagger_{\mathcal{J}} - u_\delta\right\rangle = \left\langle \mu^\dagger - \mu_\delta, f - Au_\delta\right\rangle \\ &= \left\langle \mu^\dagger, f - Au_\delta\right\rangle + \left\langle -\mu_\delta, f - Au_\delta\right\rangle. \end{aligned}$$

Multiplying this by $\alpha$ and noting that $\alpha\mu_\delta = f_\delta - Au_\delta$ from the optimality conditions (5.9), we get

$$\begin{aligned} \alpha D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) &= \alpha\left\langle \mu^\dagger, f - Au_\delta\right\rangle + \langle Au_\delta - f_\delta, f - Au_\delta\rangle \\ &\leqslant \alpha\left\langle \mu^\dagger, f_\delta - Au_\delta\right\rangle + \alpha\left\langle \mu^\dagger, f - f_\delta\right\rangle + \frac{1}{2}\delta^2 - \frac{1}{2}\|Au_\delta - f_\delta\|^2_{\mathcal{Y}} \\ &\leqslant \alpha\|\mu^\dagger\|_{\mathcal{Y}}\left(\|f_\delta - Au_\delta\|_{\mathcal{Y}} + \|f - f_\delta\|_{\mathcal{Y}}\right) + \frac{1}{2}\delta^2 - \frac{1}{2}\|Au_\delta - f_\delta\|^2_{\mathcal{Y}}. \end{aligned}$$

Noting that

$$\left(\frac{1}{2}\|Au_\delta - f_\delta\|^2_{\mathcal{Y}} - \alpha\|\mu^\dagger\|_{\mathcal{Y}}\|Au_\delta - f_\delta\|_{\mathcal{Y}} + \frac{1}{2}\alpha^2\|\mu^\dagger\|_{\mathcal{Y}}\right)^2 = \frac{1}{2}\left(\|Au_\delta - f_\delta\|_{\mathcal{Y}} - \alpha\|\mu^\dagger\|_{\mathcal{Y}}\right)^2,$$

we get that

$$\begin{aligned} \alpha D^{symm}_{\mathcal{J}}(u_\delta, u^\dagger_{\mathcal{J}}) &\leqslant \alpha\delta\|\mu^\dagger\|_{\mathcal{Y}} - \frac{1}{2}\left(\|Au_\delta - f_\delta\|_{\mathcal{Y}} - \alpha\|\mu^\dagger\|_{\mathcal{Y}}\right)^2 + \frac{1}{2}\alpha^2\|\mu^\dagger\|_{\mathcal{Y}} + \frac{1}{2}\delta^2 \\ &\leqslant \frac{1}{2}\delta^2 + \alpha\delta\|\mu^\dagger\|_{\mathcal{Y}} + \frac{1}{2}\alpha^2\|\mu^\dagger\|_{\mathcal{Y}} = \frac{1}{2}\left(\delta + \alpha\|\mu^\dagger\|_{\mathcal{Y}}\right)^2, \end{aligned}$$

which yields the desired estimate upon dividing by $\alpha$. □

# Chapter 6

# Bayesian approach to discrete inverse problems

The idea of Bayesian inversion is to rephrase the deterministic inverse problem studied above as a question of statistical inference. We consider model

$$F = AU + N, \tag{6.1}$$

where the measurement, unknown and noise are now modelled as random variables. Let $\Omega = \Omega_1 \times \Omega_2$ be our probability space. Then $U : \Omega_1 \to \mathbb{R}^d$, $N : \Omega_2 \to \mathbb{R}^k$ and $F : \Omega \to \mathbb{R}^k$. We denote the realisations of the random variables by lower case letter, e.g. for a fixed $\omega$ we write $U(\omega) = u$.

This approach allows us to model the noise through its statistical properties. We can also encode our *a priori* knowledge of the unknown in form of a probability distribution that assigns higher probability to those values of $U$ we expect to see. Note that the above discussed regularisation methods produce a single estimate of the unknown while the solution to (6.1) is so-called *posterior distribution*, which is the conditional probability distribution of $U$ given a measurement $f$. This distribution can then be used to obtain estimates that are most likely in some sense. The great advance of the method is, however, that it automatically delivers a quantification of uncertainty, obtained by assessing the spread of the posterior distribution.

We recall the Bayes' formula that states

$$\mathbb{P}(u \in A \,|\, f \in B) = \frac{\mathbb{P}(f \in B \,|\, u \in A)\mathbb{P}(u \in A)}{\mathbb{P}(f \in B)},$$

where $A$ and $B$ are some measurable sets. We would like to solve an inverse problem "approximate $U$ when a measurement $f$ is given", that is, we would like to condition $u \in A$ with a single realisation of $F$. To do this we need to we start with some modern probability theory.

## 6.1 A brief introduction to probability theory

A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega$ is the sample space, $\mathcal{F}$ the $\sigma$-algebra of events and $\mathbb{P}$ the probability measure. A measure is called $\sigma$-finite if $\Omega$ is a countable union of measurable sets with finite measure. Lebesgue measure on $\mathbb{R}^d$ is an example of a

$\sigma$-finite measure. One intuitive way of thinking $\sigma$-algebras in probability theory is that they describe information. The $\sigma$-algebra contains the subsets representing the events for which we can decide, after the observation, whether they happened or not. Hence $\mathcal{F}$ represents all the information we can get from an experiment in $(\Omega, \mathcal{F}, \mathbb{P})$ while a sub-$\sigma$-algebra $\mathcal{G} \subset \mathcal{F}$ represents partial information.

Let $(X, \mathcal{B}(X))$ be a measurable space, with $\mathcal{B}(X)$ denoting the Borel $\sigma$-algebra generated by the open sets. We call a measurable mapping $U : \Omega \to X$ a random variable. The random variable $U$ induces the following probability measure on $X$

$$\mu(A) = \mathbb{P}(U^{-1}(A)) = \mathbb{P}(\omega \in \Omega : U(\omega) \in A), \quad A \in \mathcal{B}(X).$$

The measure $\mu$ is called the probability distribution of $U$ and we will denote $U \sim \mu$.

Let $\mu$ and $\nu$ be two measures on the same measure space. Then $\mu$ is absolutely continuous with respect to (dominated by) $\nu$ if $\nu(A) = 0$ implies that $\mu(A) = 0$. We denote this by $\mu \ll \nu$. Measures $\mu$ and $\nu$ are said to be equivalent if $\mu \ll \nu$ and $\nu \ll \mu$. If $\mu$ and $\nu$ are supported on disjoint sets they are called mutually singular.

**Theorem 6.1.1** (Radon-Nikodym Theorem)**.** *Let $\mu$ and $\nu$ be two measures on the same measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and $\nu$ is $\sigma$-finite then there exists a unique function $g \in L^1_\nu$ such that for any measurable set $A \in \mathcal{F}$,*

$$\mu(A) = \int_A g \, d\nu.$$

The unique $g \in L^1_\nu$ in the above theorem is called the Radon-Nikodym derivative of $\mu$ with respect to $\nu$ and is denoted by $\frac{d\mu}{d\nu}$. The following example shows how Radon-Nikodym Theorem can be used to define probability density for a measure on a finite space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

**Example 6.1.2.** Let $\mu$ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mu \ll \nu_L$, where $\nu_L$ is the standard Lebesgue measure on $\mathbb{R}^d$. Since $\nu_L$ is $\sigma$-finite we can use Theorem 6.1.1 and conclude that there exists such $g \in L^1(\mathbb{R}^d)$ that, for any $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu(A) = \int_A g(t) dt.$$

The function $g$ is called the probability density of $U \sim \mu$.

The $\sigma$-algebras we use are often generated by random variables. If $U : \Omega \to X$ then $\sigma(U)$ denotes the smallest $\sigma$-algebra containing preimages $U^{-1}(A)$ of measurable sets $A \in \mathcal{B}(X)$. Observing the value of $U$ corresponds of knowing, with every $A \in \mathcal{B}(X)$, whether $U(\omega) = u \in A$. Note that $\sigma(U) \subset \mathcal{F}$ where, according to the information interpretation, $\mathcal{F}$ represents "full information" (all events on our probability space).

**Definition 6.1.3.** *Let $\mathcal{G} \subset \mathcal{F}$ be a sub-$\sigma$-algebra. We call a $\mathcal{G}$-measurable function $V : \Omega \to X$ a conditional expectation of $U : \Omega \to X$ with respect to $\mathcal{G}$ if*

$$\int_G U d\mathbb{P} = \int_G V d\mathbb{P}$$

*for all $G \in \mathcal{G}$ and write $\mathbb{E}(U \,|\, \mathcal{G}) = V$.*

Note that the measurability with respect to $\mathcal{G}$ is a stronger assumption than measurability with respect to $\mathcal{F}$ since there are fewer choices for the preimages of $V$. Even though the definition of $\mathbb{E}(U \,|\, \mathcal{G})$ resembles that of $\mathbb{E}(U \,|\, G)$ for an event $G$ these are very different objects. The first is a $\mathcal{G}$-measurable function $\Omega \to X$ while the second is an element in $X$.

We can also consider conditional expectation of the form $\mathbb{E}(g(U) \,|\, \mathcal{G})$ which leads us to conditional probability.

**Definition 6.1.4.** *Let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$. The conditional probability for $A \in \mathcal{B}(X)$, given $\mathcal{G}$ is defined by*

$$\mathbb{P}(A \,|\, \mathcal{G}) = \mathbb{E}(\mathbb{1}_A \,|\, \mathcal{G}).$$

It is tempting to try to interpret the map $A \to \mathbb{P}(A \,|\, \mathcal{G})(\omega)$ as a probability measure for a fixed $\omega \in \Omega$. However $\mathbb{P}(A \,|\, \mathcal{G})$ is defined only up to $\mathbb{P}$ almost everywhere.

**Definition 6.1.5.** *A family of probability distributions $(\mu(\cdot, \omega))_{\omega \in \Omega}$ on $(X, \mathcal{B}(X))$ is called a regular conditional distribution of $U$, given $\mathcal{G} \subset \mathcal{F}$, if*

$$\mu(A, \cdot) = \mathbb{E}(\mathbb{1}_A(U) \,|\, \mathcal{G}) \ a.s$$

*for every $A \in \mathcal{B}(X)$.*

**Theorem 6.1.6.** *Let $U : \Omega \to X$ be a random variable and $\mathcal{G} \in \mathcal{F}$ a sub-$\sigma$-algebra. Then there exists a regular conditional distribution $(\mu(\cdot, \omega))_{\omega \in \Omega}$ of $U$ given $\mathcal{G}$.*

Let $\sigma(F) \subset \mathcal{F}$ be the $\sigma$-algebra generated by a random measurement $F$. We can then use the regular conditional probability measure

$$\pi_{post}(A, F(\omega)) = \mathbb{E}(\mathbb{1}_A(U) \,|\, \sigma(F))(\omega)$$

as a posterior measure and identify this with $\pi_{post}(A, f) = \pi_{prior}(A \,|\, f)$.

For further information see e.g. [23].

## 6.2 Bayes' formula

We can now return to the problem of "approximate $U$ given a measurement $f$" using a posterior distribution that is a regular conditional distribution. We assume that $U$ follows a prior $\Pi$ with Lebesgue density $\pi(u)$. The noise $N$ is assumed to be independent of $U$ and distributed according to $P_0$ with Lebesgue density $\rho(n)$. Then $F \,|\, u$ can be found by simply shifting $P_0$ by $Au$ to measure $P_u$, which has Lebesgue density $\rho_u(f) = \rho(f - Au)$. It follows that $(U, F) \in \mathbb{R}^d \times \mathbb{R}^k$ is a random variable with Lebesgue density $\nu(u, f) = \rho(f - Au)\pi(u)$.

**Theorem 6.2.1** (Bayes' Theorem)**.** *Assume that*

$$Z(f) = \int_{\mathbb{R}^d} \rho(f - Au)\pi(u)du > 0.$$

*Then $U \,|\, f$ is a random variable with Lebesgue density $\pi^f(u)$ given by*

$$\pi^f(u) = \pi(u \,|\, f) = \frac{1}{Z(f)}\rho(f - Au)\pi(u).$$

Let us take a closer look to what the above theorem means in our inverse problems settings.

i) $\pi(u)$ is called *prior density*. The prior should be independent of the measurement and assign higher probability to those values of $u$ we expect to see.

ii) $\rho(f - Au)$ is the *likelihood* which measures the data misfit.

iii) $\pi^f(u)$ is called *posterior density* and it gives a solution to the inverse problem (6.1) by updating the prior with a given measurement.

iv) $Z(f)$ is the probability of the measurement and plays the role of a normalising constant.

v) We define

$$\Phi(u; f) = -\log \rho(f - Au)$$

and call $\Phi$ *potential*.

vi) Let $\Pi^f$ and $\Pi$ be measures on $\mathbb{R}^d$ with densities $\pi^f$ and $\pi$ respectively. Then Theorem 6.2.1 can be rewritten as

$$\frac{d\Pi^f}{d\Pi}(u) = \frac{1}{Z(f)} \exp(-\Phi(u; f)),$$
$$Z(f) = \int_{\mathbb{R}^d} \exp(-\Phi(u; f)) d\Pi(u).$$

Note that this means the posterior is absolutely continuous with respect to the prior and the Radon-Nikodym derivative is proportional to the likelihood.

When stated as in vi) the formula has a natural generalisation to infinite dimensions where there are no densities $\rho$ and $\pi$ with respect to Lebesgue measure (since there is no analogue of Lebesgue measure on infinite dimensional spaces) but where $\Pi^f$ has a Radon-Nikodym derivative with respect to $\Pi$.

**Remark 6.2.2.** In Example 6.1.2 we defined density $g$ of a measure $\mu$ in $\mathbb{R}^d$, which is absolutely continuous with respect to Lebesgue measure $\nu_L$. Strictly speaking $g(x) = g_L(x) = \frac{d\mu}{d\nu_L}(x)$ is a probability density function with respect to Lebesgue measure.

It is also possible to find the density of $\mu$ with respect to a Gaussian measure. Let $\mu_0 \sim \mathcal{N}(0, I)$ denote the standard Gaussian measure in $\mathbb{R}^d$. Then

$$\mu_0(dx) = \frac{1}{(2\pi)^{d/2}} \exp\left( -\frac{1}{2}|x|^2 \right) dx.$$

Thus the density of $\mu$ with respect to $\mu_0$ is

$$g_G(x) = (2\pi)^{d/2} \exp\left(\frac{1}{2}|x|^2\right) g_L(x).$$

We then have identities

$$\mu(A) = \int_A g_G(x)\mu_0(dx) \quad \text{and} \quad \frac{d\mu}{d\mu_0}(x) = g_G(x). \tag{6.2}$$

Note that unlike Lebesgue measure infinite-dimensional Gaussian measure is well-defined (we return to this later). Many measures have a Radon–Nikodym derivative with respect to an infinite-dimensional Gaussian measure and hence formulation (6.2) can be generalised to infinite-dimensional settings while the Lebesgue density can not.

**Example 6.2.3.** We start by studying the case $U \in \mathbb{R}$ and $F \in \mathbb{R}^k$, $k \geqslant 1$. The measurement is defined by

$$F = AU + N,$$

where $A \in \mathbb{R}^k \setminus \{0\}$ and $N \sim \mathcal{N}(0, \delta^2 I)$. We model the unknown $U$ by a Gaussian measure $\mathcal{N}(0, 1)$. Then

$$\pi^f(u) \propto \exp\bigg( -\frac{1}{2\delta^2}\|f - Au\|^2 - \frac{1}{2}|u|^2 \bigg).$$

The notation $h \propto g$ means that functions $h$ and $g$ coincide up to a constant, i.e., there is some $c > 0$ such that $h = cg$. The posterior is Gaussian and its mean and covariance, which can be found by completing the square, are given by

$$\theta_\delta = \frac{\langle A, f\rangle}{\delta^2 + \|A\|^2} \quad \text{and} \quad \sigma_\delta^2 = \frac{\delta^2}{\delta^2 + \|A\|^2}.$$

When the noise tends to zero we see that

$$\overline{\theta} = \lim_{\delta\to 0} \theta_\delta = \frac{\langle A, f_0\rangle}{\|A\|^2} \quad \text{and} \quad \overline{\sigma}^2 = \lim_{\delta^2\to 0} \sigma_\delta^2 = 0.$$

The point $\overline{\theta}$ is the least-square solution for the linear equation $f_0 = Au$. We see that the prior plays no role on the limit of zero observational noise.

Next we study the case $U \in \mathbb{R}^d$, $d \geqslant 2$, and $F \in \mathbb{R}$. The measurement is given by

$$F = \langle A, U\rangle + N,$$

with some $A \in \mathbb{R}^d \setminus \{0\}$. We assume that $N \sim \mathcal{N}(0, \delta^2)$ and $U \sim \mathcal{N}(0, \Sigma_0)$. Then

$$\pi^f(u) \propto \exp\bigg( -\frac{1}{2\delta^2}|f - \langle A, u\rangle|^2 - \frac{1}{2}\langle u, \Sigma_0^{-1}u\rangle \bigg).$$

We known that, as an exponential of a quadratic form, the posterior is a Gaussian measure with mean and covariance

$$\theta_\delta = \frac{f\Sigma_0 A}{\delta^2 + \langle A, \Sigma_0 A\rangle} \quad \text{and} \quad \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)(\Sigma_0 A)^*}{\delta^2 + \langle A, \Sigma_0 A\rangle}.$$

When the noise tends to zero we get

$$\overline{\theta} = \lim_{\delta\to 0} \theta_\delta = \frac{f_0\Sigma_0 A}{\langle A, \Sigma_0 A\rangle} \quad \text{and} \quad \overline{\Sigma} = \lim_{\delta\to 0} \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)(\Sigma_0 A)^*}{\langle A, \Sigma_0 A\rangle}.$$

We note that $\langle\overline{\theta}, A\rangle = f_0$ and $\overline{\Sigma}A = 0$. That is, when the observational noise decreases knowledge of $u$ in the direction of $A$ becomes certain. However, the uncertainty remains in directions not aligned with $A$. The magnitude of this uncertainty is determined by interaction between the properties of the prior and forward operator $A$. We see that in the underdetermined case the prior plays an important role even when the observational noise disappears.

**Definition 6.2.4.** *Let $\mu_n$, $n \in \mathbb{N}$, and $\mu$ be two probability measures on $(X, \mathcal{B}(X))$. We say that $\mu_n$ converges weakly to $\mu$ if, for all bounded and continuous functions $g$, it holds that*

$$\lim_{n\to\infty} \int_X g(x) d\mu_n(x) = \int_X g(x) d\mu(x).$$

*If this is the case, we write $\mu_n \rightharpoonup \mu$.*

**Lemma 6.2.5.** *Let $\mu_n = \mathcal{N}(\theta_n, \Sigma_n)$ and $\mu = \mathcal{N}(\theta, \Sigma)$ on $\mathbb{R}^d$. If $\theta_n \to \theta$ and $\Sigma_n \to \Sigma$, as $n \to \infty$, then $\mu_n \rightharpoonup \mu$.*

**Example 6.2.6.** Let us return to the deblurring Example 1.2.1. In real life we only observe the signal $f$ at finite number of observation points on a finite interval

$$f_i = f(t_i) = \int_0^1 a(t_i - s)u(s)ds + n(t_i), \quad 1 \leqslant i \leqslant k,$$

where we assume $a$ to be of the form

$$a(t-s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(t-s)^2\Big).$$

We will also discretise the unknown $u$ on the same mesh and approximate the integral as

$$\int_0^1 a(t_i - s)u(s)ds \approx \sum_{j=1}^k \frac{1}{k} a(t_i - s_j)u(s_j) = \sum_{j=1}^k a_{ij}u_j,$$

where we have denoted $s_j = \frac{j-1}{k-1}$, $u_j = u(s_j)$ and $a_{ij} = \frac{1}{k}a(t_i - s_j)$.

We have now discrete model $f = Au + n$, where $f, u, n \in \mathbb{R}^k$. To employ the Bayesian approach we will consider the stochastic model

$$F = AU + N,$$

where $F, U$ and $N$ are treated as random variables. We assume that $N$ is Gaussian noise with variance $\delta^2 I$,

$$N \sim \mathcal{N}(0, \delta^2 I), \quad \rho(n) \propto \exp\Big(-\frac{1}{2\delta^2}\|n\|^2\Big).$$

Then the likelihood density is given as

$$\rho_u(f) = \rho(f - Au) \propto \exp\Big(-\frac{1}{2\delta^2}\|f - Au\|^2\Big).$$

Next we have to choose a prior for the unknown. Assume that we know that $u(0) = u(1) = 0$ and $u$ is quite smooth, that is, the value of $u(t)$ in a point is more or less the same as its neighbour. We will then model the unknown as

$$u_j = \frac{1}{2}(u_{j-1} + u_{j+1}) + W_j$$

where the innovation term $W_j$ follows Gaussian distribution $\mathcal{N}(0, \gamma^2)$. The variance $\gamma^2$ determines how much the reconstructed function $u$ departs from the smoothness model $u_j = \frac{1}{2}(u_{j-1} + u_{j+1})$. We can then write in matrix form

$$Lu = W, \quad \text{where} \quad L = \frac{1}{2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

Therefore the prior can be written as

$$\pi(u) \propto \exp\Big( -\frac{1}{2\gamma^2}\|Lu\|^2\Big).$$

Using the Bayes' formula we get the posterior distribution

$$\pi^f(u) \propto \exp\Big( -\frac{1}{2\delta^2}\|f - Au\|^2 - \frac{1}{2\gamma^2}\|Lu\|^2\Big).$$

In the next Section we will see how to extract useful information from the posterior distribution.

## 6.3 Estimators

The dimension of the inverse problem can be large and consequently the posterior distribution lives in a high dimensional space which makes its visualisation difficult. However, we can calculate different point estimators and spread or region estimators. The point estimators approximate the most probable value of the unknown given the data and the prior. The spread estimators give a region that contain the unknown with some high probability.

One of the most used statistical estimators is the maximum a posterior estimate (MAP), which is the mode of the posterior distribution. That is, given the posterior density $\pi^f(u)$ the MAP estimate $u_{MAP}$ satisfies

$$u_{MAP} = \arg\max_{u\in\mathbb{R}^d} \pi^f(u),$$

if such maximiser exists. Note that MAP estimator may not be unique.

Another widely used point estimate is the conditional mean (CM) of the unknown $U$ given the data $f$, which is defined by

$$u_{CM} = \mathbb{E}(U \,|\, f) = \int_{\mathbb{R}^d} u\pi(u \,|\, f)du,$$

provided that the integral converges. The main problem with CM estimation is that solving the integral in high-dimensional space is often very difficult.

As an example of spread estimate we can consider Bayesian credible sets. A level $1-\alpha$ credible set $\mathcal{C}_\alpha$, with some small $\alpha \in (0,1)$, is defined as

$$\Pi(\mathcal{C}_\alpha \,|\, f) = \int_{\mathcal{C}_\alpha} \pi(u \,|\, f)du = 1 - \alpha.$$

Hence a credible set $\mathcal{C}_\alpha$ is a region that contains a large fraction of the posterior mass.

Another way of quantifying uncertainty is to consider problem $F^\dagger = Au^\dagger + N$, where $u^\dagger$ is though to be a deterministic 'true' unknown. We would then like to find random sets $\mathcal{C}_\alpha$ that frequently contain the 'true' unknown $u^\dagger$, that is, $\mathbb{P}(u^\dagger \in \mathcal{C}_\alpha) = 1 - \alpha$. The set $\mathcal{C}_\alpha$ is called a frequentist confidence region of level $1 - \alpha$.

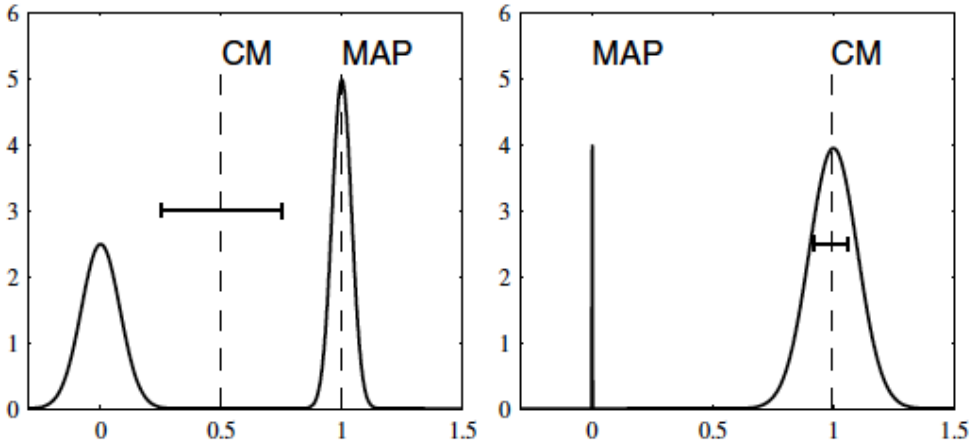



Figure 6.1: We can not say that one point estimator is better than the other in all applications.

**Example 6.3.1.** Let us return to Example 6.2.6 where we got the posterior distribution

$$\pi^f(u) \propto \exp\Big( -\frac{1}{2\delta^2}\|f - Au\|^2 - \frac{1}{2\gamma^2}\|Lu\|^2\Big).$$

Since the posterior distribution is also Gaussian we know that the MAP and CM estimators coincides and we have an estimator

$$u^\delta_{MAP} = \arg\max_{u\in\mathbb{R}^k} \pi(u \,|\, f) = \arg\min_{u\in\mathbb{R}^k} \Big\{ \frac{1}{2\delta^2}\|f - Au\|^2 + \frac{1}{2\gamma^2}\|Lu\|^2\Big\}.$$

Notice that $u_{MAP}$ is of the same form as a Tikhonov estimator.

Completing the square we can write the posterior in form

$$\pi^f(u) \propto \exp\Big( -\frac{1}{2}\big\|u - \frac{1}{\delta^2}\Gamma^{-1}A^\top f\big\|^2_\Gamma\Big),$$

where we have used the weighted norm $\|\cdot\|_\Gamma = \|\Gamma^{\frac{1}{2}}\cdot\|$ with $\Gamma = \frac{1}{\delta^2}A^\top A + \frac{1}{\gamma^2}L^\top L$. Hence we see that the MAP estimator is given by

$$u_{MAP} = \frac{1}{\delta^2}\Gamma^{-1}A^\top f = \Big(A^\top A + \frac{\delta^2}{\gamma^2}L^\top L\Big)^{-1}A^\top f$$

and the posterior covariance is $\Sigma = \Gamma^{-1}$.

## 6.4 Prior models

Constructing a good prior density is one of the most challenging parts of solving a Bayesian inverse problem. The main problem is transforming our qualitative information into a quantitative form that can be coded as a prior density. The prior probability distribution should be concentrated on those values of $U$ we expect to see and assigns a clearly higher probability to them than to the unexpected ones.

### 6.4.1 Gaussian prior

Gaussian probability densities are the most used priors in statistical inverse problems. They are easy to construct and form a versatile class of densities. They also often lead to explicit estimators. Due to the central limit theorem the Gaussian densities are often good approximation to inherently non-Gaussian distributions when the observation is based on a large number of mutually independent random events. This is also the reason why the noise is often assumed to be Gaussian.

**Definition 6.4.1.** *Let $\theta \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix. A Gaussian $d$-variate random variable $U$ with mean $\theta$ and covariance $\Sigma$ is a random variable with the probability density*

$$\pi(u) = \frac{1}{(2\pi|\Sigma|)^{d/2}} \exp\Big( -\frac{1}{2}(u-\theta)^\top \Sigma^{-1}(u-\theta)\Big),$$

*where $|\Sigma| = \det(\Sigma)$. We then denote $U \sim \mathcal{N}(\theta, \Sigma)$.*

The Gaussian distribution is completely characterised by its mean and covariance. Notice that the expression $(u-\theta)^\top\Sigma^{-1}(u-\theta)$ can also be written in form $\|\Sigma^{-1/2}(u-\theta)\|_2^2$, since due to our assumptions on $\Sigma$ the inverse square root $\Sigma^{-1/2}$ is well-defined.

If we consider linear inverse problems and assume Gaussian prior and Gaussian noise model the posteriori distribution is of the form $c\cdot\exp(-G(u))$, where $G$ can be rewritten as a sum of a quadratic form and constant term in order to show that the posterior is Gaussian. This method is called completing the square. In order to analyse the Gaussian posterior, we need some machinery from linear algebra.

**Definition 6.4.2.** *Let*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \in \mathbb{R}^{d\times d}$$

*be a positive definite symmetric matrix, where $\Sigma_{11} \in \mathbb{R}^{n\times n}$, $\Sigma_{22} \in \mathbb{R}^{(d-n)\times(d-n)}$, $n < d$ and $\Sigma_{21} = \Sigma_{12}^\top$. We define the Schur complements $\widetilde{\Sigma}_{jj}$ of $\Sigma_{jj}$, $j = 1, 2$, as*

$$\widetilde{\Sigma}_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \quad \textit{and} \quad \widetilde{\Sigma}_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The positive definiteness of $\Sigma$ implies that $\Sigma_{jj}$, $j = 1, 2$, are also positive definite and hence the Schur complements are well defined. The following matrix inversion lemma is useful when calculating the conditional covariance.

**Lemma 6.4.3.** *Let $\Sigma$ be a matrix satisfying the assumptions of Definition 6.4.2. Then the Schur complements $\widetilde{\Sigma}_{jj}$ are invertible matrices and*

$$\Sigma^{-1} = \begin{bmatrix} \widetilde{\Sigma}_{22}^{-1} & -\widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\widetilde{\Sigma}_{11}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \widetilde{\Sigma}_{11}^{-1} \end{bmatrix}.$$

Let $U \sim \mathcal{N}(\theta_u, \Sigma_u)$ and $N \sim \mathcal{N}(0, \Sigma_n)$ with $U$ and $N$ independent. The distribution of $F = AU + N$ is Gaussian with $\theta_f = \mathbb{E}(f) = A\theta_u$ and

$$\Sigma_f = \mathbb{E}\big((F - \theta_f)(F - \theta_f)^\top\big) = A\Sigma_u A^\top + \Sigma_n.$$

We can also calculate

$$\mathbb{E}\big((U - \theta_u)(F - \theta_f)^\top\big) = \Sigma_u A^\top$$

The joint distribution of $U$ and $F$ then has a covariance

$$\operatorname{Cov}\begin{bmatrix} U \\ F \end{bmatrix} = \begin{bmatrix} \Sigma_u & \Sigma_u A^\top \\ A\Sigma_u & A\Sigma_u A^\top + \Sigma_n \end{bmatrix}$$

Hence the joint probability density of $U$ and $F$ is given by

$$\nu(u, f) \propto \exp\bigg( -\frac{1}{2} \begin{bmatrix} u - \theta_u \\ f - \theta_f \end{bmatrix}^\top \begin{bmatrix} \Sigma_u & \Sigma_u A^\top \\ A\Sigma_u & A\Sigma_u A^\top + \Sigma_n \end{bmatrix}^{-1} \begin{bmatrix} u - \theta_u \\ f - \theta_f \end{bmatrix} \bigg).$$

**Theorem 6.4.4.** *Assume that $U : \Omega \to \mathbb{R}^d$ and $N : \Omega \to \mathbb{R}^k$ are mutually independent Gaussian random variables*

$$U \sim \mathcal{N}(\theta_u, \Sigma_u), \quad N \sim \mathcal{N}(0, \Sigma_n),$$

*where $\Sigma_u \in \mathbb{R}^{d\times d}$ and $\Sigma_n \in \mathbb{R}^{k\times k}$ are positive definite. The noisy measurement is given by $F = AU + N$, where $A \in \mathbb{R}^{k\times d}$ is a known matrix. Then the posterior probability density of $U$ given the measurement $f$ is*

$$\pi(u \,|\, f) \propto \exp\Big( -\frac{1}{2}(u - \overline{u})^\top \Sigma^{-1}(u - \overline{u})\Big),$$

*where*

$$\overline{u} = \theta_u + \Sigma_u A^\top (A\Sigma_u A^\top + \Sigma_n)^{-1}(f - A\theta_u)$$

*and*

$$\Sigma = \Sigma_u - \Sigma_u A^\top (A\Sigma_u A^\top + \Sigma_n)^{-1} A\Sigma_u.$$

*Proof.* By shifting the coordinate origin to $[\theta_u, \theta_f]$ we may assume that $\theta_u = \theta_f = 0$. By Bayes' formula we have $\pi(u \,|\, f) \propto \nu(u, f)$ and hence we will consider the joint density as a function of $U$. We denote the components of $\operatorname{Cov}([U\ F]^\top)$ by $\Sigma_{ij}$, $i, j = 1, 2$. Then, by Lemma 6.4.3 and the fact that $\Sigma_{22}^{-1}\Sigma_{21}\widetilde{\Sigma}_{22}^{-1} = \widetilde{\Sigma}_{11}^{-1}\Sigma_{21}\Sigma_{11}^{-1}$ (the covariance is symmetric), we have

$$\begin{aligned} \nu(u, f) &\propto \exp\Big( -\frac{1}{2}(u^\top \widetilde{\Sigma}_{22}^{-1} u - 2u^\top \widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} f + f^\top \widetilde{\Sigma}_{11}^{-1} f)\Big) \\ &= \exp\Big( -\frac{1}{2}(u - \Sigma_{12}\Sigma_{22}^{-1} f)^\top \widetilde{\Sigma}_{22}^{-1}(u - \Sigma_{12}\Sigma_{22}^{-1} f) + c\Big), \end{aligned}$$

where

$$c = f^\top(\widetilde{\Sigma}_{11}^{-1} - \Sigma_{22}^{-1}\Sigma_{21}\widetilde{\Sigma}_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1}) f$$

is a constant independent of $u$ and can hence be factored out of the density. □

Note that the posterior covariance is independent of the prior mean $\theta$ (and mean of the noise even if that would be non-zero). We have a more compact expression for the posterior mean and variance

**Lemma 6.4.5.** *Assume that $F, U, N$ are as in Theorem 6.4.4. We then have*

$$\pi^f(u) \propto \exp\Big( -\frac{1}{2}(u-\overline{u})^\top \Sigma^{-1}(u-\overline{u})\Big),$$

*where*

$$\Sigma = (A^\top \Sigma_n^{-1} A + \Sigma_u^{-1})^{-1}$$

*and*

$$\overline{u} = \Sigma(A^\top \Sigma_n^{-1} f + \Sigma_u^{-1}\theta_u).$$

Proof left as an exercise.

Consider next a problem where the unknown is a two-dimensional pixel image. Let $u \in \mathbb{R}^d$ be the pixel image (which we have arranged as a vector), where a component $u_j$ represents the intensity of the $j^{th}$ pixel. Since we consider images it is natural to add a positivity constraint to our prior. Gaussian white noise density with positivity constraint is

$$\pi(u) \propto \mu_+(u) \exp\left( -\frac{1}{2\alpha^2}\|u\|_2^2\right)$$

where $\mu_+(u) = 1$ if $u_j > 0$ for all $j$, and $\mu_+(u) = 0$ otherwise. We assume that each component is independent of the others and hence the random draws can be performed componentwise. The one-dimensional cumulative distribution function can be defined as

$$\Phi(t) = \frac{1}{\alpha}\sqrt{\frac{2}{\pi}} \int_0^t \exp\Big( -\frac{1}{2\alpha^2}s^2\Big) ds = \mathrm{erf}\Big(\frac{t}{\alpha\sqrt{2}}\Big),$$

where erf stands for the error function

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp(-s^2) ds.$$

The mutually independent components $u_j$ are then drawn by

$$u_j = \Phi^{-1}(t_j) = \mathrm{erf}^{-1}\Big(\frac{t_j}{\alpha\sqrt{2}}\Big),$$

where $t_j$s are drawn randomly from the uniform distribution $\mathcal{U}([0,1])$. The proof that this really produces draws from the prior is left as an exercise.

### 6.4.2 Impulse Prior

We assume again that the unknown is a two-dimensional pixel image. We have prior information is that the image contains small and well localised objects (for example a tumour in X-ray image). We can then use impulse prior model. These priors favour images with low average amplitude with few outliers. One example of such a prior is $\ell^1$ prior. Let $u \in \mathbb{R}^d$

represent the pixel image, where a component $u_j$ represents the intensity of the $j^{th}$ pixel. The $\ell^1$ prior is defined as

$$\pi(u) = \Big(\frac{\alpha}{2}\Big)^d \exp(-\alpha\|u\|_1),$$

where $\alpha > 0$ and $\|\cdot\|_1$ is the $\ell^1$-norm. We can enhance the impulse noise effect by taking a smaller power of the components of $u$, that is, using $\sum |u_j|^p$, $p \in (0,1)$ instead of the $\ell^1$-norm.

Another density that produces few distinctly different pixels with a low-amplitude background is the Cauchy density, which is defined as

$$\pi(u) = \Big(\frac{\alpha}{\pi}\Big)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2 u_j^2}.$$

Let us take a closer look at the $\ell^1$ prior and to what kind of draws it produces. Since we consider images we add a positivity constraint to our prior and write

$$\pi(u) = \alpha^d \mu_+(u) \exp(-\alpha\|u\|_1),$$

where $\mu_+(u) = 1$ if $u_j > 0$ for all $j$, and $\mu_+(u) = 0$ otherwise. The one-dimensional distribution function can be defined as

$$\Phi(t) = \alpha \int_0^t e^{-\alpha s} ds = 1 - e^{-\alpha t}.$$

The mutually independent components $u_j$ are then drawn by

$$u_j = \Phi^{-1}(t_j) = -\frac{1}{\alpha}\log(1-t_j),$$

where $t_j$s are drawn randomly from the uniform distribution $\mathcal{U}([0,1])$. As mentioned before the proof that this really produces draws from the prior is left as an exercise.

Similarly when we draw independent components from the Cauchy distribution with positivity constraint we use the distribution function

$$\Phi(t) = \frac{2\alpha}{\pi} \int_0^t \frac{1}{1+\alpha^2 s^2} ds = \frac{2}{\pi}\arctan(\alpha t)$$

meaning that the inverse cumulative distribution is

$$\Phi^{-1}(t) = \frac{1}{\alpha}\tan\Big(\frac{\pi t}{2}\Big).$$

### 6.4.3 Discontinuities

Assume next that we want to estimate one-dimensional signal $u : [0,1] \to \mathbb{R}$, $u(0) = 0$, from indirect observations. Our prior knowledge is that the signal is usually relatively stable but can have large jumps every now and then. We may also have information on the size of the jumps or the rate of occurrence of the discontinuities. One possible prior is the finite difference approximation of the derivative of $u$ with assumption that the derivative follows

an impulse noise probability distribution. Let us discretise the interval $[0,1]$ by points $t_j = j/N$ and write $u_j = u(t_j)$. Consider the density

$$\pi(u) = \Big(\frac{\alpha}{\pi}\Big)^N \prod_{j=1}^{N} \frac{1}{1+\alpha^2(u_j - u_{j-1})^2}.$$

To draw from the above distribution let us define new random variables

$$x_j = u_j - u_{j-1}, \quad 1 \leqslant j \leqslant N.$$

The probability distribution of these variables is

$$\pi(x) = \Big(\frac{\alpha}{\pi}\Big)^N \prod_{j=1}^{N} \frac{1}{1+\alpha^2 x_j^2},$$

that is, they are independent of each other and can hence be drawn from the one-dimensional Cauchy density. Note that $u = [u_1, \cdots, u_N]^\top \in \mathbb{R}^N$ satisfies $u = Bx$, where $B \in \mathbb{R}^{N\times N}$ is a lower triangular matrix such that $B_{ij} = 1$ for $i \geqslant j$. The idea of the above prior can be generalised to higher dimensions which brings us to total variation prior.

We start by defining the concept of total variation for functions. Let $u : D \to \mathbb{R}$ be a function in $L^1(D)$, $D \subset \mathbb{R}^d$. We define the total variation of $u$, denoted by $\mathrm{TV}(u)$, as

$$\mathrm{TV}(u) = \sup_g \Big\{ \int_D u\nabla\cdot g\,dx \,|\, g = (g_1, \cdots, g_d) \in C_0^1(D, \mathbb{R}^d),\ \|g\|_{L^\infty} \leqslant 1 \Big\}.$$

The test function space $C_0^1(D, \mathbb{R}^d)$ consist of continuously differentiable vector-valued functions on $D$ that vanish at the boundary. A function is said to have bounded variation if $\mathrm{TV}(u) < \infty$. To understand the definition let us consider the following simple example

**Example 6.4.6.** Let $D \subset \mathbb{R}^2$ be an open set and $B \subset D$ be a set bounded by a smooth curve $\partial B = S$, which does not intersect with the boundary of $D$. Let $u : D \to \mathbb{R}$ be 1 when $x \in B$ and zero otherwise. Let $g \in C_0^1(D, \mathbb{R}^2)$ be an arbitrary test function. By the divergence theorem we obtain

$$\int_D u\nabla\cdot g\,dx = \int_B \nabla\cdot g\,dx = \int_{\partial B} \widetilde{n}\cdot g\,dS,$$

where $\widetilde{n}$ is the exterior unit normal vector of $\partial B$. This integral attains its maximum, under the constraint $\|g\|_{L^\infty} \leqslant 1$, if we set $\widetilde{n}\cdot g = 1$ identically. Hence

$$\mathrm{TV}(u) = \mathrm{length}(\partial B).$$

Notice that the weak derivative of $u$ is the Dirac delta of the boundary curve, which cannot be be presented by an integrable function. Therefore, the space of functions with bounded variation differs from the corresponding Sobolev space.

We will next consider two dimensional problem and define a discrete analogue for TV. Let $D \in \mathbb{R}^2$ be bounded and divided in $d$ pixels. We define two pixels as neighbours if they share a common edge. The total variation of the discrete image $u = [u_1, \cdots, u_d]^\top$ is then defined

$$\mathrm{TV}(u) = \sum_{j=1}^{d} V_j(u), \quad V_j(u) = \frac{1}{2}\sum_{i\in\mathcal{N}_j} |u_i - u_j|,$$

where $\mathcal{N}_j$ is the neighbourhood of pixel $u_j$ ($j \notin \mathcal{N}_j$). The discrete total variation density is then given by

$$\pi(u) \propto \exp(-\alpha \mathrm{TV}(u)).$$

The total variation density is concentrated on images that are 'blocky' consisting of blocks with short boundaries and small variation within each block.

The total variation prior is an example of a structural prior. Different structural priors, depending on different neighbourhood systems, can be derived from the theory of Markov random fields. For more prior choices and examples see e.g. [22, Section 3.3]

## 6.5 Sampling methods

An important part of Bayesian inversion techniques is to develop methods for exploring the posterior probability densities. We will next discus a random sampling methods known as the Markov chain Monte Carlo (MCMC) techniques. We saw previously in Section 6.3 that finding a MAP estimate leads to an optimisation problem, whereas the conditional mean requires integration over the space $\mathbb{R}^d$ where the posterior density is defined. Since the dimension of the problem can be large instead of calculating the full integral we want to sample from the posterior and then use these sample points to approximate the integral.

Let $\mu$ denote a probability measure on $\mathbb{R}^d$ and let $g$ be a measurable function integrable over $\mathbb{R}^d$ with respect to $\mu$, that is, $g \in L^1_\mu$. We want to estimate the integral of $g$ with respect to the measure $\mu$. In numerical quadrature methods one defines a set of support points $x_j \in \mathbb{R}^d$, $1 \leqslant j \leqslant N$ and the corresponding weights $w_j$ to get an approximation

$$\int_{\mathbb{R}^d} g(x) d\mu(x) \approx \sum_{j=1}^{N} w_j g(x_j).$$

The above method is designed for computing one-dimensional integrals. To compute integrals in multiple dimensions, we could phrase the integral as repeated one-dimensional integrals by applying Fubini's theorem. However, this approach requires the function evaluations to grow exponentially as the number of dimensions increases which makes it infeasible in high dimensions.

In Monte Carlo integration the support points $x_j$ are generated randomly by drawing from some probability density (ideally determined by $\mu$) and the weights are then determined from the distribution $\mu$. Assume that $x \sim \mu$. If we had a random generator such that repeated realisations of $x$ could be produced we could generate a set of points distributed according to $\mu$. We could then approximate the integral of $g$ by the so called ergodic average,

$$\int_{\mathbb{R}^d} g(x) d\mu(x) = \mathbb{E}\big(g(x)\big) \approx \frac{1}{N} \sum_{j=1}^{N} g(x_j),$$

where $\{x_1, \cdots, x_N\} \subset \mathbb{R}^d$ is a representative collection of samples distributed according to $\mu$.

The MCMC methods are systematic way of generating sample collection so that the above approximation holds. We start with some basic tools from probability theory

**Definition 6.5.1.** *A mapping $P : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ is called a probability transition kernel if*

1. *for each* $B \in \mathcal{B}(\mathbb{R}^d)$ *the mapping* $\mathbb{R}^d \to [0,1]$, $x \mapsto P(x,B)$ *is a measurable function;*
2. *for each* $x \in \mathbb{R}^d$ *the mapping* $\mathcal{B}(\mathbb{R}^d) \to [0,1]$, $B \mapsto P(x,B)$ *is a probability distribution.*

A discrete time stochastic process is an ordered set $\{x_j\}_{j=1}^{\infty}$ of random variables $x_j \in \mathbb{R}^d$. A time homogeneous Markov chain with the transition kernel $P$ is a stochastic process $\{x_j\}_{j=1}^{\infty}$ with the properties

$$\mu_{x_{j+1}}(B_{j+1} \,|\, x_1, \cdots, x_j) = \mu_{x_{j+1}}(B_{j+1} \,|\, x_j) = P(x_j, B_{j+1}).$$

The first equality means that the probability $x_{j+1} \in B_{j+1}$ depends on the past only through the previous state $x_j$. The second equality states that time is homogeneous in the sense that the dependence of consecutive moments does not vary in time since the kernel $P$ does not depend on time $j$.

We define the transition kernel that propagates $k$ steps forward as

$$P^k(x_j, B_{j+k}) = \mu_{x_{j+k}}(B_{j+k} \,|\, x_j) = \int_{\mathbb{R}^d} P(x_{j+k-1}, B_{j+k}) P^{k-1}(x_j, dx_{j+k-1}), \quad k \geqslant 2,$$

where $P^1(x_j, B_{j+1}) = P(x_j, B_{j+1})$. if $\mu_{x_j}$ denotes the probability distribution of $x_j$ the distribution of $x_{j+1}$ is given by

$$\mu_{x_{j+1}}(B_{j+1}) = \mu_{x_j} P(B_{j+1}) = \int_{\mathbb{R}^d} P(x_j, B_{j+1}) d\mu_{x_j}(x_j).$$

We will next introduce few concepts concerning the transition kernels

1. The measure $\mu$ is an *invariant measure* of $P(x_j, B_{j+1})$ if
$$\mu P = \mu.$$
This means that the distribution of the random variable $x_j$ before the time step $j \to j+1$ is the same as the variable $x_{j+1}$ after the step.

2. The transition kernel $P$ is *irreducible* (with respect to a given measure $\mu$) if for each $x \in \mathbb{R}^d$ and $B \in \mathcal{B}(\mathbb{R}^d)$, with $\mu(B) > 0$, there exists an integer $k$ such that $P^k(x,B) > 0$. This means that regardless of the starting point the Markov chain generated by $P$ visits any set of positive measure with positive probability.

3. Let $P$ be irreducible kernel. We say that $P$ is *periodic* if, for some integer $m \geqslant 2$, there is a set of disjoint non-empty sets $\{E_1, \cdots, E_m\} \subset \mathbb{R}^d$ such that $P(x, E_{j+1(\bmod m)}) = 1$ for all $j = 1, \cdots, m$ and all $x \in E_j$. This means that a periodic $P$ generates a Markov chain that remains in a periodic loop for ever. We say that $P$ is an *aperiodic* kernel if it is not periodic.

The following theorem is of crucial importance for MCMC methods.

**Theorem 6.5.2.** *Let* $\mu$ *be a probability measure on* $\mathbb{R}^d$ *and* $\{x_j\}$ *a time homogeneous Markov chain with transition kernel* $P$*. Assume further that* $\mu$ *is an invariant measure of the transition kernel, and that* $P$ *is irreducible and aperiodic. Then for all* $x \in \mathbb{R}^d$*,*

$$\lim_{N\to\infty} P^N(x,B) = \mu(B), \quad \text{for all } B \in \mathcal{B}(\mathbb{R}^d),$$

*and for* $g \in L^1_\mu(\mathbb{R}^d)$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} g(x_j) = \int_{\mathbb{R}^d} g(x) d\mu(x)$$

*almost certainly.*

### 6.5.1 Metropolis-Hastings

Let $\Pi$ denote the target probability measure in $\mathbb{R}^d$. We assume that $\Pi$ is absolutely continuous with respect to Lebesgue measure and has density $\pi(x)$. We want to determine a transition kernel $P(x, B)$ so that $\Pi$ is its invariant measure.

Let $P$ denote any transition kernel. If we start from a point $x \in \mathbb{R}^d$ the kernel either proposes to move to another point $y \in \mathbb{R}^d$ or to stay in $x$. Hence we can split the kernel in two parts,

$$P(x, B) = \int_B K(x, y)dy + r(x)\mathbb{1}_B(x),$$

where $\mathbb{1}_B(x)$ is 1 when $x \in B \in \mathcal{B}(\mathbb{R}^d)$ and zero otherwise. Loosely speaking $K(x, y) \geqslant 0$ describes the probability for moving and $r(x) \geqslant 0$ the probability for staying put.

The condition $P(x, \mathbb{R}^d) = 1$ implies that

$$r(x) = 1 - \int_{\mathbb{R}^d} K(x, y)dy. \tag{6.3}$$

We assume that the $K$ satisfies the *detailed balance condition*

$$\pi(y)K(y, x) = \pi(x)K(x, y), \tag{6.4}$$

for all $x, y \in \mathbb{R}^d$. This guarantees that $\Pi$ is an invariant measure of $P$ since using (6.3) we can then write

$$\begin{aligned}
\Pi P(B) &= \int_{\mathbb{R}^d} \left( \int_B K(x, y)dy + r(x)\mathbb{1}_B(x) \right) \pi(x)dx \\
&= \int_B \left( \int_{\mathbb{R}^d} \pi(x)K(x, y)dx + r(y)\pi(y) \right) dy \\
&= \int_B \pi(y)dy
\end{aligned}$$

Our goal now is to construct a transition kernel that $K$ that satisfies the detailed balance equation 6.4. Let $q : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a given functional with property $\int q(x, y)dy = 1$. The function $q$ is called the *proposal distribution* and it defines a transition kernel

$$Q(x, A) = \int_A q(x, y)dy.$$

If $q$ satisfies the detailed balance condition we can simply choose $K(x, y) = q(x, y)$ and $r(x) = 0$. Otherwise we have to correct the kernel and define

$$K(x, y) = \alpha(x, y)q(x, y), \tag{6.5}$$

where $\alpha$ is a correction term.

Assume that instead of the detailed balance condition we have

$$\pi(y)q(y, x) < \pi(x)q(x, y),$$

for some $x, y \in \mathbb{R}^d$. Our aim is to choose $\alpha$ so that

$$\pi(y)\alpha(y, x)q(y, x) = \pi(x)\alpha(x, y)q(x, y).$$

We can achieve this by setting

$$\alpha(y,x) = 1 \quad \text{and} \quad \alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1.$$

We then see that the kernel $K$ defined in (6.5) satisfies the detailed balance condition (6.4) if we define

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right).$$

This transition kernel is called the Metropolis-Hastings kernel.

We can implement the method derived above using an algorithm that is carried out through the following steps;

1. Pick initial value $x_1 \in \mathbb{R}^d$ and set $j = 1$.

2. Draw $y \in \mathbb{R}^d$ from the proposal kernel $q(x_j, y)$ and calculate the acceptance ratio

$$\alpha(x_j,y) = \min\left(1, \frac{\pi(y)q(y,x_j)}{\pi(x_j)q(x_j,y)}\right).$$

3. Draw $t \in [0,1]$ from uniform probability density.

4. If $t \leqslant \alpha(x_j, y)$, set $x_{j+1} = y$, otherwise $x_{j+1} = x_j$. Increase $j \to j+1$ and go to step 2. until $j = J$, the desired sample size.

Note that if the candidate generating the kernel is symmetric $q(x,y) = q(y,x)$ for all $x, y \in \mathbb{R}^d$ then the acceptance ration simplifies to

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

This means that we accept immediately moves towards higher probability and sometimes also moves that take us to lower probability.

**Example 6.5.3.** Consider a two-dimensional density

$$\pi(x) \propto \exp\Big( -10(x_1^2 - x_2)^2 - \Big(x_2 - \frac{1}{4}\Big)^4\Big).$$

In what follows, we assume to have random number generators for $W \sim \mathcal{N}(0,1)$ and $T \sim \mathcal{U}([0,1])$ at our disposal (in Matlab the command randn and rand respectively).

We construct Metropolis–Hastings sequence using the random walk proposal distribution. We define

$$q(x,y) = \exp\Big( -\frac{1}{2\gamma^2}\|x-y\|^2\Big).$$

This means that we assume that the scaled random step from $x$ to $y$ is distributed as white noise $W = (y-x)/\gamma \sim \mathcal{N}(0, I)$. Using the above proposal distribution we get the following updating algorithm;

**Algorithm 1:** Simple Metropolis–Hastings update scheme

Pick initial value $x_1$. Set $x = x_1$;
**for** $j = 2 : J$ **do**
  Calculate $\pi(x)$;
  Draw $W \sim \mathcal{N}(0, I)$, set $y = x + \gamma W$;
  Calculate $\pi(y)$;
  Calculate $\alpha(x, y) = \min(1, \pi(y)/\pi(x))$;
  Draw $T \sim \mathcal{U}([0, 1])$;
  **if** $t < \alpha(x, y)$ **then**
    Accept: Set $x = y$, $x_j = x$;
    **else**
      Reject: Set $x_j = x$
    **end if**
  **end if**
**end for**

### 6.5.2 Single component Gibbs sampler

Gibbs sampling is used to sample multivariate distributions. The proposal kernel is defined using the density $\pi$ to sample each component $x_i$ of the vector $x = (x_1, \cdots, x_d)$ from the distribution of that component conditioned on all other components sampled so far.

If $x$ is a $d$-variate random variable with the probability density $\pi$ the probability density of the $i^{th}$ component $x_i$ conditioned on all $x_j$, for which $i \neq j$, is given by

$$\pi(x_i \,|\, x_{-i}) = C_i \pi(x)$$

where $x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d)$ and $C_i$ is a normalisation constant. We can then define a transition kernel $K$ as

$$K(x, y) = \prod_{i=1}^{d} \pi(y_i \,|\, y_1, \cdots, y_{i-1}, x_{i+1}, \cdots, x_d)$$

and set $r(x) = 0$. This kernel does not usually satisfy the detailed balance condition but it satisfies a weaker but sufficient balance condition $\int_{\mathbb{R}^d} \pi(y) K(y, x) dx = \int_{\mathbb{R}^d} \pi(x) K(x, y) dx$.

The steps needed for implementing the algorithm can be summarised as follows;

1. Pick an initial value $x^1 \in \mathbb{R}^d$ and set $j = 1$.

2. Set $x = x^j$. For $1 \leqslant i \leqslant d$, draw $y_i \in \mathbb{R}$ from the one-dimensional distribution $\pi(y_i \,|\, y_1, \cdots, y_{i-1}, x_{i+1}, \cdots, x_d)$.

3. Set $x^{j+1} = y$. Increase $j \to j + 1$ and repeat from step 2. until $j$ reaches the desired sample size $J$.

**Example 6.5.4.** We want to sample from the two-dimensional distribution

$$\pi(x) = \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \ \rho > 0.$$

In order to sample from this distribution using Gibbs sampler, we need to calculate the conditional distributions for directions $x_1, x_2$. We see that

$$\pi(x_1^j \,|\, x_2^{(j-1)}) = \mathcal{N}(\rho x_2^{(j-1)}, \sqrt{1-\rho^2}) \quad \text{and} \quad \pi(x_2^j \,|\, x_1^j) = \mathcal{N}(\rho x_1^j, \sqrt{1-\rho^2}).$$

We can write the algorithm as follows;

**Algorithm 2:** Simple Gibbs sampler update scheme

Pick initial value $x^1$. Set $x = x^1$;
**for** $j = 2 : J$ **do**
  Draw $X_1 \sim \mathcal{N}(\rho x_2 \sqrt{1-\rho^2})$;
  Draw $X_2 \sim \mathcal{N}(\rho x_1, \sqrt{1-\rho^2})$;
  Set $x^j = x$
**end for**

## 6.6 Hierarchical models

The prior densities we use depend on some parameters, such as variance or mean. So far we have assumed that these parameters are known. However, we often do not know how to choose them. If a parameter is not know, it can be estimated as a part of the statistical inference problem based on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter $\theta$, which is assumed to be unknown. We then write the prior as a conditional density

$$\pi(u \,|\, \theta).$$

We model the unknown $\theta$ with a hyperprior $\pi_h(\theta)$ and write the joint distribution of $U$ and $\Theta$ as

$$\pi(u, \theta) = \pi(u \,|\, \theta)\pi_h(\theta).$$

Assuming we have a likelihood model $\rho(f \,|\, u, \theta)$ for the measurement $F$, we get the posterior density for $U$ and $\Theta$ given $f$ using the Bayes' formula

$$\pi(u, \theta \,|\, f) \propto \rho(f \,|\, u, \theta)\pi(u, \theta) = \rho(f \,|\, u, \theta)\pi(u \,|\, \theta)\pi_h(\theta).$$

The hyperprior density $\pi_h$ may depend on some hyperparameter $\theta_0$. The main reason for the use of a hyperprior model is that the construction of the posterior is assumed to be more robust with respect to fixing a value for the hyperparameter $\theta_0$ than fixing a value for $\theta$. Sometimes we might want to treat also $\theta_0$ as a random variable with a respective probability density. We can then write $\pi(\theta \,|\, \theta_0)$ which leads to nested hypermodels.

**Example 6.6.1.** We return to the deblurring example 6.2.6, where we assumed prior $\pi(u) \propto \exp\big(-\|Lu\|^2/2\theta\big)$, with $L$ being the second order finite difference matrix and $\theta = \gamma^2$ was assumed to be known.

We will next assume that we do not know the value of $\theta$. The prior for $U$ given $\theta$ is

$$\pi(u \,|\, \theta) = C_\theta \exp\Big(-\frac{1}{2\theta}\|Lu\|^2\Big).$$

The integral of a density is 1 and hence, with the change of variables $u = \sqrt{\theta}z$, $du = \theta^{d/2}dz$, we see that

$$1 = C_\theta \int_{\mathbb{R}^d} \exp\Big( -\frac{1}{2\theta}\|Lu\|^2\Big)du = \theta^{d/2} C_\theta \int_{\mathbb{R}^d} \exp\Big( -\frac{1}{2}\|Lz\|^2\Big)dz.$$

The last integral does not depend on $\theta$ so we deduce that $C_\theta \propto \theta^{-d/2}$ and write

$$\pi(u\,|\,\theta) \propto \exp\Big( -\frac{1}{2\theta}\|Lu\|^2 - \frac{d}{2}\log(\theta)\Big).$$

Since $\theta$ is not known we will treat it as a random variable. Any information concerning $\Theta$ is then coded in the prior probability density $\pi_h$. The inverse problem is to approximate pair of unknowns $(U, \Theta)$. If we only know that $\theta > 0$ we can use the improper prior density

$$\pi_h(\theta) = \pi_+(\theta) = \begin{cases} 0 & \text{if } \theta < 0, \\ 1 & \text{if } \theta \geqslant 0. \end{cases}$$

Note that $\pi_+$ is an improper density, since it is not integrable. In practice, we can assume an upper bound that we hope will never play a role. The posterior density is then given as

$$\pi^f(u, \theta) = \pi_+(\theta) \exp\left( -\frac{1}{2\delta^2}\|f - Au\|^2 - \frac{1}{2\theta}\|Lu\|^2 - \frac{d}{2}\log(\theta)\right).$$

To find the MAP estimator we can use sequential optimisation, where we update the value for $u$ using the value for $\theta$ from previous step and then use this value of $u$ to update $\theta$;

1. Initialise $\theta = \theta_0$, set $k = 1$.

2. Update $u$,

$$u^k = \arg\max_{u\in\mathbb{R}^d}\{\pi^f(u\,|\,\theta^{k-1})\} = \arg\min_{u\in\mathbb{R}^d}\left\{\frac{1}{2\delta^2}\|f - Au\|^2 + \frac{1}{2\theta}\|Lu\|^2\right\}.$$

3. Update $\theta$,

$$\theta^k = \arg\max_{\theta\geqslant 0}\{\pi^f(\theta\,|\,u^k)\} = \arg\min_{\theta\geqslant 0}\left\{\frac{1}{2\theta}\|Lu\|^2 + \frac{d}{2}\log(\theta)\right\}.$$

4. Increase $k$ by one and repeat from 2. until convergence.

We calculated the update for $u$ in Example 6.3.1. For $\theta$ we notice that the derivative is zero in the minimum of the function, that is,

$$-\frac{1}{\theta^2}\|Lu\|^2 + \frac{d}{\theta} = 0 \quad \Rightarrow \quad \theta = \frac{\|Lu\|^2}{d}.$$

Assume next that we know that the signal varies slowly except for unknown number of jumps of unknown size and location. The jumps should be sudden, suggesting that the variances should be mutually independent. This means that instead of assuming $W_j \sim \mathcal{N}(0, \theta)$ we should assume $W_j \sim \mathcal{N}(0, \theta_j)$. Then

$$\pi(u\,|\,\theta) \propto \exp\left( -\frac{1}{2}\|D_\theta^{-1/2}Lu\|^2 - \frac{1}{2}\sum_{j=1}^{d}\log(\theta_j)\right),$$

where $D_\theta = \mathrm{diag}(\theta)$. Only a few variances can be significantly large, while most of them should be small, suggesting a hyperprior that allows rare outliers.

One option is to use Gamma distribution as a prior for $\theta_j$

$$\theta_j \sim \mathrm{Gamma}(\alpha, \theta_0), \quad \pi_h(\theta) \propto \prod_{j=1}^{d} \theta_j^{\alpha-1} \exp\Big( - \frac{\theta_j}{\theta_0}\Big).$$

Then, if $\mathcal{L}(u, \theta \,|\, f) = -\log(\pi(u, \theta \,|\, f))$, we see that

$$\mathcal{L}(u, \theta \,|\, f) \propto \frac{1}{2\delta^2}\|Au - f\|^2 + \frac{1}{2}\|D_\theta^{-1/2} Lu\|^2 + \frac{1}{\theta_0}\sum_{j=1}^{d}\theta_j - \Big(\alpha - \frac{3}{2}\Big)\sum_{j=1}^{d}\log(\theta_j).$$

We then get the following update step for $u$

$$u^k = \arg\min_{u\in\mathbb{R}^d}\left(\frac{1}{2\delta^2}\|Au - f\|^2 + \frac{1}{2}\|D_{\theta^{k-1}}^{-1/2} Lu\|^2\right).$$

To update $\theta$ we notice that $\theta_j^k$ satisfies

$$\frac{\partial}{\partial\theta_j}\mathcal{L}(u^k, \theta) = -\frac{1}{2}\left(\frac{(Lu^k)_j}{\theta_j}\right)^2 + \frac{1}{\theta_0} - \Big(\alpha - \frac{3}{2}\Big)\frac{1}{\theta_j} = 0,$$

which has explicit solution

$$\theta_j^k = \theta_0\left(a + \sqrt{\frac{(Lu^k)_j^2}{2\theta_0} + a^2}\right), \quad a = \frac{1}{2}\Big(\alpha - \frac{3}{2}\Big).$$

# Chapter 7

# Infinite dimensional Bayesian inverse problems

In this section we prove a version of Bayes' theorem that can be used when the likelihood and prior are measures on separable Banach spaces. Note that there is no equivalent to Lebesgue measure in infinite dimensions (as it could not be $\sigma$-additive), and so we cannot define a measure by prescribing the form of its density. In our setting the posterior will always be absolutely continuous with respect to the prior. It is possible to construct examples even in purely Gaussian setting where this is not true. Hence working under this assumption is not strictly necessary but it is quite natural. Absolute continuity ensures that almost sure properties of the prior will be inherited by the posterior. To change such properties by data the data would have to contain infinite amount of information, which is unnatural in most applications. We follow the program laid out in [13] and [29].

Let $X$ and $Y$ denote measurable spaces and let $\nu$ and $\mu$ be probability measures on $X \times Y$. We assume that $\nu \ll \mu$. Using Theorem 6.1.1 we know that there exist a $\mu$-measurable function $\varphi : X \times Y \to \mathbb{R}$ with $\varphi \in L^1_\mu$ such that

$$\frac{d\nu}{d\mu}(x, y) = \varphi(x, y).$$

**Theorem 7.0.1.** *Assume that the conditional random variable $x \,|\, y$ exists under $\mu$ with probability distribution denoted by $\mu^y(dx)$. Then the conditional random variable $x \,|\, y$ under $\nu$ exists with probability distribution denoted by $\nu^y(dx)$. Furthermore $\nu^y \ll \mu^y$ and if $z(y) = \int_X \varphi(x, y) d\mu^y(x) > 0$ we can write*

$$\frac{d\nu^y}{d\mu^y}(x) = \frac{1}{z(y)}\varphi(x, y).$$

We will proceed to use the above theorem to construct the conditional distribution of the unknown $U$ given data $f$ from their joint probability distribution. We will need the following lemma to establish the measurability of the likelihood.

**Lemma 7.0.2.** *Let $X$ be a Borel measurable topological space and assume that $g \in C(X; \mathbb{R})$, and that $\mu(X) = 1$ for some probability measure $\mu$ on $X$. Then $g$ is a $\mu$ measurable function.*

## 7.1 Bayes' theorem for inverse problems

Let $X$, $\widetilde{Y}$ and $Y$ be separable Banach spaces, equipped with the Borel $\sigma$-algebra, and let $A : X \to \widetilde{Y} \subset Y$ be a measurable linear mapping. We are interested in approximating $U$ from a measurement

$$F = AU + N,$$

where $N \in Y$ denotes noise. We assume $(U, F) \in X \times Y$ to be a random variable and want to compute $U \mid f$. The random variable $(U, F)$ is specified as follows:

- **Prior:** $U \sim \Pi$ measure on $X$.
- **Noise:** $N \sim P_0$ measure on $Y$ and $N \perp U$.

The random variable $F \mid u$ is then distributed according to measure $P_u$, the translation of $P_0$ by $Au$. In the following we assume that $P_u \ll P_0$ for $\Pi$-a.s. Thus there exists a potential $\Phi : X \times Y \to \mathbb{R}$

$$\frac{dP_u}{dP_0}(f) = \exp(-\Phi(u; f)).$$

The mapping $\Phi(u; \cdot) : Y \to \mathbb{R}$ is measurable for a fixed $u$ and $\mathbb{E}^{P_0} \exp(-\Phi(u; f)) = 1$. For a given realisation $f$ of the data the function $-\Phi(\cdot; f)$ is called the *log likelihood*.

We define $Q_0$ to be the product measure

$$Q_0(du, df) = \Pi(du)P_0(df).$$

We assume that $\Phi(\cdot, \cdot)$ is $Q_0$ measurable. Then the random variable $(U, F) \in X \times Y$ is distributed according to measure $Q(du, df) = \Pi(du)P_u(df)$ and $Q \ll Q_0$ with

$$\frac{dQ}{dQ_0}(u, f) = \exp(-\Phi(u; f)).$$

We have the following infinite dimensional version of Theorem 6.2.1.

**Theorem 7.1.1** (Bayes' Theorem)**.** *Assume that $\Phi : X \times Y \to \mathbb{R}$ is $Q_0$ measurable and that*

$$Z(f) = \int_X \exp(-\Phi(u; f))d\Pi(u) > 0$$

*for $P_0$-a.s. Then the conditional distribution of $U \mid f$ exists under $Q$ and is denoted by $\Pi^f$. Furthermore $\Pi^f \ll \Pi$ and*

$$\frac{d\Pi^f}{d\Pi}(u) = \frac{1}{Z(f)} \exp(-\Phi(u; f)),$$

*for $f$ $Q$-a.s.*

*Proof.* The positivity of $Z(f)$ holds $Q_0$ almost surely, and hence by the absolutely continuity of $Q$ with respect to $Q_0$, it also holds $Q$ almost surely. We can then use Theorem 7.0.1. Note that since $\mu = Q_0(du, df)$ has a product form the conditional distribution of $U \mid f$ under $Q_0$ is simply $\Pi$. □

## 7.2 Well-posedness

In inverse problems small changes in data can cause large changes in the solution and hence some form of regularisation is needed to stabilise the problems. We will next show that Bayesian approach can be used to combat the ill-posedness of inverse problems so that small changes in the data will lead to small changes in the posterior measure.

In order to measure the changes in the posterior measure $\Pi^f$ caused by the changes in the data we need a metric in measures. Let $\mu$ and $\mu'$ be probability measures on separable Banach space $X$ and assume that they are both absolutely continuous with respect some reference measure $\nu$ defined in the same measure space (we can take for example $\nu = 1/2(\mu + \mu')$).

**Definition 7.2.1.** *We define the total variation distance between $\mu$ and $\mu'$ as*

$$d_{TV}(\mu, \mu') = \frac{1}{2} \int \left| \frac{d\mu}{d\nu} - \frac{d\mu'}{d\nu} \right| d\nu.$$

If $\mu' \ll \mu$ we can simplify the above and write

$$d_{TV}(\mu, \mu') = \frac{1}{2} \int \left| 1 - \frac{d\mu'}{d\mu} \right| d\mu.$$

**Definition 7.2.2.** *We define the Hellinger distance between $\mu$ and $\mu'$ as*

$$d_{Hell}(\mu, \mu') = \sqrt{\frac{1}{2} \int \left( \sqrt{\frac{d\mu}{d\nu}} - \sqrt{\frac{d\mu'}{d\nu}} \right)^2 d\nu}.$$

If $\mu' \ll \mu$ we can simplify the above and write

$$d_{Hell}(\mu, \mu') = \sqrt{\frac{1}{2} \left( 1 - \sqrt{\frac{d\mu'}{d\mu}} \right)^2 d\mu}.$$

Note that we have

$$0 \leqslant d_{TV}(\mu, \mu') \leqslant 1 \quad \text{and} \quad 0 \leqslant d_{Hell}(\mu, \mu') \leqslant 1$$

Hellinger and total variation distances generate the same topology and we have the following inequalities.

**Lemma 7.2.3.** *The total variation and Hellinger metrics are related by the inequalities*

$$\frac{1}{\sqrt{2}} d_{TV}(\mu, \mu') \leqslant d_{Hell}(\mu, \mu') \leqslant \sqrt{d_{TV}(\mu, \mu')}.$$

Let $X$ and $Y$ be separable Banach spaces, equipped with the Borel $\sigma$-algebra, and let $\Pi$ be a measure on $X$. We want to study the posterior distribution defined in the previous section. To make sense of it we need the following assumption.

**Assumption 7.2.4.** Let $X' \subset X$ and assume that $\Phi \in C(X' \times Y; \mathbb{R})$. Assume further that there are functions $M_i : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$, $i = 1, 2$, which are monotonic non-decreasing separately in each argument, and with $M_2$ strictly positive, such that for all $u \in X'$ and $f, f' \in B_Y(0, r)$,

$$-\Phi(u; f) \leqslant M_1(r, \|u\|_X),$$
$$|\Phi(u; f) - \Phi(u; f')| \leqslant M_2(r, \|u\|_X) \|f - f'\|.$$

**Theorem 7.2.5.** *Let Assumption 7.2.4 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(X' \cap B) > 0$ *for some bounded set* $B \subset X$*. We also assume that*

$$\exp(M_1(r, \|u\|_X)) \in L^1_\Pi(X; \mathbb{R}), \tag{7.1}$$

*for every fixed* $r > 0$*. Then* $Z(f) = \int_X \exp(-\Phi(u; f))d\Pi(u)$ *is positive and finite for every* $f \in Y$ *and the posterior probability measure* $\Pi^f$ *given by Theorem 7.1.1 is well defined.*

*Proof.* The boundedness of $Z(f)$ follows directly from the lower bound on $\Phi$ in Assumption 7.2.4 together with the integrability condition assumed in the theorem.

If $U \sim \Pi$ then $u \in X'$ a.s. and we can write

$$Z(f) = \int_{X'} \exp(-\Phi(u; f))d\Pi(u).$$

We also note that, since $B' = B \cap X'$ is bounded by assumption, $\sup_{u \in B'} \|u\|_X = R_1 < \infty$. Since $\Phi : X' \times Y \to \mathbb{R}$ is continuous it is finite at every point in $B' \times \{f\}$. Thus we see that

$$\sup_{(u,f) \in B' \times B_Y(0,r)} \Phi(u; f) = R_2 < \infty.$$

Hence

$$Z(f) \geqslant \int_{B'} \exp(-R_2)d\Pi(u) = \exp(-R_2)\Pi(B') > 0.$$

□

The above theorem shows that the measure $\Pi^f$ is well-defined and normalisable. We did not need to check normalisability in Theorem 7.1.1 because $\Pi^f$ was defined as a regular conditional probability via Theorem 7.0.1 which makes it automatically normalisable.

**Theorem 7.2.6.** *Let Assumption 7.2.4 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(B \cap X') > 0$ *for some bounded set* $B$ *in* $X$*. We assume also that*

$$\exp(M_1(r, \|u\|_X))\Big(1 + M_2(r, \|u\|_X)^2\Big) \in L^1_\Pi(X; \mathbb{R}),$$

*for every fixed* $r > 0$*. Then there exists* $c = c(r) > 0$ *such that*

$$d_{Hell}(\Pi^f, \Pi^{f'}) \leqslant c\|f - f'\|_Y,$$

*for all* $f, f' \in B_Y(0, r)$*.*

*Proof.* Let $Z(f)$ and $Z(f')$ denote the normalisation constants for $\Pi^f$ and $\Pi^{f'}$ so that

$$Z(f) = \int_{X'} \exp(-\Phi(u; f))d\Pi(u) > 0 \quad \text{and}$$
$$Z(f') = \int_{X'} \exp(-\Phi(u; f'))d\Pi(u) > 0$$

by Theorem 7.2.5. Using the local Lipschitz property of the exponential and the assumed Lipschitz continuity of $\Phi(u; \cdot)$ together with fact that $M_2(r, \|u\|_X) \leqslant 1 + M_2(r, \|u\|_X)^2$ we

get

$$\begin{aligned}
|Z(f) - Z(f')| &\leqslant \int_{X'} |\exp(-\Phi(u;f)) - \exp(-\Phi(u;f'))| d\Pi(u) \\
&\leqslant \int_{X'} \exp(M_1(r, \|u\|_X)) |\Phi(u;f) - \Phi(u;f')| d\Pi(u) \\
&\leqslant \Big( \int_{X'} \exp(M_1(r, \|u\|_X)) M_2(r, \|u\|_X) d\Pi(u) \Big) \|f - f'\|_Y \\
&\leqslant \Big( \int_{X'} \exp(M_1(r, \|u\|_X)) \Big(1 + M_2(r, \|u\|_X)\Big) d\Pi(u) \Big) \|f - f'\|_Y \\
&\leqslant c \|f - f'\|_Y .
\end{aligned}$$

We use $c = c(r)$ to denote a constant independent of $u$ and the value may change from occurrence to occurrence.

Using the definition of Hellinger distance and the fact that $(ab - cd)^2 \leqslant 2a^2(b-d)^2 + 2(a-c)^2 d^2$ we get

$$\begin{aligned}
\Big(d_{Hell}(\Pi^f, \Pi f')\Big)^2 &= \int_X \Big( Z(f)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\Phi(u;f)\Big) - Z(f')^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\Phi(u;f')\Big) \Big)^2 d\Pi(u) \\
&\leqslant I_1 + I_2,
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \frac{1}{Z(f)} \int_{X'} \Big( \exp\Big(-\frac{1}{2}\Phi(u;f)\Big) - \exp\Big(-\frac{1}{2}\Phi(u;f')\Big) \Big)^2 d\Pi(u) \quad \text{and} \\
I_2 &= (Z(f)^{-\frac{1}{2}} - Z(f')^{-\frac{1}{2}})^2 \int_{X'} \exp\big(-\Phi(u;f')\big) d\Pi(u).
\end{aligned}$$

Using the Assumption 7.2.4 and the fact that $Z(f) > 0$ we can use similar Lipschitz calculation as before and write

$$\begin{aligned}
I_1 &\leqslant \frac{1}{4Z(f)} \int_{X'} \exp\big(M_1(r, \|u\|_X)\big) |\Phi(u;f) - \Phi(u;f')|^2 d\Pi(u) \\
&\leqslant \frac{\|f - f'\|_Y^2}{4Z(f)} \int_{X'} \exp\big(M_1(r, \|u\|_X)\big) M_2(r, \|u\|_X)^2 d\Pi(u) \\
&\leqslant c \|f - f'\|_Y^2 .
\end{aligned}$$

We note that Assumption 7.2.4 with (7.1) implies

$$\int_{X'} \exp\big(-\Phi(u;f')\big) d\Pi(u) \leqslant \int_{X'} \exp\big(M_1(r, \|u\|_X)\big) d\Pi(u) < \infty.$$

Hence

$$I_2 \leqslant \frac{c\big(Z(f) - Z(f')\big)^2}{\min\big(Z(f)^3, Z(f')^3\big)} \leqslant c \|f - f'\|_Y^2,$$

which completes the proof. □

Hellinger distance has the desirable property of giving bounds for expectations.

**Lemma 7.2.7.** *Let $\mu$ and $\mu'$ be two probability measures on a separable Banach space $X$. Assume that $g : X \to E$, where $(E, \|\cdot\|)$ is a separable Banach space, is measurable and has second moments with respect to both $\mu$ and $\mu'$. Then*

$$\|\mathbb{E}^{\mu} g - \mathbb{E}^{\mu'} g\| \leqslant 2\sqrt{\mathbb{E}^{\mu}\|g\|^2 + \mathbb{E}^{\mu'}\|g\|^2}\; d_{Hell}(\mu, \mu')$$

The proof of the above lemma is left as an exercise.
Using Lemma 7.2.7 we see that, for $f, f' \in B_Y(0, r)$,

$$|\mathbb{E}^{\Pi^f} g(u) - \mathbb{E}^{\Pi^{f'}} g(u)| \leqslant c_{g,r}\|f - f'\|_Y$$

If $\Pi$ is Gaussian we can use the following Fernique theorem to establish the integrability conditions in the above theorems.

**Theorem 7.2.8** (Fernique)**.** *Let $\Pi$ be a Gaussian probability measure on a separable Banach space $X$. Then there exists $\alpha > 0$ such that*

$$\int_X \exp(\alpha\|u\|_X^2) d\Pi(u) < \infty.$$

## 7.3 Approximation of the potential

In this section we will examine the continuity properties of the posterior measure with respect to approximation of the potential $\Phi$. The data $f$ is assumed to be fixed in this section so we will write $Z(f) = Z$ and $\Phi(u; f) = \Phi(u)$. Let $X$ be a separable Banach space and $\Pi$ a measure on $X$. Assume that $\Pi^f$ and $\Pi_N^f$ are both absolutely continuous with respect to $\Pi$ and given by

$$\begin{aligned} \frac{d\Pi^f}{d\Pi}(u) &= \frac{1}{Z}\exp(-\Phi(u)), \quad Z = \int_X \exp(-\Phi(u)) d\Pi(u) \quad \text{and} \\ \frac{d\Pi_N^f}{d\Pi}(u) &= \frac{1}{Z_N}\exp(-\Phi_N(u)), \quad Z_N = \int_X \exp(-\Phi_N(u)) d\Pi(u). \end{aligned} \tag{7.2}$$

The measure $\Pi_N$ can arise e.g. when approximating the forward map $A$ in (6.1). It is important to know whether closeness of the forward map and its approximation imply closeness of the posterior measure.

**Assumption 7.3.1.** Let $X' \subset X$ and assume that $\Phi \in C(X'; \mathbb{R})$. Assume further that there exists functions $M_i : \mathbb{R}^+ \to \mathbb{R}^+$, $i = 1, 2$ that are independent of $N$, non-decreasing and $M_2$ being strictly positive, such that for all $u \in X'$,

$$\begin{aligned} \Phi(u) \geqslant -M_1(\|u\|_X), \quad \Phi_N(u) \geqslant -M_1(\|u\|_X) \quad \text{and} \\ |\Phi(u) - \Phi_N(u)| \leqslant M_2(\|u\|_X)\psi(N) \end{aligned}$$

where $\psi(N) \to 0$ as $N \to \infty$.

The following theorems are similar to the ones in the previous section but they estimate changes in the posterior caused by changes in the potential $\Phi$ rather than data $f$.

**Theorem 7.3.2.** *Let Assumption 7.3.1 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(B \cap X') > 0$ *for some bounded set* $B$ *in* $X$*. We also assume that*

$$\exp(M_1(\|u\|_X)) \in L^1_\Pi(X; \mathbb{R}).$$

*Then* $Z$ *and* $Z_N$ *defined in 7.2 are positive and finite, and the probability measures* $\Pi^f$ *and* $\Pi^f_N$ *are well defined. Furthermore, for sufficiently large* $N$*,* $Z_N$ *is bounded below by a positive constant independent of* $N$*.*

*Proof.* The finiteness of $Z$ and $Z_N$ follows from the lower bounds on $\Phi$ and $\Phi_N$ given in Assumption 7.3.1 combined with the integrability condition assumed in the theorem. Since $U \sim \Pi$ satisfies $U \in X'$ a.s. we have

$$Z = \int_{X'} \exp(-\Phi(u)) d\Pi(u)$$

Note that $B' = B \cap X'$ is bounded in $X$ and hence $\sup_{u \in B'} \|u\|_X = R_1 < \infty$. Since $\Phi : X' \to \mathbb{R}$ is continuous it is finite in every point of $B'$. Using the Assumption 7.3.1 for large enough $N$ we can write

$$\sup_{u \in B'} |\Phi(u) - \Phi_N(u)| \leqslant R_2 < \infty.$$

This implies

$$\sup_{u \in B'} \Phi(u) = 2R_2 < \infty \quad \text{and} \quad \sup_{u \in B'} \Phi_N(u) = 2R_2 < \infty.$$

Hence

$$Z \geqslant \int_{B'} \exp(-2R_2) d\Pi(u) = \exp(-2R_2)\Pi(B') > 0.$$

We get the same lower bound for $Z_N$ and note that it is independent of $N$ as required.

□

**Theorem 7.3.3.** *Let Assumption 7.3.1 hold. Assume that* $\Pi(X') = 1$ *and that* $\Pi(B \cap X') > 0$ *for some bounded set* $B$ *in* $X$*. We assume furthermore that*

$$\exp(M_1(\|u\|_X))\Big(1 + M_2(\|u\|_X)^2\Big) \in L^1_\Pi(X; \mathbb{R}).$$

*Then there exists* $c > 0$ *such that*

$$d_{Hell}(\Pi^f, \Pi^f_N) \leqslant c\psi(N)$$

*for all sufficiently large* $N$*.*

*Proof.* Let $N$ be sufficiently large so that by Theorem 7.3.3 $Z > 0$ and $Z_N > 0$ with positive lower bounds independent of $N$. Using the local Lipschitz property of exponential, Assumption 7.3.1 and the fact that $M_2(\|u\|_X) \leqslant 1 + M_2(\|u\|_X)^2$ we can write

$$\begin{aligned}
|Z - Z_N| &\leqslant \int_{X'} |\exp(-\Phi(u)) - \exp(-\Phi_N(u))| d\Pi(u) \\
&\leqslant \int_{X'} \exp(M_1(\|u\|_X)) |\Phi(u) - \Phi_N(u)| d\Pi(u) \\
&\leqslant \psi(N) \int_{X'} \exp(M_1(\|u\|_X)) M_2(\|u\|_X) d\Pi(u) \\
&\leqslant \psi(N) \int_{X'} \exp(M_1(\|u\|_X))(1 + M_2(\|u\|_X)^2) d\Pi(u) \\
&\leqslant C\psi(N),
\end{aligned}$$

where $C$ is a constant that does not depend on $u$ or $N$. As in the proof of Theorem 7.2.6 we can write

$$\Big(d_{Hell}(\Pi^f, \Pi^f_N)\Big)^2 = I_1 + I_2,$$

where

$$I_1 = \frac{1}{Z}\int_{X'} \Big(\exp\big(-\frac{1}{2}\Phi(u)\big) - \exp\big(-\frac{1}{2}\Phi_N(u)\big)\Big)^2 d\Pi(u) \quad \text{and}$$
$$I_2 = \big(Z^{-\frac{1}{2}} - Z_N^{-\frac{1}{2}}\big)^2 \int_{X'} \exp\big(\Phi_N(u)\big) d\Pi(u).$$

Using similar arguments as above we see that

$$\begin{aligned} I_1 &\leqslant \frac{1}{4Z}\int \exp(M_1(\|u\|_X))|\Phi(u) - \Phi_N(u)|^2 d\Pi(u) \\ &\leqslant \frac{\psi(N)^2}{Z}\int \exp(M_1(\|u\|_X)) M_2(\|u\|_X)^2 d\Pi(u) \\ &\leqslant C\psi(N)^2. \end{aligned}$$

We also notice that

$$\int_{X'} \exp\big(\Phi_N(u)\big) d\Pi(u) \leqslant \int_{X'} \exp\big(M_1(\|u\|_X)\big) d\Pi(u) < \infty$$

and the upper bound is independent of $N$. Hence

$$I_2 \leqslant \frac{c\big(Z - Z_N\big)^2}{\min\big(Z^3, Z_N^3\big)} \leqslant C\psi(N)^2,$$

which concludes the proof. □

## 7.4 Infinite dimensional Gaussian measure (non examinable)

We start by introducing infinite dimensional Gaussian random variables and some of their key properties. For more details see e.g. [19, Section 3] or [12, Section 2], and if you feel brave [9].

Let $X$ be a separable Banach space and denote by $X^*$ its dual space of linear functionals on $X$. We define the characteristic function of a probability distribution $\mu$ on a separable Banach Space $X$ as

$$\varphi_\mu(\psi) = \mathbb{E}\exp(i\psi(x)),$$

for $\psi \in X^*$.

**Theorem 7.4.1.** *If $\mu$ and $\nu$ are two probability measures on a separable Banach space $X$ and if $\varphi_\mu(\psi) = \varphi_\nu(\psi)$, for all $\psi \in X^*$, then $\mu = \nu$.*

A function $\theta \in X$ is called the mean of $\mu$ if $\psi(\theta) = \int_X \psi(v)d\mu(v)$ for all $\psi \in X^*$. A linear operator $\Sigma : X^* \to X$ is called the covariance operator if $\psi(\Sigma\varphi) = \int_X \psi(v-\theta)\varphi(v-\theta)d\mu(v)$ for all $\psi, \varphi \in X^*$. If we assume that $X = \mathcal{H}$ is a Hilbert space then $\theta = \mathbb{E}(V)$ and the covariance operator is characterised by identity $\mathbb{E}\big(\langle\varphi, (V-\theta)\rangle\langle\psi, (V-\theta)\rangle\big) = \langle\Sigma\varphi, \psi\rangle$.

A measure $\mu$ on $(X, \mathcal{B}(X))$ is Gaussian if, for any $\psi \in X^*$, $\psi(V) \sim \mathcal{N}(\theta_\psi, \sigma_\psi^2)$ for some $\theta_\psi \in \mathbb{R}$ and $\sigma_\psi \in \mathbb{R}$. We allow $\sigma_\psi = 0$, so that the measure may be a Dirac mass at $\theta_\psi$. Note that it is expected that $\theta_\psi = \psi(\theta)$ and $\sigma_\psi^2 = \psi(\Sigma\psi)$ for all $\psi \in X^*$.

**Theorem 7.4.2.** *A Gaussian measure $\mu$ on $(X, \mathcal{B}(X))$ has a mean $\theta$ and covariance operator $\Sigma$. The characteristic function of the measure is*

$$\varphi_\mu(\psi) = \exp\left(i\psi(\theta) - \frac{1}{2}\psi(\Sigma\psi)\right).$$

Using the above Theorem and Theorem 7.4.1 we see that the mean and covariance completely characterise the Gaussian measure and hence we can simply write $\mathcal{N}(\theta, \Sigma)$.

**Definition 7.4.3.** *Let $\{\varphi_i\}_{i=1}^\infty$ denote an orthonormal basis for a separable Hilbert space $\mathcal{H}$. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is trace-class if*

$$Tr(A) = \sum_{i=1}^{\infty} \langle A\varphi_i, \varphi_i\rangle < \infty.$$

*The sum is independent of the choice of basis. The operator $A$ is Hilbert–Schmidt if*

$$Tr(A^*A) = \sum_{i=1}^{\infty} \|A\varphi_i\|_{\mathcal{H}}^2 < \infty.$$

We can construct random draws from a Gaussian measure on a Hilbert space $\mathcal{H}$ using Karhunen–Loève expansion.

**Theorem 7.4.4.** *Let $\Sigma$ be a self-adjoint, positive semi-definite, trace class operator in a Hilbert space $\mathcal{H}$, and let $\theta \in \mathcal{H}$. Let $\{\varphi_k, \gamma_k\}$ be an orthonormal set of eigenvectors and eigenvalues for $\Sigma$ ordered so that $\gamma_1 \geqslant \gamma_2 \geqslant \cdots$. Take $\{\xi_k\}_{k=1}^\infty$ to be an i.i.d. sequence with $\xi_1 \sim \mathcal{N}(0, 1)$. Then the random variable $V \in \mathcal{H}$ given by the Karhunen–Loève expansion*

$$V = \theta + \sum_{k=1}^{\infty} \sqrt{\gamma_k}\xi_k\varphi_k \tag{7.3}$$

*is distributed according to $\mu = \mathcal{N}(\theta, \Sigma)$.*

The proof is left as an exercise.

**Example 7.4.5.** A random variable $N$ is said to be white Gaussian noise on $L^2(\mathbb{T}^d)$ if $\mathbb{E}(N) = 0$ and $\mathbb{E}\big(\langle N, \varphi\rangle\langle N, \psi\rangle\big) = \langle \varphi, \psi\rangle$, in which case we denote $N \sim \mathcal{N}(0, I)$. Note that $I : L^2(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$ is not a trace class operator in $L^2(\mathbb{T}^d)$, and hence white noise does not take values in $L^2(\mathbb{T}^d)$. Let $e_{\vec{\ell}} \in L^2(\mathbb{T}^d)$, $\vec{\ell} = (\ell_1, \ell_2, \ldots, \ell_d) \in \mathbb{Z}^d$ be an orthonormal basis of $L^2(\mathbb{T}^d)$ consisting of eigenfunctions of Laplacian, numbered so that $-\Delta e_{\vec{\ell}} = |\vec{\ell}|^2 e_{\vec{\ell}}$. Such functions $e_{\vec{\ell}}(x)$ can be chosen to be normalised products of the sine and cosine functions $\sin(\ell_j x_j)$ and $\cos(\ell_j x_j)$ that form the standard Fourier basis of $L^2(\mathbb{T}^d)$. The Fourier coefficients of $N$ with respect to this basis are independent, normally distributed $\mathbb{R}$-valued random variables with variance one, that is, $\langle N, e_{\vec{\ell}}\rangle \sim N(0, 1)$. Then

$$\mathbb{E}\|N\|_{L^2(\mathbb{T}^d)}^2 = \sum_{\vec{\ell}\in\mathbb{Z}^d} \mathbb{E}|\langle N, e_{\vec{\ell}}\rangle|^2 = \sum_{\vec{\ell}\in\mathbb{Z}^d} 1 = \infty.$$

This implies that realisations of $N$ are in $L^2(\mathbb{T}^d)$ with probability zero. However, when $s > d/2$

$$\mathbb{E}\|N\|_{H^{-s}(\mathbb{T}^d)}^2 = \sum_{\vec{\ell}\in\mathbb{Z}^d} (1 + |\vec{\ell}|^2)^{-s}\mathbb{E}|\langle N, e_{\vec{\ell}}\rangle|^2 < \infty \tag{7.4}$$

and hence $N$ takes values in $H^{-s}(\mathbb{T}^d)$ a.s. For more details about Sobolev spaces see Appendix 8.

The above result can be generalised to show that if $V \sim \mathcal{N}(0, \Sigma)$ and the eigenvalues of $\Sigma$ satisfy $\gamma_j \asymp j^{-\frac{2s}{d}}$ (e.g. $\Sigma = (I - \Delta)^{-s}$) then, for $t < s - d/2$, we have $v \in H^t$ a.s. We can also generalise the results for more general domains than the torus or $\mathbb{R}^d$ using Hilbert scales. These spaces do not, in general, coincide with Sobolev spaces, because of the effect of the boundary conditions.

The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ of a Gaussian on $\mathcal{H}$ is a compact operator and its inverse is densely defined unbounded operator on $\mathcal{H}$. We call this inverse precision operator. Bot the covariance and the precision operator are self-adjoint on appropriate domains and the fractional powers of them can be defined via spectral theorem.

Given a Gaussian measure $\mu$ on a separable Banach space $X$, we define the Cameron–Martin space $H_\mu \subset X$ of $\mu$ to be the intersection of all linear spaces of full measure. The main importance of the Cameron-Martin space is that it characterises exactly the directions in $X$ in which a centred Gaussian measure can be shifted to obtain an equivalent Gaussian measure. When $\dim(X) = \infty$ the measure of the Cameron–Martin space is zero, that is, $\mu(H_\mu) = 0$. Compare this to the case of finite dimensional Lebesgue measure which is invariant under translations in any direction. This is a striking illustration of the fact that measures in infinite-dimensional spaces have a strong tendency of being mutually singular.

**Lemma 7.4.6.** *For a Gaussian measure on Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ the Cameron–Martin space $H_\mu$ consists of the image of $\mathcal{H}$ under $\Sigma^{\frac{1}{2}}$ and the Cameron–Martin norm is given by $\|h\|_\mu^2 = \|\Sigma^{-\frac{1}{2}} h\|_{\mathcal{H}}^2$.*

**Theorem 7.4.7.** *Let $\mu = \mathcal{N}(0, \Sigma)$ be a Gaussian measure on a separable Banach space $X$. The Cameron–Martin space $H_\mu$ of $\mu$ can be endowed with Hilbert space structure and $H_\mu$ is compactly embedded in all separable spaces $X'$ such that $\mu(X') = 1$.*

**Theorem 7.4.8** (Special case of the Cameron-Martin theorem)**.** *Let $\mu = \mathcal{N}(0, \Sigma)$ be a Gaussian measure on a separable Banach space $X$. Denote by $\mu_h$ the translation of $\mu$ by $h$, $\mu_h = \mu(\cdot - h)$. If $h \in H_\mu$ then $\mu_h$ is absolutely continuous with respect to $\mu$ and*

$$\frac{d\mu_h}{d\mu}(v) = \exp\left( -\frac{1}{2}\|h\|_{H_\mu}^2 + \langle h, v\rangle_{H_\mu} \right)$$

*$v \in X$, $\mu$-a.s. If $h \notin H_\mu$, then $\mu$ and $\mu_h$ are mutually singular.*

**Example 7.4.9.** Consider two Gaussian measures $\mu_i$, $i = 1, 2$, on $\mathcal{H} = L^2((0,1))$ both with precision operator (the densely defined inverse covariance operator $\Sigma^{-1} = \mathcal{L}$) $\mathcal{L} = -d^2/dx^2$, the domain of $\mathcal{L}$ being $H_0^1((0,1)) \cap H^2((0,1))$. We assume that $\mu_1 \sim \mathcal{N}(\theta, \Sigma)$ and $\mu_2 \sim \mathcal{N}(0, \Sigma)$. Then $H_\mu = \mathrm{Im}(\Sigma^{1/2}) = H_0^1((0,1))$. Hence the measures are equivalent if and only if $\theta \in H_\mu$. If this is satisfied then the Radon–Nikodym derivative between the two measures is given by

$$\frac{d\mu_1}{d\mu_2}(v) = \exp\left( \langle \theta, v\rangle_{H_0^1} - \frac{1}{2}\|\theta\|_{H_0^1}^2 \right).$$

## 7.5 MAP estimators and Tikhonov regularisation

In this section we assume that the prior $\Pi$ is Gaussian. We show that MAP estimators (point of maximal probability) coincide with the minimisers of Tikhonov regularised least

squares functions with regularisation term being given by the Cameron-Martin norm of the Gaussian prior.

The classical deterministic way for solving inverse problem is to try and minimise the potential $\Phi$ with some regularisation. If we have finite data and Gaussian observational noise $N \sim N(0, \Gamma)$ we can write

$$\Phi(u; f) = \frac{1}{2}\left\|\Gamma^{-1/2}(f - Au)\right\|^2.$$

Thus $\Phi$ is covariance weighted data misfit least square function.

We assume that $\Pi$ is a Gaussian probability measure on a separable Banach space $(X, \|\cdot\|_X)$ and $\Pi(X) = 1$. We denote the Cameron-Martin space of $\Pi$ by $(H_\Pi, \|\cdot\|_{H_\Pi})$. In this section we want to show that maximising $\Pi^f$ is equivalent to minimising

$$I(u) = \begin{cases} \Phi(u; f) + \frac{1}{2}\|u\|^2_{H_\Pi} & \text{if } u \in H_\Pi, \text{ and} \\ \infty & \text{else.} \end{cases} \tag{7.5}$$

The realisation $f$ of the data does not play role in this section and we will write $\Phi(u; f) = \Phi(u)$.

We note that the properties of $\Phi$ we assume below are typically determined by the forward operator, which maps the unknown function $u$ to the data $f$. Probability theory does not play a direct role in verifying these properties of $\Phi$. Probability becomes relevant when choosing the prior measure $\Pi$ so that it charges the Banach space $X$, on which the desired properties of $\Phi$ hold, with full measure.

**Assumption 7.5.1.** The function $\Phi : X \to \mathbb{R}$ satisfies the following conditions:

1) For every $\varepsilon > 0$ there is an $R = R(\varepsilon) \in \mathbb{R}$, such that for all $u \in X$,

$$\Phi(u) \geqslant R - \varepsilon\|u\|^2_X.$$

2) $\Phi$ is locally bounded above, that is, for every $r > 0$ there exists $K = K(r) > 0$ such that, for all $u \in X$ with $\|u\|_X < r$, we have

$$\Phi(u) \leqslant K.$$

3) $\Phi$ is locally Lipschitz continuous i.e. for every $r > 0$ there exists $L = L(r) > 0$ such that, for all $u_1, u_2 \in X$ with $\|u_1\|_X, \|u_2\|_X < r$, we have

$$|\Phi(u_1) - \Phi(u_2)| \leqslant L\|u_1 - u_2\|_X.$$

In finite dimensions there is an obvious notion of most likely points for measures which have a continuous density with respect to Lebesgue measure: the points at which the Lebesgue density is maximised. Unfortunately we can not translate this idea to infinite dimensions. To fix this we will restate the idea in a way that will work also in infinite dimensional settings. Fix a small radius $\delta > 0$ and identify centres of balls of radius $\delta$ which have maximal probability. Letting $\delta \to 0$ then recovers the maximums when there is continuous Lebesgue density. We will use this small ball approach in infinite dimensional settings.

Let $z \in H_\Pi$ and $B_\delta(z) \subset X$ be the open ball centred at $z \in X$ with radius $\delta$ in $X$. Let

$$J^f_\delta(z) = \Pi^f(B_\delta(z))$$

be the mass of the ball $B_\delta(z)$ under the posterior measure $\Pi^f$. Similarly we define

$$J_\delta(z) = \Pi(B_\delta(z))$$

to be the mass of the ball $B_\delta(z)$ under the Gaussian prior. We note that all balls in a separable Banach space have positive Gaussian measure. Thus $J_\delta(z)$ is finite and positive for any $z \in H_\Pi$. By the above assumptions on $\Phi$ and the Fernique Theorem 7.2.8 the same is true for $J^f_\delta(z)$. We will next prove that the probability is maximised where $I$ is minimised.

**Theorem 7.5.2.** *Let Assumption 7.5.1 hold and assume that* $\Pi(X) = 1$. *Then, for any* $z_1, z_2 \in H_\Pi$,

$$\lim_{\delta\to 0} \frac{J^f_\delta(z_1)}{J^f_\delta(z_2)} = \exp\big(I(z_2) - I(z_1)\big),$$

*where the function* $I$ *is defined by* (7.5).

Before moving to prove the above theorem we state a result about the small ball probabilities under Gaussian measure

**Theorem 7.5.3.** *Let* $z \in H_\Pi$ *and* $B_\delta(z) \subset X$ *be the open ball centred at* $z \in X$ *with radius* $\delta$ *in* $X$. *The ratio of small ball probabilities under Gaussian measure* $\Pi$ *satisfies*

$$\lim_{\delta\to 0} \frac{\Pi(B_\delta(z_1))}{\Pi(B_\delta(z_2))} = \exp\left(\frac{1}{2}\|z_2\|^2_{H_\Pi} - \frac{1}{2}\|z_1\|^2_{H_\Pi}\right).$$

*Proof of theorem 7.5.2.* The ratio is finite and positive since $J^f_\delta(z)$ is finite and positive for any $z \in H_\Pi$. The estimate given in Theorem 7.5.3 transfers the question about probability into statement concerning the Cameron-Martin norm of $\Pi$. Note that if $U \sim \Pi$ then its realisation is in $H_\Pi$ only with probability zero and hence $\|u\|_{H_\Pi} = \infty$ almost surely.

We can write

$$\begin{aligned}\frac{J^f_\delta(z_1)}{J^f_\delta(z_2)} &= \frac{\int_{B_\delta(z_1)} \exp(-\Phi(u))d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(v))d\Pi(v)} \\ &= \frac{\int_{B_\delta(z_1)} \exp(-\Phi(u)+\Phi(z_1))\exp(-\Phi(z_1))d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(v)+\Phi(z_2))\exp(-\Phi(z_2))d\Pi(v)}.\end{aligned}$$

By Assumption 7.5.1 there exists $L = L(r)$ such that

$$-L\|u_1 - u_2\|_X \leqslant \Phi(u_1) - \Phi(u_2) \leqslant L\|u_1 - u_2\|_X$$

for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$. We can then write

$$\begin{aligned}\frac{J^f_\delta(z_1)}{J^f_\delta(z_2)} &\leqslant e^{2\delta L}\frac{\int_{B_\delta(z_1)} \exp(-\Phi(z_1))d\Pi(u)}{\int_{B_\delta(z_2)} \exp(-\Phi(z_2))d\Pi(v)} \\ &\leqslant e^{2\delta L}e^{-\Phi(z_1)+\Phi(z_2)}\frac{\Pi(B_\delta(z_1))}{\Pi(B_\delta(z_2))},\end{aligned}$$

Using Theorem 7.5.3 we get

$$\frac{J_\delta^f(z_1)}{J_\delta^f(z_2)} \leqslant r_1(\delta)e^{2\delta L}e^{-I(z_1)+I(z_2)}$$

where $r_1(\delta) \to 1$ as $\delta \to 0$. Thus

$$\limsup_{\delta\to 0} \frac{J_\delta^f(z_1)}{J_\delta^f(z_2)} \leqslant e^{-I(z_1)+I(z_2)}.$$

We can deduce in the same way that

$$\frac{J_\delta^f(z_1)}{J_\delta^f(z_2)} \geqslant r_2(\delta)e^{-2\delta L}e^{-I(z_1)+I(z_2)}$$

with $r_2(\delta) \to 1$ as $\delta \to 0$ and furthermore

$$\liminf_{\delta\to 0} \frac{J_\delta^f(z_1)}{J_\delta^f(z_2)} \geqslant e^{-I(z_1)+I(z_2)},$$

which concludes the proof. □

We will next show that the minimisation problem for $I$ is well-defined when Assumption 7.5.1 holds.

**Definition 7.5.4.** *Let $E$ be a Hilbert space. The function $I : E \to \mathbb{R}$ is weakly lower semicontinuous if*

$$\liminf_{j\to\infty} I(u_j) \geqslant I(u)$$

*whenever $u_j \rightharpoonup u$ in $E$. The function $I : E \to \mathbb{R}$ is weakly continuous if*

$$\lim_{j\to\infty} I(u_j) = I(u)$$

*whenever $u_j \rightharpoonup u$ in $E$.*

**Lemma 7.5.5.** *Let $(E, \langle\cdot,\cdot\rangle_E)$ be a Hilbert space with induced norm $\|\cdot\|_E$. Then the quadratic form $J(u) = \frac{1}{2}\|u\|_E^2$ is weakly lower semicontinuous.*

*Proof.* We can write

$$\begin{aligned} J(u_j) - J(u) &= \frac{1}{2}\|u_j\|_E^2 - \frac{1}{2}\|u\|_E^2 \\ &= \frac{1}{2}\langle u_j - u, u_j + u\rangle_E \\ &= \frac{1}{2}\langle u_j - u, 2u\rangle_E + \frac{1}{2}\|u_j - u\|_E^2 \\ &\geqslant \frac{1}{2}\langle u_j - u, 2u\rangle_E \to 0, \end{aligned}$$

when $u_j \rightharpoonup u$ in $E$. □

**Theorem 7.5.6.** *Suppose that Assumption 7.5.1 holds and let $E$ be a Hilbert space compactly embedded in $X$. Then there exists $\overline{u} \in E$ such that*

$$I(\overline{u}) = \overline{I} := \inf\{I(u) \ : \ u \in E\}.$$

*Furthermore if $\{u_j\}$ is a minimising sequence satisfying $I(u_j) \to I(\overline{u})$ then there exists a subsequence $\{u_{j'}\}$ that converges strongly to $\overline{u}$ in $E$.*

*Proof.* Compactness of $E \subset X$ implies that $\|u\|_X \leqslant C\|u\|_E$. Hence by Assumption 7.5.1 1) it follows that for any $\varepsilon > 0$ there is $R(\varepsilon) \in \mathbb{R}$ such that

$$I(u) \geqslant \Big(\frac{1}{2} - \varepsilon C\Big)\|u\|_E^2 + R(\varepsilon).$$

We can choose $\varepsilon$ small enough so that

$$I(u) \geqslant \frac{1}{4}\|u\|_E^2 + R \tag{7.6}$$

for all $u \in E$ with some $R \in \mathbb{R}$.

Let $u_j$ be minimising sequence satisfying $I(u_j) \to I(\overline{u})$ as $j \to \infty$. For any $\delta > 0$ there is $N = N(\delta)$, such that for all $j \geqslant N$

$$\overline{I} \leqslant I(u_j) \leqslant \overline{I} + \delta. \tag{7.7}$$

We can then use (7.6) to conclude that $\{u_j\}$ is bounded in $E$. We assumed that $E$ is a Hilbert space so there exists $\overline{u} \in E$ and a subsequence $u_{j'}$ such that $u_{j'} \rightharpoonup \overline{u}$ in $E$. Since $E$ is compactly embedded in $X$ we can deduce that there is a subsubsequence (also denoted by $u_{j'}$) so that$u_{j'} \to \overline{u}$ strongly in $X$. By the Assumption 7.5.1 3) the potential $\Phi$ is Lipschitz continuous and hence $\Phi(u_{j'}) \to \Phi(\overline{u})$. Thus $\Phi$ is weakly continuous on $E$. Using Lemma 7.5.5 we see that $I(u) = J(u) + \Phi(u)$ is weakly lower semicontinuous on $E$. Using (7.7) we can then conclude that, for any $\delta > 0$,

$$\overline{I} \leqslant I(\overline{u}) \leqslant \overline{I} + \delta.$$

Since $\delta$ can be chosen arbitrarily small the first result follows.

Next we study a subsequence of $u_{j'}$. For large enough $n, \ell$ we can write

$$\begin{aligned}
\frac{1}{4}\|u_n - u_\ell\|_E^2 &= \frac{1}{2}\|u_n\|_E^2 + \frac{1}{2}\|u_\ell\|_E^2 - \frac{1}{4}\|u_n + u_\ell\|_E^2 \\
&= I(u_n) + I(u_\ell) - 2I\Big(\frac{1}{2}(u_n + u_\ell)\Big) - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big) \\
&\leqslant 2(\overline{I} + \delta) - 2\overline{I} - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big) \\
&\leqslant 2\delta - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\Big(\frac{1}{2}(u_n + u_\ell)\Big).
\end{aligned}$$

The subsequences $u_n$, $u_\ell$ and $\frac{1}{2}(u_n + u_\ell)$ converge strongly to $\overline{u} \in X$. Since $\Phi$ is continuous we see that for large enough $n, \ell$

$$\frac{1}{4}\|u_n - u_\ell\|_E^2 \leqslant 3\delta.$$

We have shown that the subsequence is Cauchy in $E$ which completes the proof. □

Note that by Theorem 7.4.7 the Cameron–Martin space $H_\Pi$ is a Hilbert space that is compactly embedded in $X$ and hence we can find a minimiser in $H_\Pi$.

# Chapter 8

# Appendix; Sobolev spaces

Sobolev spaces constitute one of the most relevant functional settings for the treatment of PDEs and boundary value problems. This appendix gives a short introduction to the topic. Sobolev spaces are covered properly on course *Analysis of Partial Differential Equations*. For more a more detailed treatment of Sobolev spaces and applications to PDEs see [17]. For a comprehensive study of Sobolev spaces see e.g. [2].

We start by introducing the notion of a weak derivatives that generalises the classical partial derivatives.

**Definition 8.0.1** (Test functions)**.** *Let $\mathcal{O} \in \mathbb{R}^d$. We set*

$$C_0^\infty(\mathcal{O}) = \{\varphi \in C^\infty(\mathcal{O}) : supp(\varphi) \in V \subset \mathcal{O}\},$$

*the smooth functions with compact support. This space is often referred as the space of test functions and denoted by $\mathcal{D}(\mathcal{O})$.*

If $u \in C^1(\mathbb{R})$ then we can define $\frac{\partial u}{\partial x}$ by

$$\int \frac{\partial u}{\partial x}(x)\varphi(x)dx = -\int u(x)\frac{\partial \varphi}{\partial x}(x)dx,$$

for all $\varphi \in \mathcal{D}(\mathbb{R})$. We notice that the right hand side is well-defined for all $u \in L^1_{loc}(\mathbb{R})$.

**Definition 8.0.2.** *Let $\alpha = \alpha_1, \cdots, \alpha_d$ be a multi-index, $\alpha_i \in \mathbb{N}$, and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. A function $u \in L^1_{loc}(\mathcal{O})$ has a weak derivative $v = D^\alpha u \in L^1_{loc}(\mathcal{O})$ if*

$$\int_{\mathcal{O}} v(x)\varphi(x)dx = (-1)^{|\alpha|}\int_{\mathcal{O}} u(x)D^\alpha\varphi(x)dx,$$

For all test functions $\varphi \in \mathcal{D}(\mathcal{O})$. Above $D^\alpha\varphi = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}\varphi$. Note that when the weak derivative $D^\alpha u$ exists, it is defined only up to a set of measure zero. So any point-wise statements to be made about $D^\alpha u$ is understood to only hold almost surely. Most of the classical differential calculus can be reproduced for weak derivatives (e.g. the product rule and the chain rule).

**Definition 8.0.3.** *The Sobolev space $H^s(\mathcal{O})$, $s \in \mathbb{N}$, is defined as the set of all functions $u \in L^2(\mathcal{O})$ with weak derivatives $D^\alpha u \in L^2(\mathcal{O})$ up to the order $|\alpha| \leqslant s$.*

The above definition can be generalised for functions $u \in L^p(\mathcal{O})$, $1 \leqslant p \leqslant \infty$, and the resulting Sobolev spaces are usually denoted by $W^{s,p}(\mathcal{O})$. In this course we only consider $L^2(\mathcal{O})$ Sobolev spaces. The Sobolev spaces $H^s(\mathcal{O})$ are Banach spaces with the norm

$$\|u\|_{H^s} = \left( \sum_{|\alpha| \leqslant s} \|D^\alpha u\|^2_{L^2(\mathcal{O})} dx \right)^{\frac{1}{2}}. \tag{8.1}$$

The Sobolev spaces are separable Hilbert spaces with inner product

$$\langle u, v \rangle_{H^s} = \sum_{|\alpha| \leqslant s} \langle D^\alpha u, D^\alpha v \rangle_{L^2} = \sum_{|\alpha| \leqslant s} \int_{\mathcal{O}} D^\alpha u(x) D^\alpha v(x) dx,$$

for all $u, v \in H^s(\mathcal{O})$.

**Definition 8.0.4.** *The space $H^s_0(\mathcal{O})$ are the closure of $C^\infty_0(\mathcal{O})$ under the Sobolev norm (8.1).*

The spaces $H^s_0(\mathcal{O})$ is a closed subspace of $H^s(\mathcal{O})$. If $\mathcal{O} = \mathbb{R}^d$ then $H^s_0(\mathcal{O}) = H^s(\mathcal{O})$. We can define $H^1_0(\mathcal{O})$ also through Trace Theorem (see [17, Section 5.5]) which states that there is a continuous linear mapping $\mathrm{tr} : H^1(\mathcal{O}) \to L^2(\partial\mathcal{O})$ called the trace operator. In this sense, we say that functions from $H^1(\mathcal{O})$ have traces (boundary values) in $L^2(\partial\mathcal{O})$ and

$$H^1_0(\mathcal{O}) = \{u \in H^1(\mathcal{O}) : u = 0 \text{ in } \partial\mathcal{O}\}.$$

As defined above, Sobolev spaces concern integer numbers of derivatives. However, the concept can be extended to fractional derivatives using Fourier transform.

**Definition 8.0.5.** *Assume $0 \leqslant s < \infty$ and $u \in L^2(\mathbb{R}^d)$. Then $u \in H^s(\mathbb{R}^d)$ if $(1 + |\xi|^s)\widehat{u} \in L^2(\mathbb{R}^d)$. The Sobolev norm is given by*

$$\|u\|_{H^s} = \|(1 + |\cdot|^s)\widehat{u}\|_{L^2},$$

where $\widehat{u} = \mathcal{F}(u)$ is the Fourier transform. Note that for a positive integer $s$, the above definition agrees with the definition given by the weak derivatives. For $s < 0$, we define $H^s(\mathbb{R}^d)$ via duality. The resulting spaces are separable for all $s \in \mathbb{R}$. If $\mathcal{O} \subset \mathbb{R}^d$ then $H^{-1}(\mathcal{O})$ is the dual space of $H^1_0(\mathcal{O})$.

In these notes we often consider $u \in L^2(\mathbb{T}^d)$, $\mathbb{T}^d$ being the $d$-dimensional unit torus, found by identifying opposite faces of the unit cube $[0, 1]^d$. In this periodic case the Sobolev norm of the space $H(\mathbb{T}^d)$ can be written as

$$\|u\|_{H^s} = \sum_{\ell \in \mathbb{Z}^d} (1 + |\ell|^2)^s \widehat{u}(\ell)^2.$$

We define the Laplace operator $\Delta = \nabla \cdot \nabla$ as $\Delta u = \sum_{i=1}^d \frac{\partial_i u}{\partial x_i^2}$ and note that the eigenvalues of $(I - \Delta)$ with domain $H^2(\mathbb{T}^d)$ are simply $1 + 4\pi^2|\ell|^2$, for $\ell \in \mathbb{Z}^d$. The fractional powers of $(I + \Delta)$ are defined as follows

$$(I - \Delta)^\gamma u = \sum_{\ell \in \mathbb{Z}^d} (1 + |\ell|^2)^\gamma \widehat{u}(\ell) \varphi_\ell,$$

where $\varphi_k$ are the eigenvectors of $-\Delta$ in $\mathbb{T}^d$, that form the orthonormal basis of $L^2(\mathbb{T}^d)$. We see that on the torus $H^s = \mathcal{D}((I + \Delta)^{\frac{s}{2}})$ and we have $\|u\|_{H^s} = \|(I + \Delta)^{\frac{s}{2}} u\|_{L^2}$. We also note that that $(1 - \Delta)^{-r} : H^t(\mathbb{T}^d) \to H^{t+r}(\mathbb{T}^d)$ for all $t, r \in \mathbb{R}$.

# Bibliography


[1] Y. A. Abramovich and C. D. Aliprantis, *An Invitation to Operator Theory*, Graduate Studies in Mathematics, American Mathematical Society, 2002.

[2] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[3] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, Elsevier Science, Singapore, 2003.

[4] C. D. Aliprantis and K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 2006.

[5] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Clarendon Press, 2000.

[6] A. B. Bakushinskii, *Remarks on the choice of regularization parameter from quasioptimality and relation tests*, Zhurnal Vychislitel'noï Matematiki i Matematicheskoï Fiziki, 24 (1984), pp. 1258–1259.

[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.

[8] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numerica, 27 (2018), pp. 1–111.

[9] V. I. Bogachev, *Gaussian measures*, vol. 62 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1998.

[10] B. Bollobás, *Linear Analysis: An Introductory Course*, Cambridge University Press, Cambridge, second ed., 1999.

[11] M. Burger and S. Osher, *Convergence rates of convex variational regularization*, Inverse Problems, 20 (2004), p. 1411.

[12] G. Da Prato and J. Zabczyk, *Stochastic equations in infinite dimensions*, vol. 152, Cambridge university press, 2014.

[13] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, R. Ghanem, D. Higdon, and H. Owhadi, eds., Springer International Publishing, 2016, pp. 311–428.

[14] N. Dunford and J. T. Schwartz, *Linear Operators, Part 1: General Theory*, Wiley Interscience Publishers, 1988.

[15] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, 1976.

[16] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, vol. 375, Springer Science & Business Media, 1996.

[17] L. C. Evans, *Partial differential equations*, American Mathematical Society, Providence, RI, 1998.

[18] C. W. Groetsch, *Stable approximate evaluation of unbounded operators*, Springer, 2006.

[19] M. Hairer, *An introduction to stochastic PDEs*, arXiv preprint arXiv:0907.4178, (2009).

[20] J. Hunter and B. Nachtergaele, *Applied Analysis*, World Scientific Publishing Company Incorporated, 2001.

[21] J. A. Iglesias, G. Mercier, and O. Scherzer, *A note on convergence of solutions of total variation regularized linear inverse problems*, Inverse Problems, 34 (2018), p. 055011.

[22] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[23] O. Kallenberg, *Foundations of modern probability theory*, Springer, 1997.

[24] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, Springer Science & Business Media, 2000.

[25] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.

[26] W. Rudin, *Functional Analysis*, International series in pure and applied mathematics, McGraw-Hill, 1991.

[27] K. Saxe, *Beginning Functional Analysis*, Springer, 2002.

[28] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging*, Springer, 2009.

[29] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.

[30] T. Tao, *Epsilon of Room, One*, vol. 1, American Mathematical Soc., 2010.

[31] E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, vol. 108 of Applied Mathematical Sciences Series, Springer, 1995.

[32] ——, *Applied Functional Analysis: Main Principles and Their Applications*, vol. 109 of Applied Mathematical Sciences Series, Springer, 1995.