Tan Bui-Thanh

From Bayesian Inference to Machine Learning: An Introduction

– Monograph –

November 29, 2018

Springer

Use the template dedic.tex *together with the Springer document class SVMono for monograph-type books or SVMult for contributed volumes to style a quotation or a dedication at the very beginning of your book in the Springer layout*

Foreword

Use the template *foreword.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your foreword in the Springer layout.

The foreword covers introductory remarks preceding the text of a book that are written by a *person other than the author or editor* of the book. If applicable, the foreword precedes the preface which is written by the author or editor of the book.

Place, month year *Firstname Surname*

Preface

Use the template *preface.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your preface in the Springer layout.

A preface is a book's preliminary statement, usually written by the *author or editor* of a work, which states its origin, scope, purpose, plan, and intended audience, and which sometimes includes afterthoughts and acknowledgments of assistance.

When written by a person other than the author, it is called a foreword. The preface or foreword is distinct from the introduction, which deals with the subject of the work.

Customarily *acknowledgments* are included as last part of the preface.

Place(s), *Firstname Surname*
month year *Firstname Surname*

Acknowledgements

Use the template *acknow.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) if you prefer to set your acknowledgement section as a separate chapter instead of including it as last part of your preface.

Contents

Acronyms

Use the template *acronym.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your list(s) of abbreviations or symbols in the Springer layout.

Lists of abbreviations, symbols and the like are easily formatted with the help of the Springer-enhanced `description` environment.

ABC	Spelled-out abbreviation and definition
BABI	Spelled-out abbreviation and definition
CABR	Spelled-out abbreviation and definition

Part I
Introduction

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Chapter 1
An Introduction to the Mathematical Theory of Ill-posed Problems

Abstract This chapter studies why inverse problems are ill-posed and the characterization of the ill-posedness. Many inverse problems require to invert a compact operator, and we will learn why inverting a compact operator is an ill-posed problem. The key tool is the singular value decomposition theory that allows us to precisely uncover the ill-posedness due to instability. We then cast the inverse problem into an optimization problem, which provides us an intuitive way to fix the ill-posedness. In particular, we study Tikhonov regularization method and show that it leads to well-posed optimization problems. This is a classical and deterministic way to construct well-posed inverse problems. In chapter **??**, we will learn how to form well-posed inverse problems from a statistical point of view and we then study the relation between the two. In particular we will show why the latter is more useful.

1.1 Introduction

Consider the linear operator

$$\mathscr{A} : \mathbb{X} \ni m \mapsto \mathscr{A}m = g \in \mathbb{Y},$$

where $\mathbb{X}$ and $\mathbb{Y}$ are two Hilbert spaces[1]. We are interested in solving the following problem for m

$$\mathscr{A}m = g. \tag{1.1}$$

Definition 1.1 (Well-posedness). In Hadamard's sense, (1.1) is well-posed if

1. $\mathscr{A}$ is surjective (*there exists a solution*: **existence**),
2. $\mathscr{A}$ is injective (*there is at most one solution*: **uniqueness**), and
3. $\mathscr{A}^{-1}$ is continuous (*the solution depends continuously on the data*: **stability**).

[1] Banach spaces are possible, but not considered in this note.

Definition 1.2 (Ill-posedness). Problem (1.1) is ill-posed if at least one of Hadamard's conditions is violated.

We are interested in ill-posed problems arising from inverse problems. Let us start with a "simple" example of inverse problem

Example 1.1. Consider the following quadratic equation

$$m_1 u^2 + m_2 u + m_3 = 0, \tag{1.2}$$

whose solution is

$$u = \frac{-m_2 \pm \sqrt{m_2^2 - 4m_1 m_3}}{2m_1} \stackrel{\text{def}}{=} \mathscr{A}\mathbf{m}.$$

(1.2) is called the "forward" problem, i.e., we are given the parameters m_1, m_2, m_3, and the task is is to solve for u.

The *inverse* problem reverses the process, that is, we are given, say, $u = 2$, and the task is to infer the parameters m_1, m_2, m_3. Clearly, this nonlinear inverse problem

$$\mathbf{m} = \mathscr{A}^{-1}(u)$$

Which conditions do not hold? is ill-posed since $\mathscr{A}$ violates at least one of the Hadamard's conditions.

Next, let us present a PDE-constrained inverse problem.

Example 1.2. The foward problem of interest in this case reads

$$\begin{aligned} \beta \cdot \nabla u &= 0, \quad \text{in } \Omega, \\ u &= m, \quad \text{in } \partial\Omega_{\text{in}}, \end{aligned}$$

where the transport velocity is $\beta = (1,1)$, and $\partial\Omega_{\text{in}}$ denotes the inflow boundary.

In the forward problem, we are given the inflow boundary data m, and the task is to solve for *forward state* u. Now consider the following inverse problem: given an observation of u at a finite number of points on the outflow boundary $\partial\Omega \setminus \partial\Omega_{textin}$, and the task is to infer the boundary data m. To understand why the inverse problem is ill-posded, let us denote by $\mathscr{A}$ the *parameter-to-observable* map from m to $\mathbf{x}_i, i = 1, \ldots, N^{\text{obs}}$, where $\mathbf{x}_i$ are observational points and n is the total number of observations. Clearly, the inverse problem

$$m = \mathscr{A}^{-1}\mathbf{u},$$

Why? where $\mathbf{u} \stackrel{\text{def}}{=} \{u(\mathbf{x}_1), \ldots, u(\mathbf{x}_{N^{\text{obs}}})\}$, is ill-posed.

We have seen from the above two inverse problems that the ill-posedness is due to the *parameter-to-observable* map, which is obviously not injective. *The more subtle problem is the stability*. Both of these problems depend on the definition of the domain $\mathbb{X}$ and the image $\mathbb{Y}$. To rigorously demonstrate this, let us consider the following example.

Example 1.3 (The fundamental theorem of calculus). The forward problem is given by

$$g(x) = \mathscr{A}m \stackrel{\text{def}}{=} \int_0^x m(y)\,dy, \quad 0 \le x \le 1,$$

and the inverse problem reads

$$m = \mathscr{A}^{-1} g.$$

We are going to show that, depending on $\mathbb{X}, \mathbb{Y}$, the inverse problem can be ill-posded or well-posed.

- $\mathbb{X} = \mathbb{C}[0,1], \mathbb{Y} = \mathbb{C}[0,1]$. Let us consider a perturbation around m:

  $$\tilde{m} = m + \alpha \sin(kx),$$

  which implies

  $$\tilde{g} \stackrel{\text{def}}{=} \mathscr{A}\tilde{m} = g - \frac{\alpha}{k} + \frac{\alpha}{k}\cos(kx).$$

  Clearly

  $$\|\tilde{g} - g\|_{\mathbb{C}[0,1]} \to 0, \quad \text{as } k \to \infty,$$

  but

  $$\|\tilde{m} - m\|_{\mathbb{C}[0,1]} = \alpha \quad \forall k.$$

  We conclude that $\mathscr{A}$ does not distinguish m and $\tilde{m}$, and as the result the inverse problem does not have a unique solution. In fact, $\mathscr{A}$ is a compact operator[2] and it "smoonthes" out the difference in m and $\tilde{m}$ so that observation g remains the same. Intuitively, a compact operator "squeezes" its domain into (smaller) range: for the above example $\mathscr{A}$, as an integral operator, maps $\mathbb{C}[0,1]$ into $\mathbb{C}^1[0,1] \subset \mathbb{C}[0,1]$. Since inverse of a compact operator is unbounded (more below), the stability is also violated.
- $\mathbb{X} = \mathbb{C}[0,1], \mathbb{Y} = \mathbb{C}^1[0,1]$. In this case we have

  $$\|\tilde{g} - g\|_{\mathbb{C}^1[0,1]} \stackrel{\text{def}}{=} \|\tilde{g} - g\|_{\mathbb{C}[0,1]} + \|\tilde{g}' - g'\|_{\mathbb{C}[0,1]} = \alpha, \quad \forall k,$$

  and

  $$\|\tilde{m} - m\|_{\mathbb{C}[0,1]} = \alpha \quad \forall k.$$

  As can be seen, a small change in m leads to a small (in fact the same) change in g. The inverse problem is thus stable. The uniqueness is also trivial due to the fact that $g' = m$. The surjectivity is also clear. Consequently, the inverse problem is well-posed in the Hadamard's sense.[3]

  Exercise 1.1. Again, consider the inverse of the fundamental theorem of calculus. Is the inverse problem well-posed if $\mathbb{X} = \mathbb{L}^2(0,1), \mathbb{Y} = \mathbb{L}^2(0,1)$? How about

[2] Ascoli-Arzela theorem.

[3] Note that this is an instance of the Tikhonov theorem [**?**] since $\mathbb{C}^1[0,1]$ is compactly embedded in $\mathbb{C}[0,1]$.

$\mathbb{X} = \mathbb{L}^2(0,1), \mathbb{Y} = \mathbb{H}_0^1(0,1) = \left\{ g \in \mathbb{H}^1(0,1) : g(1) = 0 \right\}$? **Hint**: Redo Example 1.3 with the following corresponding norms

$$\|f\|_{\mathbb{L}^2(0,1)} \stackrel{\text{def}}{=} \sqrt{\int_0^1 [f(x)]^2 \, dx},$$

$$\|f\|_{\mathbb{H}^1(0,1)} \stackrel{\text{def}}{=} \sqrt{\int_0^1 [f(x)]^2 \, dx + \int_0^1 [f'(x)]^2 \, dx}$$

1.2 Ill-posedness of inverting a compact operator

We have seen that the ill-posedness of inverse problem seems to be due to the compactness of the the parameter-to-observable map, that is, *it can be not injective (multiple solutions) and it is not stable (solutions do not depend continuously on the data)*. Indeed, one can show, for many practial inverse problems, that the parameter-to-observable map (or its Fréchet derivative) is compact [8, 7, 9]. Thus, in this section we will study why inverting $\mathscr{A}$ when it is a compact operator is not a well-posed problem. To begin, we recap some results from functional analysis [22, 3]. We denote by $\mathbb{C}(\mathbb{X},\mathbb{X})$ the space of all compact operators from $\mathbb{X}$ to $\mathbb{X}$. *A direct consequence of a compact operator is that its eigenvalues decay to zero*.

Theorem 1.1 (Hilbert-Schmidt theorem for self-adjoint compact operators). *Let $\mathscr{B} \in \mathbb{C}(\mathbb{X},\mathbb{X})$, and $\mathscr{B} = \mathscr{B}^*$, i.e. $\mathscr{B}$ is self-adjoint*[4]*. Then there exists an orthonormal set of eigen-functions φ_i corresponding to non-zero eigenvalues λ_i of $\mathscr{B}$ such that for any $x \in \mathbb{X}$ we have a unique expansion of the form*

$$x = \sum_i \alpha_i \varphi_i + \mathscr{P} x,$$

where $\mathscr{P}$ is an orthogonal projection from $\mathbb{X}$ to the nullspace of $\mathscr{B}$, i.e., $N(\mathscr{B})$. Here,

$$\alpha_i = (\varphi_i, x)_{\mathbb{X}},$$

where $(\cdot,\cdot)_{\mathbb{X}}$ is the inner product in $\mathbb{X}$.

Furthermore, we have

$$\mathscr{B} x = \sum_i \lambda_i (\varphi_i, x)_{\mathbb{X}} \varphi_i.$$

Now let $\mathscr{A} \in \mathbb{C}(\mathbb{X},\mathbb{Y})$, we know [22, 3] that $\mathscr{B} \stackrel{\text{def}}{=} \mathscr{A}^*\mathscr{A} \in \mathbb{C}(\mathbb{X},\mathbb{X})$ is a self-adjoint operator. Hilbert-Schmidt theorem 1.1 says that there exists orthonormal eigenfunctions φ_i corresponding to nonzero eigenvalues λ_i of $\mathscr{B}$ such that

[4] Recall for an operator $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ its adjoint is defined as

$$(\mathscr{A}x, y)_{\mathbb{Y}} = (x, \mathscr{A}^* y)_{\mathbb{X}}, \quad \forall x \in \mathbb{X}, y \in \mathbb{Y}.$$

A more general definition of adjoint operators is presented in Chapter 2.

$$\mathscr{B}\varphi_i = \mathscr{A}^*\mathscr{A}\varphi_i = \lambda_i\varphi_i,$$

which implies

$$(\mathscr{A}^*\mathscr{A}\varphi_i, \varphi_i)_{\mathbb{X}} = \lambda_i (\varphi_i, \varphi_i)_{\mathbb{X}},$$

which in turn can be written as

$$\|\mathscr{A}\varphi_i\|_{\mathbb{X}}^2 = \lambda_i \|\varphi_i\|_{\mathbb{X}}^2,$$

where we have defined the induced norm $\|\cdot\|_{\mathbb{X}}$ from the inner product in $\mathbb{X}$. The above result shows that $\lambda_i \geq 0$. Let us define the *singular value* μ_i of $\mathscr{A}$ as

$$\mu_i \stackrel{\text{def}}{=} \sqrt{\lambda_i}.$$

We are now in the position to study the singular value decomposition for compact operators [12].

Theorem 1.2 (Singular value decomposition). *Let $\{\mu_i\}$ be the sequence of non-zero singular values of a compact operator $\mathscr{A} \in \mathbb{C}(\mathbb{X}, \mathbb{Y})$ and be ordered as*

$$\mu_1 \geq \mu_2 \geq \dots,$$

then there exist two orthonormal sequences $\{\varphi_i\}$ and $\{\phi_i\}$ such that

1. *$\mathscr{A}\varphi_i = \mu_i\phi_i$ and $\mathscr{A}^*\phi_i = \mu_i\varphi_i$.*
2. *$\forall \varphi \in \mathbb{X}$, we have*
$$\varphi = \sum (\varphi, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P}\varphi,$$
$\mathscr{P} : \mathbb{X} \to N(\mathscr{A})$ is an orthogonal projection.
3. *There holds*
$$\mathscr{A}\varphi = \sum \mu_i (\varphi, \varphi_i)_{\mathbb{X}} \phi_i.$$

We call $(\mu_i, \varphi_i, \phi_i)$ **the singular system** *of $\mathscr{A}$.*

Proof. The key that we exploit is the Hilbert-Schmidt theorem 1.1.

1. By theorem 1.1 we know that
$$\mathscr{A}^*\mathscr{A}\varphi_i = \mu_i^2\varphi_i,$$
where $\{\varphi_i\}$ is an orthonormal set. Let us define
$$\mu_i\phi_i \stackrel{\text{def}}{=} \mathscr{A}\varphi_i.$$
Then, $\{\phi_i\}$ is also an orthonormal set. By definition we have *see Exercise 1.2.*
$$\mathscr{A}^*\phi_i = \frac{1}{\mu_i}\mathscr{A}^*\mathscr{A}\varphi_i = \mu_i\varphi_i.$$
2. Again, by theorem 1.1 we have

$$\forall \varphi \in \mathbb{X}: \quad \varphi = \sum (\varphi, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P}\varphi,$$

where $\mathscr{P}$ is an orthonormal projection from $\mathbb{X}$ to $N(\mathscr{A}^*\mathscr{A})$, but $N(\mathscr{A}) = N(\mathscr{A}^*\mathscr{A})$.

See Exercise 1.3.

3. We start with the partial sum

$$s_N \stackrel{\text{def}}{=} \sum_{i=1}^{N} (\varphi, \varphi_i)_{\mathbb{X}} \varphi_i,$$

and we have

$$\mathscr{A} s_N = \sum_{i=1}^{N} \mu_i (\varphi, \varphi_i)_{\mathbb{X}} \phi_i.$$

Do you see the second equality?

Now taking the limit of the left hand side gives

$$\lim_{N\to\infty} \mathscr{A} s_N = \mathscr{A}(\varphi - \mathscr{P}\varphi) = \mathscr{A}\varphi.$$

Consequently,

$$\mathscr{A}\varphi = \sum \mu_i (\varphi, \varphi_i)_{\mathbb{X}} \phi_i.$$

□

Exercise 1.2. Show that $\{\phi_i\}$ is an orthonormal set.

Exercise 1.3. Show that $N(\mathscr{A}) = N(\mathscr{A}^*\mathscr{A})$.

The next theorem [12], due to Picard, tells us the conditions under which inverting a compact operator is well-defined.

Theorem 1.3 (Picard). *Suppose $\mathscr{A} \in \mathbb{C}(\mathbb{X}, \mathbb{Y})$. The equation*

$$\mathscr{A}\varphi = g$$

is solvable iff

i) $g \in N(\mathscr{A}^*)^{\perp}$, *and*[5]

ii) $\sum \frac{1}{\mu_i^2} |(g, \phi_i)_{\mathbb{Y}}|^2 < \infty$,

where $(\mu_i, \varphi_i, \phi_i)$ is a singular system of $\mathscr{A}$. In this case the solution is given by

$$\varphi = \sum \frac{1}{\mu_i} (g, \phi_i)_{\mathbb{Y}} \varphi_i.$$

Proof. The SVD theorem 1.2 provides a simple proof for this theorem.

[5] Note that $\perp$ denotes the orthogonal complement of a set. For example $N(\mathscr{A}^*)^{\perp} = \{y \in \mathbb{Y} : (y, w)_{\mathbb{Y}} = 0, \quad \forall w \in N(\mathscr{A}^*)\}$.

$\Rightarrow$ Solvability implies that g belongs to the range of $\mathscr{A}$, i.e. $g \in R(\mathscr{A})$. From the closed range theorem[6] [22, 3] we know that $R(\mathscr{A}) \subset \overline{R(\mathscr{A})} = N(\mathscr{A}^*)^\perp$ (see Theorem 2.4 of Chapter 2), and hence *i)* holds. On the other hand, we have

$$\varphi = \sum (\varphi, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P}\varphi,$$

which implies

$$\|\varphi\|_{\mathbb{X}}^2 = \sum |(\varphi, \varphi_i)_{\mathbb{X}}|^2 + \|\mathscr{P}\varphi\|_{\mathbb{X}}^2,$$

which in turns implies

$$\sum |(\varphi, \varphi_i)_{\mathbb{X}}|^2 \leq \|\varphi\|_{\mathbb{X}}^2 < \infty.$$

But

$$(\varphi, \varphi_i)_{\mathbb{X}} = \frac{1}{\mu_i} (\varphi, \mathscr{A}^* \phi_i)_{\mathbb{X}} = \frac{1}{\mu_i} (\mathscr{A}\varphi, \phi_i)_{\mathbb{Y}} = \frac{1}{\mu_i} (g, \phi_i)_{\mathbb{Y}}.$$

Thus *ii)* holds.

$\Leftarrow$ Since $g \in N(\mathscr{A}^*)^\perp$, Hilbert-Schmidt theorem gives *Why?*

$$g = \sum (g, \phi_i)_{\mathbb{Y}} \phi_i.$$

Now, from *ii)* we can define

$$\varphi = \sum \frac{1}{\mu_i} (g, \phi_i)_{\mathbb{Y}} \varphi_i.$$

Whence,

$$\mathscr{A}\varphi = \sum \frac{1}{\mu_i} (g, \phi_i)_{\mathbb{Y}} \mathscr{A}\varphi_i = \sum (g, \phi_i)_{\mathbb{Y}} \phi_i = g,$$

and this concludes the proof.

□

We now discuss some important consequences of the Picard theorem 1.3 which are the main goals of this chapter. We observe that

$$g = \mathscr{A}\varphi = \sum \mu_i (\varphi, \phi_i)_{\mathbb{X}} \phi_i,$$

where we have used the second assertion of theorem 1.2, i.e.,

$$\varphi = \sum (\varphi, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P}\varphi.$$

[6] Note that $R(\mathscr{A})$ cannot be closed since $\mathscr{A}$ is compact. Assume, on the contrary, it is, then by the bounded inverse theorem [25] we know that $\mathscr{A}^{-1}$ is continuous and hence $\mathrm{i} = \mathscr{A}^{-1}\mathscr{A}$ is also a compact operator. But, this is a contradiction since identity operator in infinite dimensional space cannot be a compact operator [22, 3]. If $R(\mathscr{A})$ is closed, then *i)* is both necessary and sufficient. Sine this is not true for compact operators, we have to replace the closedness by the smooth property *ii)* of the right hand side.

Since $\mu_i \to 0$ as $i \to \infty$, $\mathscr{A}$ smoothes out the contribution from φ_i for large i. In other words, the output g is not sensitive to the components of φ in φ_i when $i \to \infty$.

Conversely, let us perturb the right hand side g as

$$\tilde{g} = g + \delta\phi_N,$$

why? then the corresponding solution reads

$$\tilde{\varphi} = \sum \frac{1}{\mu_i} (\tilde{g}, \phi_i)_{\mathbb{Y}} \varphi_i + \mathbb{P}(\varphi) = \varphi + \frac{\delta}{\mu_N} \varphi_N.$$

Thus,

$$\frac{\|\tilde{\varphi} - \varphi\|_{\mathbb{X}}}{\|\tilde{g} - g\|_{\mathbb{Y}}} = \frac{1}{\mu_N} \to \infty,$$

which is exactly the subtle instability problem of inverting a compact operator, namely, small changes in the input can lead to very large change in the solution.

Exercise 1.4. let us consider the following one dimensional deblurring (deconvolution) problem

$$g(s_j) = \int_0^1 a(s_j,t) f(t)\, dt + e(s_j), \quad 0 \le j \le N^{\text{obs}},$$

where $a(s,t) = \dfrac{1}{\sqrt{2\pi\beta^2}} \exp(-\dfrac{1}{2\beta^2}(t-s)^2)$ is a given kernel, and $s_j = j/N^{\text{obs}}$, $j = 0, \ldots, N^{\text{obs}}$ the mesh points. Our task is to reconstruct $f(t) : [0,1] \to \mathbb{R}$ from the noisy observations $g(s_j)$, $j = 0, \ldots, N^{\text{obs}}$. To cast the function reconstruction problem, which is in infinite dimensional space, into a reconstruction problem in $\mathbb{R}^{N^{\text{obs}}}$, we discretize $f(t)$ on the same mesh and use simple rectangle method for the integral. Let us define $y^{obs} = [g(s_0), \ldots, g(s_{N^{\text{obs}}})]^T$, $m = (f(s_0), \ldots, f(s_{N^{\text{obs}}}))^T$, and $\mathscr{A}_{i,j} = a(s_i, s_j)/N^{\text{obs}}$, then the discrete deconvolution problem reads

$$y^{obs} = \mathscr{A} m + e.$$

Let's pick the "exact" synthetic truth as $f(t) = sin(2\pi t)$. From this truth, compute the synthetic observation without noise, i.e. ignoring e from the above equation. Now **numerically solve** for $m = \mathscr{A}^{-1} y^{obs}$ using the synthetic observation with, say, $N^{\text{obs}} = 100$ and $\beta = 0.1$. What do you see? What if β is very small, i.e. $\beta = 10^{-10}$? What do you see? What if β is large, i.e. $\beta = 10$? Based on what we discussed above, can you explain why?

Repeat the computation with $e \sim \mathscr{N}(0, \sigma^2 I)$, i.e. generating the synthetic observation with the above equation, where I is the identity matrix in $\mathbb{R}^{(N^{\text{obs}}+1)\times(N^{\text{obs}}+1)}$. Here, the noise standard deviation σ is taken to be the 5% of the maximum value of $f(s)$, i.e. $\sigma = 0.05 \max_{s\in[0,1]} |f(s)|$. Do you get better results or not? Why?

We have seen that there could be multiple solutions to the linear inverse problem of interest (1.1) (or there is none). The main reason is that the nullspace of $\mathscr{A}$ is non-trivial (or g is not in the range of $\mathscr{A}$). The question is which solution is the most "useful"? One way to address this question is to look for the solution that optimizes some quantity of interest. In many practical setting, the quantity of interest is some norm and the task is typically to solve for a minimum norm solution. That is, *we need to solve an optimization to find the most useful solution*, e.g.,

$$\min_m \frac{1}{2} \|\mathscr{A}m - g\|_{\mathbb{X}}^2. \tag{1.3}$$

However, the ill-conditioning nature of our inverse problem does not go away. To see this let us recall the following important theorem [19].

Theorem 1.4. *Let $\mathbb{X}$ be a linear vector space endowed with an inner product and the induced norm $\|\cdot\|_{\mathbb{X}}$, and $\mathbb{M}$ be a subspace of $\mathbb{X}$. Then for any $x \in \mathbb{X}$, if there is a vector $x_0 \in \mathbb{M}$ such that $\|x - x_0\|_{\mathbb{X}} \leq \|x - y\|_{\mathbb{X}}$ for all $y \in \mathbb{M}$, then x_0 is unique. A necessary and sufficient condition that $x_0 \in \mathbb{M}$ be a unique minimizing vector in $\mathbb{M}$ is that the error vector $x - x_0$ is orthogonal to $\mathbb{M}$.*

Lemma 1.1. *m is the solution of the optimization problem* (1.3) *iff*

$$\mathscr{A}^*\mathscr{A}m = \mathscr{A}^*g. \tag{1.4}$$

Proof. The problem (1.3) is equivalent to minimizing $\|g - y\|$ where $y \in R(\mathscr{A})$. Since $R(A)$ is a subspace of $\mathbb{X}$, from theorem 1.4 we know that y is the minimizing vector iff $(g - y) \in R(\mathscr{A})^{\perp} = N(\mathscr{A}^*)$. That is

$$0 = \mathscr{A}^*(g - y) = \mathscr{A}^*(g - \mathscr{A}m),$$

and this ends the proof. □

Lemma 1.1 shows that though we can write down the equation for a minimizer m, we cannot solve for it since $\mathscr{A}$ is compact and so is $\mathscr{A}^*\mathscr{A}$. In other words, we still have problem with uniqueness if $N(A)$ is not empty, and (bigger) problem with instability due to inverting the compact operator $\mathscr{A}^*\mathscr{A}$. Thus, recasting the original (e.g. $\mathscr{A}$ is compact) problem (1.1) into an optimization problem (1.3) does not seem resolve the illposedness. However, the optimization idea paves the way for using optimization technique to overcome the problem. Clearly, if the cost function in (1.3) is a "nice parabola", then the minimizer is unique. This immediately suggests that one should add a quadratic term to the cost function to make it more like a parabola, and hence removing the uniqueness issue (as will be shown, this also addresses the stability). This is essentially the idea behind the *Tikhonov regularization*, which proposes to solve the nearby problem

$$\min_m \frac{1}{2} \|\mathscr{A}m - g\|_{\mathbb{X}}^2 + \frac{\kappa}{2} \|m - m_0\|_{\mathbb{X}}^2, \tag{1.5}$$

where m_0 is some "prior" reference function and κ is known as the *regularization parameter*. To show that the *regularized optimization* problem (1.5) is well-posed, we need the following projection theorem [19] and a key result from the Riesz-Fredholm theory [11].

Lemma 1.2 (The classical projection theorem). *Let $\mathbb{X}$ be a Hilbert space and $\mathbb{Y}$ be a* closed subspace *of $\mathbb{X}$. For any $x \in \mathbb{X}$, there exist a unique vector $y_0 \in \mathbb{Y}$ such that*

$$\|x - y_0\|_{\mathbb{X}} \leq \|x - y\|_{\mathbb{X}}, \quad \forall y \in \mathbb{Y}.$$

Moreover, the necessary and sufficient condition that $y_0 \in \mathbb{Y}$ be the unique minimizing vector is that $x - y_0$ be orthogonal to $\mathbb{Y}$.

Lemma 1.3. *Let $\mathscr{A}$ be a compact operator from $\mathbb{X}$ to $\mathbb{X}$. If $(I + \mathscr{A})$ is injective, then $(I + \mathscr{A})$ is continuously invertible.*

Theorem 1.5. *For any $\kappa > 0$, the regularized optimization problem* (1.5) *is well-posed.*

Proof. Without loss of generality, assume $\kappa = 1$. We begin by rewrite the optimization (1.5) as

$$\min_m \frac{1}{2} \|\mathscr{B}m - z\|^2_{\mathbb{X}\times\mathbb{X}}, \tag{1.6}$$

where we have defined $\mathscr{B} : \mathbb{X} \ni m \mapsto (\mathscr{A}m, m) \in \mathbb{X} \times \mathbb{X}$, and $z := (z_1, z_2) := (g, m_0)$. The intrinsic norm of $z \in \mathbb{X} \times \mathbb{X}$ is defined as $\|y\|^2_{\mathbb{X}\times\mathbb{X}} := \|y_1\|^2_{\mathbb{X}} + \|y_2\|^2_{\mathbb{X}}$. Though $R(\mathscr{A})$ is not closed (see the discussion in the proof of Theorem 1.3), $R(B)$ is closed as we now show. Let $z^n \in R(\mathscr{B})$ and assume that $z^n \to z$, we need to show that $z \in R(\mathscr{B})$. Clearly, there exists x^n such that $z^n := (z^n_1, z^n_2) = \mathscr{B}x^n = (\mathscr{A}x^n, x^n)$. It follows that $x^n \to z_2$ and $\mathscr{A}x^n \to z_1$, which in turns implies $\mathscr{A}z_2 = z_1$ due to the continuity of $\mathscr{A}$. In other words, $z = (\mathscr{A}z_2, z_2) \in R(\mathscr{B})$.

Do you see that?

Now, a simple application of Theorem 1.2 with $\mathbb{M} = R(\mathscr{B})$ concludes that there exists a unique minimizer for (1.6), and hence (1.5). Next from Lemma 1.1 we know that the minimizer satisfies

$$\mathscr{B}^*\mathscr{B}m = \mathscr{B}^*z,$$

which is equivalent to

$$(\mathscr{A}^*\mathscr{A} + I)\, m = \mathscr{A}^*g + m_0.$$

See Exercise 1.5.

Since $(\mathscr{A}^*\mathscr{A} + I)$ is injective, Lemma 1.3 shows that it is continuously invertible, i.e. $\left\|(\mathscr{A}^*\mathscr{A} + I)^{-1}\right\| < \infty$. Hence,

$$\|m\|_{\mathbb{X}} = \left\|(\mathscr{A}^*\mathscr{A} + I)^{-1}(\mathscr{A}^*g + m_0)\right\| \leq \alpha \|g\|_{\mathbb{X}} + \beta \|m_0\|_{\mathbb{X}},$$

where α and β are some finite positive constants. This ends the proof. □

Exercise 1.5. Show that $(\mathscr{A}^*\mathscr{A} + I)$ is injective.

Exercise 1.6. Back to the deblurring problem in Exercise 1.4 with noise-corrupted observation (the second part of Problem 1.4), but now with $\beta = 0.2$ and $N^{\text{obs}} = 100$, and hence the new synthetic observation.

1. Write the continuous problem in the form $\mathscr{A}m = g$, and identify $\mathscr{A}$ and g
2. Consider the following Tikhonov regularization

$$\min_m \frac{1}{2}\|\mathscr{A}m - g\|_{\mathbb{X}}^2 + \frac{\kappa}{2}\|\nabla m\|_{\mathbb{L}^2(0,1)}^2$$

   Then discretize the integrals with the same rule in Exercise 1.4, and discretize the derivative with any finite difference scheme (forward or backward or central). Then find the discrete minimizer of the regularized optimization problem for a regularization parameter $\kappa = 1$. Plot the inversion result with the synthetic truth.
3. In practice, we can find the "best" regularization parameter by noticing that on the one hand we like to minimize the *misfit*, i.e. $\|\mathscr{A}m - g\|_{\mathbb{X}}^2$, and on the other hand we want to regularize the solution by not making the regularization (the second part of the regularized cost function without κ, i.e. $\|\nabla m\|_{\mathbb{L}^2(0,1)}^2$) large. A minimizer of the regularized problem compromises both of these two goals by not only making each term as small as possible but also making the sum of them smallest. By plotting the misfit as a function of the regularization as κ varies, we can approximate the "best" regularization paramter κ. Determine this value. **Hint**: compute the inverse solutions $m(\kappa)^*$ for many values of κ, and then evaluate both the misfit and the regularization at these inverse solutions.
4. Another regularization strategy is by truncation. That is we solve the normal equation (1.4) directly using some iterative method such as conjugate gradient, and then stop early when the approximate solution makes the misfit comparable to the noise, i.e.,

$$\|\mathscr{A}\tilde{m} - g\|^2 \approx (N^{\text{obs}} + 1)\sigma^2,$$

   that is we don't want to overfit the noise. This stopping criteria is known as the *Morozov's discrepancy principle*. Write a program to determine a disrepancy-satisfying solution. Compare the solution with the Tikhonov solution corresponding to the "best" κ.

1.3 Appendix

This section presents some of the mathematical background in more details: necessary compact operator that is useful for the text: definitions and some of the properties and proof.

Chapter 2
Optimization

Abstract We have seen from chapter 1 that there could be multiple solutions to the linear inverse problem of interest (1.1). The main reason is that the nullspace of $\mathscr{A}$ is non-trivial. The question is which solution is the most "useful"? One way to address this question is to look for the solution that optimizes some quantity of interest. In many practical setting, the quantity of interest is some norm and the task is typically to solve for a minimum norm solution. That is, *we need to solve an optimization to find the most useful solution.* In this chapter, we provide necessary tools to study both (un-)constrained linear and nonlinear optimization problems. Of interest are the definition and derivation of the derivatives in function spaces that are needed for advanced Markov chain Monte Carlo methods discussed in chapter **??**.

2.1 Optimization of functions over $\mathbb{R}^N$

Let $f:\mathbb{R}^N \ni x \mapsto f(x) \in \mathbb{R}$ and we are interested in studying the optimization problem $\min_{x\in\mathbb{R}^N} f(x)$. It is sufficient to consider the case $N=1$ (see Exercise 2.3).

Definition 2.1. x_0 is a minimizer of $f(x)$ if there exists a open neighborhood, i.e. $B_\delta(x_0) \stackrel{\text{def}}{=} \{x : |x-x_0| < \delta\}$ (ball with radius δ), for some small δ, such that

$$f(x) \geq f(x_0), \quad \forall x \in B_\delta(x_0).$$

Now Taylor series expansion about x_0 states that there exists $0<\theta<1$ such that

$$f(x_0+\varepsilon) = f(x_0) + \varepsilon \underbrace{f'(x_0)}_{\text{gradient } g(x_0)} + \frac{1}{2}\varepsilon^2 \underbrace{f''(x_0+\theta\varepsilon)}_{\text{Hessian } H(x_0+\theta\varepsilon)}.$$

If x_0 is a minimizer, what can we say about the gradient $g(\cdot)$ and the Hessian $H(\cdot)$ at x_0? Here is the answer.

Lemma 2.1 (Necessary condition for optimality). *Suppose $f(x)$ is twice continuously differentiable in a neighborhood of a minimizer x_0. It is necessary that*

i) The gradient vanishes, i.e., $g(x_0) = 0$, and
ii) the Hessian is non-negative, i.e., $H(x_0) \geq 0$.

Proof. The proof is by contradiction. For the first assertion, we considering $g(x_0) > 0$ and together with "sufficiently small" ε leads to a contradiction. A similar contradiction can be done for the case $g(x_0) < 0$. For the second assertion, we start with

$$f(x_0 + \varepsilon) = f(x_0) + \frac{1}{2}\varepsilon^2 H(x_0 + \theta\varepsilon),$$

and assume that $H(x_0) < 0$. Due to the continuity of $H(x)$, for sufficiently small ε we have $H(x_0 + \theta\varepsilon) < 0$. Thus, $f(x) < f(x_0)$, a contradiction. □

Exercise 2.1. Prove the first assertion of Lemma 2.1 in details. •

Exercise 2.2 (Sufficient condition for optimality). Suppose $f(x)$ is twice continuously differentiable in a neighborhood of x_0. If

i) The gradient vanishes, i.e., $g(x_0) = 0$, and
ii) the Hessian is non-negative, i.e., $H(x_0) > 0$,

then show that x_0 is a (local) minimizer. •

Exercise 2.3. Derive both first- and second-order optimality conditions for $\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})$. Note that an open neighborhood (ball) with radius δ around $\mathbf{x}_0$ can be defined as $B_\delta(\mathbf{x}_0) \stackrel{\text{def}}{=} \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\|_{\mathbb{R}^N} < \delta\}$. Here the norm in $\mathbb{R}^N$ is the standard Euclidean norm. •

2.2 Optimization of functionals

The object of interest in this section is function of functions residing in a function spaces, i.e. *functional*:

$$f : \mathbb{X} \ni u \mapsto f(u) \in \mathbb{R}.$$

The classical derivatives that are used in section 2.1 are not well-defined in this case, and this asks for an extension of derivatives in function spaces. Though there are many extensions in the literature, let us focus on the Fréchet derivative.

Definition 2.2. Suppose, for any $h \in \mathbb{X}$,

$$\mathscr{D}f(u,h) = \left.\frac{df}{dt}(u+th)\right|_{t=0} = \lim_{t\to 0}\frac{f(u+th) - f(u)}{t}$$

exists and converges *uniformly* in h. Then $\mathscr{D}f(u,\cdot)$ is called the Fréchet derivative of the functional $f(\cdot)$ at u if it is *linear* and *continuous (bounded)* with respect to h. In this case, *we say the function is Fréchet differentiable*.

Note that the Fréchet derivative is a linear functional from $\mathbb{X}$ to $\mathbb{R}$, i.e.,

$$\mathscr{D}f(u,\cdot):\mathbb{X}\ni h\mapsto \mathscr{D}f(u,h)\in\mathbb{R}.$$

Exercise 2.4. Show that

$$\lim_{\|h\|_{\mathbb{X}}\to 0}\frac{|f(u+h)-f(u)-\mathscr{D}f(u,h)|}{\|h\|_{\mathbb{X}}}=0.$$

In other words, this result is equivalent to

$$f(u+h)-f(u)-\mathscr{D}f(u,h)=o(\|h\|_{\mathbb{X}}).$$

•

Exercise 2.5. Show that Fréchet derivative is unique. •

Exercise 2.6. Show that if f is Fréchet defferentiable at u, then it is continuous[1] at u. •

Theorem 2.1 (Riesz representation theorem). *Let f be a linear and bounded functional on a Hilbert space $\mathbb{X}$, i.e.,*

$$\begin{aligned} f:\mathbb{X}\ni u\mapsto f(u)\in\mathbb{R},\\ f(\alpha u+\beta v)=\alpha f(u)+\beta f(v),\\ f(u)\le c\|u\|_{\mathbb{X}},\quad 0<c<\infty. \end{aligned}$$

There exists a unique $u\in\mathbb{X}$ such that

$$f(v)=(u,v)_{\mathbb{X}},\quad \forall v\in\mathbb{X}.$$

Furthermore $\|f\|\overset{def}{=}\sup_{v\in\mathbb{X}}\dfrac{|f(v)|}{\|v\|_{\mathbb{X}}}=\|u\|_{\mathbb{X}}$.

Definition 2.3. The gradient of $f(\cdot)$ at u, $g(u)=\nabla f(u)$, is defined as a function in $\mathbb{X}$ such that

$$(\nabla f(u),h)_{\mathbb{X}}=\mathscr{D}f(u,h),\quad \forall h\in\mathbb{X}.$$

That is, the gradient is defined as the Riesz representation of the Fréchet derivative.

Exercise 2.7. Consider $\mathbb{X}\equiv(\mathbb{R}^n,\mathbf{A})$, where $\mathbb{R}^n$ is endowed with a weighted inner product $(\mathbf{x},\mathbf{y})_{(\mathbb{R}^n,\mathbf{A})}=\mathbf{x}^T\mathbf{A}\mathbf{y}$ and $\mathbf{A}$ is a symmetric positive definite matrix. Suppose

[1] Recall that f is continuous at u if, $\forall\varepsilon>0$, there exist $\delta>0$ such that

$$\|v-u\|_{\mathbb{X}}<\delta\Rightarrow|f(v)-f(u)|<\varepsilon.$$

that the (usual) partial derivatives of the cost function are continuous. Show that the Fréchet derivative can be written as

$$\mathscr{D}f(\mathbf{x},\mathbf{h}) = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i} h_i.$$

From this, deduce that

$$\nabla f(\mathbf{x}) = \mathbf{A}^{-1} \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_N} \right)^T.$$

We observe that the Fréchet derivative is a generalization of the directional derivative, and the usual gradient vector is in fact the Reisz representation of the Fréchet derivative in the standard inner product of $\mathbb{R}^n$. •

Exercise 2.8. Let $\mathbb{X} = \mathbb{L}^2(0,1)$ with a weighted inner product $(x,y)_{(\mathbb{L}^2,w)} = \int_0^1 x(t)\,y(t)\,w(t)\,dt$, where $w(t)$ is a positive function, and $f(x) = \int_0^1 g(x,t)\,dt$. Assume that $\dfrac{\partial g}{\partial x}$ exists and is continuous with respect to x and t. Derive the expression for the Fréchet derivative at x acting on h and then obtain the corresponding gradient. •

Exercise 2.9. Same as in Exercise 2.8 but now $\mathbb{X} = \mathbb{H}_0^1(0,1)$ with a weighted inner product $(x,y)_{(\mathbb{H}^1,w)} = \int_0^1 \left[x(t)\,y(t) + wx'(t)\,y'(t) \right] dt$, where $w(t)$ is a non-negative function. **Hint***: use the fact that* $\mathbb{H}_0^1(0,1)$ *is dense in* $L^2(0,1)$ *and hence*

$$\int u(t)v(t)\,dt = 0, \quad \forall v(t) \in \mathbb{H}_0^1(0,1)$$

implies that $u(t) = 0$. •

Lemma 2.2. *Suppose that* $f : \mathbb{X} \to \mathbb{R}$ *attains its extremum at u. Then it is necessary*[2] *that*

$$\mathscr{D}f(u,h) = 0, \quad \forall h \in \mathbb{X}.$$

Proof. It is sufficient to assume that f is minimized at u, i.e.,

$$f(v) \geq f(u), \quad \forall v \in B_\delta(u),$$

which implies that for any h such that $\varepsilon \|h\|_{\mathbb{X}} < \delta$, we have

$$f(u + \varepsilon h) \geq f(u).$$

If we define $F(\varepsilon) \stackrel{\text{def}}{=} f(u + \varepsilon h)$, then $F(\cdot)$ is an ordinary function, namely, $F : \mathbb{R} \to R$. Furthermore, $F(\cdot)$ attains the minimum at $\varepsilon = 0$. Thus, from the first result of Lemma 2.1, we have

[2] The second order necessary conditions for functionals are much more involved and will not be discussed here.

$$\left.\frac{dF}{d\varepsilon}\right|_{\varepsilon} = 0,$$

but this is equivalent to $\mathscr{D}f(u,h) = 0$ by definition. □

Note that by definition the equivalent necesssary condition is

$$\nabla f(u) = 0.$$

Exercise 2.10 (Euler-Lagrange Equations). Consider $f : \mathbb{L}^2(t_1,t_2) \to \mathbb{R}$ defined as

$$f(u) = \int_{t_1}^{t_2} J\left(u,u',t\right) dt.$$

Assume that f attains its extremum at u. Derive the necessary condition for the gradient $\nabla f(u)$. *This condition is in fact called the Euler-Lagrange equation.* •

Of course Fréchet derivative can be directly generalized to mappings between two different function spaces. For example, if $c : \mathbb{X} \ni u \mapsto c(u) \in \mathbb{Y}$, then the Fréchet derivative $\mathscr{D}c(u)$, when it exists, can be defined as

$$\mathscr{D}c(u,h) = \lim_{t\to 0} \frac{c(u+th) - c(u)}{t}.$$

The difference is now that $\mathscr{D}c(u)$ is a linear and bounded mapping from $\mathbb{X}$ to $\mathbb{Y}$, that is, $\mathscr{D}c(u) \in \mathscr{B}(\mathbb{X},\mathbb{Y})$.

Up to this point we have looked at unconstrained optimization problem and derive the (first order) necessary condition for optimality. We next discuss optimality conditions for constrained optimization. Let us consider the following constrained optimization problem

$$\min_{u\in\mathbb{X}} f(u), \qquad \text{subject to } \; c(u) = 0, \text{ where } c(\cdot) : \mathbb{X} \to \mathbb{Y}.$$

If there were no constraint, then the optimality condition would be

$$\mathscr{D}f(u,h) = 0, \quad \forall h \in \mathbb{X},$$

That is, the variation of f in any "direction" h vanishes. However, h can be no longer arbitrary since the contraint must still hold at $u+th$ for any small t. In other words, $u+th$ needs to be *feasible*, i.e.,

$$c(u+th) = 0.$$

It is therefore necessary that

$$\mathscr{D}c(u,h) = 0.$$

To rigorously establish this result. We need the inverse function theorem.

Theorem 2.2 (Inverse function theorem). *Let $g : \mathbb{X} \to \mathbb{Y}$. Assume that $\mathscr{D}g(u_0)$ exists and maps $\mathbb{X}$* **onto** *$\mathbb{Y}$. Then, there is a neighborhood $B_\delta(y_0)$ of $y_0 = g(u_0)$*

such that $g(u) = y$ has a solution u for every $y \in B_\delta(y_0)$ and the solution satisfies $\|u-u_0\|_{\mathbb{X}} \leq K\|y-y_0\|_{\mathbb{Y}}$, for some positive bounded constant K.

Lemma 2.3 (First order necessary condition). *Suppose f attains its extremum at u_0 subject to the constraint $c(u) = 0$. Assume that both f and c are continuously Fréchet differentiable in an open set containing u_0 and $\mathscr{D}c(u_o)$ maps $\mathbb{X}$ onto $\mathbb{Y}$. Then it is necessary that*

$$\mathscr{D}f(u_0,h) = 0, \forall h \text{ such that } \mathscr{D}c(u_0,h) = 0.$$

Proof. Without lost of generality, asssume u_0 is a minimizer. We proceed by contraction. Let us consider the transformation $g(u) = (f(u), c(u)) : \mathbb{X} \to \mathbb{R} \times \mathbb{Y}$. Assume that there exists h such that $\mathscr{D}c(u_0,h) = 0$ but $\mathscr{D}f(u_0,h) \neq 0$. Then $(\mathscr{D}f(u_0), \mathscr{D}c(u_0))$ maps $\mathbb{X}$ onto $\mathbb{R} \times \mathbb{Y}$ since $\mathscr{D}c(u_0)$ maps $\mathbb{X}$ onto $\mathbb{Y}$. By the inverse function theorem that there exists ε and u with $\|u-u_0\|_{\mathbb{X}} < \varepsilon$ such that $g(u) = (f(u_0) - \delta, 0)$ for some small $\delta > 0$. This implies $f(u) < f(u_0)$, a contradiction. □

To make the necessary optimality condition practical for large-scale computation we need a Lagrangian formalism, and this requires a few extra mathematical tools. We begin with the notion of dual space. We denote by $\mathbb{X}^*$ the dual space of a Hilbert space $\mathbb{X}$. Objects in $\mathbb{X}^*$ are linear and bounded functionals on $\mathbb{X}$. For $f \in \mathbb{X}^*$ and $u \in \mathbb{X}$, we conventionally use the following notation

$$\langle f, u\rangle_{\mathbb{X}^* \times \mathbb{X}} \equiv \langle f, u\rangle_{\mathbb{X}} \equiv f(u)$$

to denote the action of f on u (or the evaluation of f at u). The first notation is the standard duality pairing while the second is an abuse of notation due to the fact that we can identify the dual space $\mathbb{X}^*$ with $\mathbb{X}$ via the Riesz representation theorem 2.1 (see Exercise 2.11).

Example 2.1. We can denote the Fréchet derivative at u as $\mathscr{D}f(u)$, which is by definition an element of $\mathbb{X}^*$, and we can write

$$\langle \mathscr{D}f(u), h\rangle_{\mathbb{X}} \equiv \langle \mathscr{D}f(u,\cdot), h\rangle_{\mathbb{X}} \equiv \mathscr{D}f(u,h).$$

Thus, the gradient $\nabla f(u)$ is the Riesz representation of $\mathscr{D}f(u)$. △

Exercise 2.11. Let $\mathbb{X} = \mathbb{R}^n$. Show that we can identify $\mathbb{X}^* = (\mathbb{R}^n)^*$ with $\mathbb{X} = \mathbb{R}^n$. Show that we can identify $\left[\mathbb{L}^2(\Omega)\right]^*$ with $\mathbb{L}^2(\Omega)$, and more generally $\mathbb{X}^*$ with $\mathbb{X}$. •

Now let $\mathscr{A} \in \mathscr{B}(\mathbb{X}, \mathbb{Y})$, we define its adjoint $\mathscr{A}^* : \mathbb{Y}^* \to \mathbb{X}^*$ as

$$\langle \mathscr{A}x, y^*\rangle_{\mathbb{Y}} = \langle \mathscr{A}^* y^*, x\rangle_{\mathbb{X}}, \quad \forall x \in \mathbb{X} \text{ and } \forall y^* \in \mathbb{Y}^*.$$

Exercise 2.12. Show that $\mathscr{A}^* \in \mathscr{B}(\mathbb{Y}^*, \mathbb{X}^*)$. •

Theorem 2.3. *Let $\mathscr{A} \in \mathscr{B}(\mathbb{X},\mathbb{Y})$, then $\|\mathscr{A}^*\| = \|\mathscr{A}\|$.*

Proof. The result of Exercise 2.12 shows that $\|\mathscr{A}^*\| \leq \|\mathscr{A}\|$. An elegant and short proof of the reverse can be done using the Hahn-Banach theorem [22, 3]. □

Proposition 2.1. *There hold:*

1. *If $\mathscr{I}$ be the identity operator on $\mathbb{X}$, then $\mathscr{I}^* = \mathscr{I}$.*
2. *If $\mathscr{A}_1, \mathscr{A}_2 \in \mathscr{B}(\mathbb{X},\mathbb{Y})$, then $(\mathscr{A}_1 + \mathscr{A}_2)^* = \mathscr{A}_1^* + \mathscr{A}_2^*$.*
3. *If $\mathscr{A} \in \mathscr{B}(\mathbb{X},\mathbb{Y})$, then $(\alpha\mathscr{A})^* = \alpha\mathscr{A}^*$, where $\alpha \in \mathbb{R}$.*
4. *If $\mathscr{A}_1 \in \mathscr{B}(\mathbb{X},\mathbb{Y})$ and $\mathscr{A}_2 \in \mathscr{B}(\mathbb{Y},\mathbb{Z})$, then $(\mathscr{A}_2\mathscr{A}_1)^* = \mathscr{A}_1^*\mathscr{A}_2^*$.*
5. *If $\mathscr{A} \in \mathscr{B}(\mathbb{X},\mathbb{Y})$ and $\mathscr{A}$ has bounded inverse, then $\left(\mathscr{A}^{-1}\right)^* = (\mathscr{A}^*)^{-1}$.*

Exercise 2.13. Prove Proposition 2.1. •

Exercise 2.14. Recall

$$\mathbb{H}^1(\Omega) = \left\{u \in \mathbb{L}^2(\Omega) : \nabla u \in \mathbb{L}^2(\Omega)\right\},$$
$$\mathbb{H}^2(\Omega) = \left\{u \in \mathbb{H}^1(\Omega) : \Delta u \in \mathbb{L}^2(\Omega)\right\}.$$

Let us define $\mathscr{A} : \mathbb{H}^2(\Omega) \subset \mathbb{L}^2(\Omega) \to \mathbb{L}^2(\Omega)$ via

$$\mathscr{A}u \stackrel{\text{def}}{=} -\Delta u \text{ with } u = 0 \text{ on } \partial\Omega.$$

Here we see that the differential operator can only be defined on a (dense) subspace $\mathbb{H}^2(\Omega)$ of $\mathbb{L}^2(\Omega)$. It turns out that the adjoint is also a differential operator whose domain is a subspace of $\mathbb{L}^2(\Omega)$. Find $\mathscr{A}^*$. •

Here is a version of the closed range theorem that is useful for our development.

Theorem 2.4 (Closed range theorem). *Let $\mathscr{A} \in \mathscr{B}(\mathbb{X},\mathbb{Y})$, where $\mathbb{X},\mathbb{Y}$ are two Hilbert spaces. Then the following are equivalent with each other:*

1. $[R(\mathscr{A})]^\perp = N(\mathscr{A}^*)$.
2. $\overline{R(\mathscr{A})} = [N(\mathscr{A}^*)]^\perp$.
3. $[R(\mathscr{A}^*)]^\perp = N(\mathscr{A})$.
4. $\overline{R(\mathscr{A}^*)} = [N(\mathscr{A})]^\perp$.

If $R(\mathscr{A})$ is closed, then we can replace its closure $\overline{R(\mathscr{A})}$ by itself and $R(\mathscr{A}^*)$ is also closed.

Proof. We only prove the first assertion (a similar token can be done for the third). The equivalence with the second and the fourth ones requires a deeper understanding of the orthogonal complement and hence ignored. Let $y^* \in N(\mathscr{A}^*)$ and $y \in R(A)$. Then $y = \mathscr{A}x$ for some $x \in \mathbb{X}$. We have

$$\langle y^*, y\rangle_{\mathbb{Y}} = \langle y^*, \mathscr{A}x\rangle_{\mathbb{Y}} = \langle \mathscr{A}^* y^*, x\rangle_{\mathbb{X}} = 0,$$

which says that $N(\mathscr{A}^*) \subset [R(\mathscr{A})]^\perp$. Now take $y^* \in [R(\mathscr{A})]^\perp$, we have

$$\langle \mathscr{A}^* y^*, x \rangle_{\mathbb{X}} = \langle y^*, \mathscr{A} x \rangle_{\mathbb{Y}} = 0, \quad \forall x \in \mathbb{X},$$

which implies that $\mathscr{A}^* y^* = 0$, which in turn shows $[R(\mathscr{A})]^{\perp} \subset N(\mathscr{A}^*)$. □

Exercise 2.15. Prove the third assertion of Theorem 2.4. •

Exercise 2.16. Let $\mathscr{A} \in \mathbb{R}^{n \times m}$. Show that

$$\mathbb{R}^m = N(\mathscr{A}) \oplus R\left(\mathscr{A}^T\right), \quad \text{and } \mathbb{R}^n = N\left(\mathscr{A}^T\right) \oplus R(A).$$

•

We are now ready for the Lagrangian multiplier theorem.

Theorem 2.5 (Lagrangian multiplier theorem). *Assume that f is continuously Fréchet differentiable and it attains the extremum at u_0 subject to the constraint $c(u) = 0$. Suppose that $\mathscr{D}c(u_0)$ maps $\mathbb{X}$* **onto** *$\mathbb{Y}$. Then there exists an element $y^* \in \mathbb{Y}^*$ such that the following Lagrangian functional*

$$L(u) \stackrel{def}{=} f(u) + \langle y^*, c(u) \rangle_{\mathbb{Y}}$$

is stationary at u_0, i.e.,

$$\mathscr{D}L(u_0, h) = \mathscr{D}f(u_0, h) + \langle y^*, \mathscr{D}c(u_0, h) \rangle_{\mathbb{Y}} = 0, \quad \forall h \in \mathbb{X}, \tag{2.1}$$

or equivalently

$$\mathscr{D}L(u_0) = \mathscr{D}f(u_0) + [\mathscr{D}c(u_0)]^* y^* = 0. \tag{2.2}$$

Proof. From Lemma 2.3 we see that $\mathscr{D}f(u_0)$ is orthogonal to the nullspace of $\mathscr{D}c(u_0)$. Since $\mathscr{D}c(u_0)$ maps $\mathbb{X}$ onto $\mathbb{Y}$, its range space is closed. Together with the closed range Theorem 2.4 we arrive at

Why?

$$\mathscr{D}f(u_0) \in R\left([\mathscr{D}c(u_0)]^*\right),$$

which implies that there exists $y^* \in \mathbb{Y}^*$ such that

$$\mathscr{D}f(u_0) = -[\mathscr{D}c(u_0)]^* y^*,$$

and this ends the proof. □

Now using the definition of gradient, we have

$$(\nabla L(u_0), h)_{\mathbb{X}} = (\nabla f(u_0), h)_{\mathbb{X}} + (y, \mathscr{D}c(u_0, h))_{\mathbb{Y}} = 0, \quad \forall h \in \mathbb{X},$$

where y is the Riesz representation of y^*. We observe that $\mathscr{D}c(u_0)$ can be understood as a bilinear functional with respect to h and y. Again, by Riesz representation theorem we can define $\nabla c(u_0; \cdot) : \mathbb{Y} \to \mathbb{X}$:

$$(\nabla c(u_0; y), h)_{\mathbb{X}} \stackrel{\text{def}}{=} (y, \mathscr{D}c(u_0, h))_{\mathbb{Y}},$$

and this allows us to conclude

$$\nabla L(u_0) = \nabla f(u_0) + \nabla c(u_0; y) = 0. \tag{2.3}$$

Equation (2.3) *together with the constraint* $c(u_0) = 0$ *provides enough equations to solve for the stationary point* u_0 *and the Lagrange muliplier* y *(or* y^**).*

Example 2.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathscr{A} \in \mathbb{R}^{m\times n}$. We consider the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \qquad \text{subject to } \mathscr{A}\mathbf{x} = \mathbf{b}.$$

Theorem 2.5 and Exercise 2.7, the necessary condition reads

$$\nabla f^T \mathbf{h} + \mathbf{y}^T \mathscr{A}\mathbf{h} = 0, \quad \forall \mathbf{h} \in \mathbb{R}^n,$$

which implies

$$\nabla f + \mathscr{A}^T \mathbf{y} = 0,$$

which is a paticular instance of (2.3). Of course, from Lemma 2.3 we can write the optimality condition in the equivalent form, i.e. avoiding the introduction of the Lagrange multiplier,

$$\nabla f^T \mathbf{h} = 0, \quad \forall \mathbf{h} \text{ such that } \mathscr{A}\mathbf{h} = \mathbf{0},$$

i.e., *due to the constraint, the gradient is not zero but its projection onto the nullspace of* $\mathscr{A}$ *must vanish*. If we let $\mathbf{Z}$ whose columns comprises a basis for the nullspace of $\mathscr{A}$, then $\mathbf{h} = \mathbf{Z}\mathbf{r}$ for some vector $\mathbf{r}$ whose dimension is the same as that of the nullspace. Then the first order necessary condition reads

$$\nabla f^T \mathbf{Z} = 0.$$

Note that $\mathbf{g}_r(\mathbf{x}) \stackrel{\text{def}}{=} \nabla f^T \mathbf{Z}$ is known as the reduced gradient. One of the reason for its name is that the its dimension is now reduced compared to that of the original counterpart. *Another important point one can draw from the optimality condition for the reduced gradient is that in the reduced space, i.e.* $\mathbf{r}$ *instead of* $\mathbf{x}$*, the optimization problem becomes unconstrained*. Now from Exercise 2.3 we know that the derivative of the reduced gradient $\mathbf{g}_r$, namely the reduced Hessian $\mathbf{H}_r$, is necessary to be semi-positive definite in any direction in the reduced space. By the chain rule we have

$$\mathbf{z}^T \mathbf{H}_r \mathbf{r} = \left.\frac{d\mathbf{g}_r(\mathbf{x} + t\mathbf{Z}\mathbf{r})}{dt}\right|_{t=0} = \mathbf{r}^T \mathbf{Z}^T \nabla f^2 \mathbf{Z}\mathbf{r},$$

from which it follows that

$$\mathbf{H}_r = \mathbf{Z}^T \nabla f^2 \mathbf{Z}.$$

Of course, the sufficient condition is, again due to Exercise 2.3, that the reduced Hessian must be positive definite. △

Exercise 2.17. As can be seen, the nullspace of the $\mathscr{A}$ is the key of the reduced space method. Show that we can use the QR factorization to compute $\mathbf{Z}$. Now assume that $\mathscr{A} \in R^{m\times n}$, where $m \leq n$, has linearly independent rows, we can often rewrite $\mathscr{A}$ as

$$\mathscr{A} = [\mathbf{U}, \mathbf{V}],$$

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ is a nonsingular matrix. Show that a form of $\mathbf{Z}$ can be written as

$$\mathbf{Z} = \begin{bmatrix} -\mathbf{U}^{-1}\mathbf{V} \\ \mathbf{I} \end{bmatrix},$$

where $\mathbf{I}$ is the $(n-m)\times(n-m)$ identity matrix. What is the relationship between the reduced variable and the original one?

This is one of the most popular approaches in the reduced space method, especially for PDE-constrained optimization problems, as we shall show later. •

Exercise 2.18. Let $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x_1, x_2) = -x_1^2 + x_2^2$ and the constraint is $x_2 = 1$. Find the optimality condition in terms of Lagrangian mulitpliers, and then compute stationary points and the corrsponding Lagrange multipliers. Find the reduced gradient and Hessian at these points. Which point is a local minimizer? •

Exercise 2.19. Consider the linear constraint $\mathscr{A}\mathbf{x} = x_1 + x_2 + x_3 = 3$, regardless of the particular form of the cost function $f : \mathbb{R}^3 \to \mathbb{R}$. From the result of Exercise 2.16, show that any $\mathbf{x} \in \mathbb{R}^3$ can be represented as

$$\mathbf{x} = \mathbf{Y}\mathbf{y} + \mathbf{Z}\mathbf{r}.$$

where $\mathbf{Y}$ is a matrix spanning the range space of $\mathscr{A}^T$ and $\mathbf{Z}$ spanning the null space of $\mathscr{A}$. Find an instance of $\mathbf{Y}$ and $\mathbf{Z}$. Now suppose that $\mathbf{x}^* \in \mathbb{R}^3$ is a solution to the optimization, find the general form of $\mathbf{x}^*$. •

Exercise 2.20. We are interested in minimizing $f(\mathbf{x}) = x_1^2x_2^3 + 4x_1^2x_3^2 + x_2^4x_3^2 + 3x_1x_2 + 4x_2x_3 + 5x_1x_3$ subject to $x_1 + x_2 + x_3 = 3$. Find stationary points and determine if they are minimizers? •

Example 2.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{c}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$. We consider the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \qquad \text{subject to } \mathbf{c}(\mathbf{x}) = \mathbf{b}.$$

Theorem 2.5 and Exercise 2.7, the necessary condition reads

$$\nabla f(\mathbf{x}) + \nabla \mathbf{c}^T(\mathbf{x})\,\mathbf{y} = 0, \tag{2.4}$$

where ∇c is the Jacobian matrix of the contraints at $\mathbf{x}$. We thus arrive at the same first order optimality condition by simply replacing $\mathscr{A}$ with $\nabla\mathbf{c}(\mathbf{x})$, i.e.,

$$\mathbf{g}(\mathbf{x}) \stackrel{\text{def}}{=} \nabla f^T(\mathbf{x})\,\mathbf{h}(\mathbf{x}) = 0, \quad \forall \mathbf{h}(\mathbf{x}) \text{ such that } \nabla\mathbf{c}(\mathbf{x})\,\mathbf{h}(\mathbf{x}) = \mathbf{0}.$$

Note that unlike the linear problem, $\mathbf{h}$ is a function of $\mathbf{x}$ since the nullspace of $\nabla\mathbf{c}(\mathbf{x})$ depends on $\mathbf{x}$. As a result, *the reduced Hessian is different from that of the linear counterpart as we now show*. To that end, we compute the Fréchet derivative of $\mathbf{g}$ in any direction $\mathbf{p}$ in the reduced space $N(\nabla\mathbf{c}(\mathbf{x}))$:

$$\mathbf{p}^T\mathbf{H}\mathbf{h} = \mathbf{p}^T\nabla f^2\mathbf{h} + \nabla f^T\mathscr{D}\mathbf{h}(\mathbf{x},\mathbf{p}). \tag{2.5}$$

To compute $\mathscr{D}\mathbf{h}(\mathbf{x},\mathbf{p})$ we take the Fréchet derivative both sides of $\nabla\mathbf{c}(\mathbf{x})\,\mathbf{h}(\mathbf{x}) = \mathbf{0}$, row by row, in direction $\mathbf{p}$:

$$\mathbf{p}^T\nabla^2 c_i\mathbf{h} + \nabla c_i^T\mathscr{D}\mathbf{h}(\mathbf{x},\mathbf{p}) = 0. \tag{2.6}$$

Combining (2.4)–(2.6) gives *Can you see it?*

$$\mathbf{p}^T\mathbf{H}\mathbf{h} = \mathbf{p}^T\left(\nabla f^2 - \sum_{i=1}^{m}\nabla^2 c_i y_i\right)\mathbf{h}.$$

Since both $\mathbf{p}$ and $\mathbf{h}$ belong to the nullspace $\mathbf{Z}(\mathbf{x})$ of $\nabla\mathbf{c}(\mathbf{x})$, the reduced Hessian $\mathbf{H}_r$ is then given by *Work it out!*

$$\mathbf{H}_r = \mathbf{Z}^T\left(\nabla f^2 - \sum_{i=1}^{m}\nabla^2 c_i y_i\right)\mathbf{Z}.$$

As can be seen, the Hessians of the constraints (which is zero for linear constraint case) contribute to the reduced Hessian. △

Exercise 2.21. Consider $\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) = x_1x_2^2$ subject to $2 - x_1^2 - x_2^2 = 0$. Compute the stationary points and the corresponding Lagrange mulitpliers, which one is a minimizer? •

Example 2.4. Consider

$$f(x) = \int_{t_0}^{t_1} g\left(x,x',t\right)dt, \quad \text{subject to} \quad c(x,t) = 0, \text{ and } x(t_0) = x(t_1) = 0,$$

where $c(\cdot,\cdot) : \mathbb{C}^1[t_0,t_1]\times(t_0,t_1) \to \mathbb{L}^2(t_0,t_1)$.

We look for the solution in the space $\mathbb{C}_0^1[t_0,t_1] = \left\{u \in \mathbb{C}^1[t_0,t_1] : u(t_0) = u(t_1) = 0\right\} \subset \mathbb{L}^2(t_0,t_1)$. Recall that any linear functional y^* on $\mathbb{L}^2(t_0,t_1)$ can be written as

$$\langle y^*, v\rangle_{\mathbb{L}^2} = \int_{t_0}^{t_1} yv\,dt,$$

where y is the Riesz respresentation of y^*. The first order optimality condition (2.1) reads *Work it out!*

$$\int_{t_0}^{t_1}\left(\frac{\partial g}{\partial x}h + \frac{\partial g}{\partial x'}h'\right)dt + \int_{t_0}^{t_1} yc_x h\,dt = 0, \quad \forall h \in \mathbb{C}_0^1[t_0,t_1],$$

which, after integrating by parts the second term in the first integral and using the fact that $\mathbb{C}_0^1[t_0,t_1]$ is dense[3] in $\mathbb{L}^2(t_0,t_1)$, gives

$$\frac{\partial g}{\partial x} - \frac{d}{dt}\left(\frac{\partial g}{\partial x'}\right) + yc_x = 0.$$

This is exactly an Euler-Lagrange equation (see Exercise 2.10). Comparing to (2.2) we see that in the standard L^2-inner product we have

$$\nabla f = \frac{\partial g}{\partial x} - \frac{d}{dt}\left(\frac{\partial g}{\partial x'}\right), \text{ and } [\mathscr{D}c(u_0)]^* y^* = yc_x.$$

△

Exercise 2.22. Rederive the first order optimality condition when $c(\cdot,\cdot) : \mathbb{C}^1[t_0,t_1] \times (t_0,t_1) \to \mathbb{R}$. •

Exercise 2.23. Seek the curve in the $x-y$ plane having end points $(-1,0)$,$(1,0)$, length ℓ, and enclosing maximum area between itself and the x-axis. •

Example 2.5 (PDE-constrained optimization). Many PDE-constrained optimization problems can be expressed in the following form

$$\min_{u\in\mathbb{X},m\in\mathbb{Z}} f(u,m), \qquad \text{subject to } c(u,m) = 0, \text{ where } c(\cdot,\cdot) : \mathbb{X}\times\mathbb{Z} \to \mathbb{Y},$$

where *the Fréchet derivative of the contraint with respect to u, i.e.* $\mathscr{D}_u c(u,m) : \mathbb{X} \to \mathbb{Y}$*, is invertible*. The first order optimality condition (2.2), together with the constraint, can be written as

$$c(u,m) = 0, \qquad \text{Forward equation} \tag{2.7a}$$
$$\mathscr{D}_u f(u,m) + [\mathscr{D}_u c(u_0)]^* y^* = 0, \qquad \text{Adjoint equation} \tag{2.7b}$$
$$\mathscr{D}_m f(u,m) + [\mathscr{D}_m c(u_0)]^* y^* = 0, \qquad \text{Control equation.} \tag{2.7c}$$

Note that this is a generalization of Exercise 2.17 since $\mathscr{A} = \mathscr{D}c(u_0) = [\mathscr{D}_u c(u_0), \mathscr{D}_m c(u_0)]$. In other words, the structure of this problem allows us to eliminate u via the *forward equation* (2.7a) so that the reduced variable is in fact m. Again, the problem becomes unconstrained in m. The procedure for computing the reduce gradient for a given m requires three steps: 1) solve the *forward equation* (2.7a) for $u(m)$, 2) solve the adjoint equation for $y^*(u(m),m)$, and 3) evaluate the reduced gradient using the left side of the *control equation* (2.7c). *The gradient, the key object of this chapter, is what we need for advanced MCMC method later for PDE-constrained statistical inverse problems*. △

Exercise 2.24. To appreciate the adjoint approach, let us again consider the following problem

[3] Since $\mathbb{C}_0^1[t_0,t_1]$ is dense in $\mathbb{L}^2(t_0,t_1)$, $\int x(t)y(t)\,dt = 0, \forall x(t) \in \mathbb{C}_0^1[t_0,t_1]$ implies that $y(t) = 0$.

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{m}\in\mathbb{R}^p} f(\mathbf{x},\mathbf{m}), \qquad \text{subject to } \mathbf{c}(\mathbf{x},\mathbf{m}) = \mathbf{b},$$

where $f:\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}$ and $\mathbf{c}(\mathbf{x},\mathbf{m}):\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}^n$. We assume that $\det\left(\frac{\partial\mathbf{c}}{\partial\mathbf{x}}\right)\neq 0, \forall\mathbf{x},\mathbf{m}$ so that the implicit function theorem allows us to compute $\mathbf{x}$ as a function of $\mathbf{m}$ from the constraint.

1. Drive the reduced gradient vector at a point $\mathbf{m}_0$ using the framework in Example 2.5.
2. One of course can ignore the adjoint approach and use the *direct sensitivity method*. In this method, we compute each component of the gradient, i.e. $\frac{\partial f}{\partial m_i}$, using the chain rule. Compare and contrast with the adjoint approach in the cases $p=1$ and $p\gg 1$.

•

Example 2.6 (Advection-PDE-constrained optimization problem). Consider the following PDE-constrained optimization problem

$$\min_{m,u}\frac{1}{2}\int_\Omega u^2\,d\Omega$$

subject to

$$\begin{aligned}\beta\cdot\nabla u &= 0, \quad \text{in } \Omega,\\ \beta\cdot nu &= m, \quad \text{in } \partial\Omega_{\text{in}},\end{aligned}$$

where $u\in H^1_\beta(\Omega)\overset{\text{def}}{=}\left\{u: u\in\mathbb{L}^2(\Omega) \text{ and } \beta\cdot\nabla u\in\mathbb{L}^2(\Omega)\right\}$. Note that for $u\in H^1_\beta(\Omega)$, its trace (in fact weighted trace with weight $\beta\cdot n$, where n is the normal vector), for example, on $\partial\Omega_{\text{in}}$ belongs to $\mathbb{L}^2(\partial\Omega_{\text{in}})$. The correct space for m is thus $\mathbb{L}^2(\partial\Omega_{\text{in}})$. Since the constraints map $u\in H^1_\beta(\Omega)$ to $\mathbb{Y}\overset{\text{def}}{=}\mathbb{L}^2(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$, the Lagrange multiplier y^* has two components $y\in\mathbb{L}^2(\Omega)$ and $z\in\mathbb{L}^2(\partial\Omega_{\text{in}})$, respectively. Our task is to find the KKT stationary condition (2.7). For practical PDE-constrained problem, the adjoint operators $[\mathscr{D}_u c(u_0)]^* y^*$ and $[\mathscr{D}_z c(u_0)]^* y^*$ are subtlely coupled and are typically found by integration by parts. To demonstrate this, let us form the Lagrangian

$$L(m,u)=\frac{1}{2}\int_\Omega u^2\,d\Omega+\int_\Omega(\beta\cdot\nabla u)\,y\,d\Omega+\int_{\partial\Omega_{\text{in}}}(\beta\cdot nu-m)\,z\,ds.$$

Take an arbitrary direction $h\in H^1_\beta(\Omega)$ and $n\in\mathbb{L}^2(\partial\Omega_{\text{in}})$, since the constraints are linear we have

$$\langle y^*,\mathscr{D}c(h,n)\rangle_{\mathbb{Y}}=\int_\Omega(\beta\cdot\nabla h)\,y\,d\Omega+\int_{\partial\Omega_{\text{in}}}(\beta\cdot nh-n)\,z\,ds,$$

which after integration by parts becomes

$$\langle y^*, \mathscr{D}c(h,n)\rangle_{\mathbb{Y}} = -\int_\Omega (\beta\cdot\nabla y)\,h\,d\Omega + \int_{\partial\Omega_{\text{in}}} \beta\cdot n\,(z+y)\,h\,ds + \int_{\partial\Omega_{\text{out}}} \beta\cdot n y h\,ds - \int_{\partial\Omega_{\text{in}}} n z\,ds,$$

where, for simplicity, we have assumed that $\nabla\cdot\beta = 0$. Here, we have restricted y into $H^1_\beta(\Omega)$ for the differential and integral operators to make sense (see also Exercise 2.14). The first order stationary condition (2.1) in this case reads: $\forall h$ and n,

$$\int_\Omega u h\,d\Omega - \int_\Omega (\beta\cdot\nabla y)\,h\,d\Omega + \int_{\partial\Omega_{\text{in}}} \beta\cdot n\,(z+y)\,h\,ds + \int_{\partial\Omega_{\text{out}}} \beta\cdot n y h\,ds - \int_{\partial\Omega_{\text{in}}} n z\,ds = 0,$$

which is also true for $n = 0$ and any $h \in H^1_{\beta,0}(\Omega) \overset{\text{def}}{=} \left\{u \in H^1_\beta(\Omega) : u|_{\partial\Omega} = 0\right\}$, that is,

$$\int_\Omega u h\,d\Omega - \int_\Omega (\beta\cdot\nabla y)\,h\,d\Omega = 0, \quad \forall h \in H^1_{\beta,0}(\Omega),$$

which implies[4]

$$-\beta\cdot\nabla y + u = 0.$$

Consequently, the stationary condition reduces to

$$\int_{\partial\Omega_{\text{in}}} \beta\cdot n\,(z+y)\,h\,ds + \int_{\partial\Omega_{\text{out}}} \beta\cdot n y h\,ds - \int_{\partial\Omega_{\text{in}}} n z\,ds = 0, \quad \forall h, n,$$

which, by taking $h = 0$ on $\partial\Omega_{\text{out}}$ and $n = 0$, implies

$$\int_{\partial\Omega_{\text{in}}} \beta\cdot n\,(z+y)\,h\,ds = 0,$$

which in turn gives[5]

$$z = -y \text{ on } \partial\Omega_{\text{in}},$$

that is, the adjoint variables are not independent. This can be substituted into the stationary condition to yield

$$\int_{\partial\Omega_{\text{out}}} \beta\cdot n y h\,ds + \int_{\partial\Omega_{\text{in}}} n y\,ds = 0, \quad \forall h, n.$$

It follows that

$$\beta\cdot n y = 0 \text{ on } \partial\Omega_{\text{out}}.$$

and

$$y = 0 \text{ on } \partial\Omega_{\text{in}}.$$

In summary, the adjoint equation (2.7b) reads

[4] It is due to the fact that $H^1_{\beta,0}(\Omega)$ is dense in $\mathbb{L}^2(\Omega)$ assuming Ω has segment property [2].

[5] Note that this is true due to the fact that the trace operator $\gamma : H^1_{\beta,0}(\Omega) \to \mathbb{L}^2_{\beta\cdot n}(\partial\Omega)$ is a continuous surjection [6].

$$-\beta \cdot \nabla y = -u \text{ in } \Omega,$$
$$\beta \cdot ny = 0 \text{ on } \partial\Omega_{\text{out}},$$

and the control equation (2.7c) becomes

$$y = 0 \text{ on } \partial\Omega_{\text{in}}.$$

As can be seen, the adjoint equation desribes a reverse flow with (the derivative of) the cost function as the forcing. The control equation says that at the optimal the forcing of the adjoint is such that the adjoint solution on $\partial\Omega_{\text{in}}$ must vanish. This can only be true if the adjoint is identically zero, or the forcing u is identically zero. It then follows from the forward equation that $m = 0$. This is not surprising since, by observation, the quadratic optimization under consideration has a unique solution $u = 0$ and $m = 0$. △

Exercise 2.25. Consider a similar setting as in Example 2.6, but assume the inflow data m is given and the optimization variable is now β. Derive the first order stationary condition for this case. •

Example 2.7 (Elliptic-PDE-constrained optimization problem).

$$\min_{m,u} J(u) \overset{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \left(u(\mathbf{x}_i) - u_i^{obs}\right)^2 = \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \int_\Omega \left[u(\mathbf{x}) - u^{obs}(\mathbf{x})\right]^2 \delta(\mathbf{x} - \mathbf{x}_i)\, d\Omega$$

subject to

$$-\nabla \cdot (e^m \nabla u) = 0, \quad \text{in } \Omega,$$
$$u = g, \quad \text{in } \partial\Omega,$$

where we have assume that the forward solution u is sufficiently smooth[6], for example $u \in \mathbb{H}^2(\Omega)$, so that the pointwise evaluation $u(\mathbf{x}_i)$ make sense. We assume that $m \in \mathbb{C}(\Omega) \subset \mathbb{L}^2(\Omega)$ and be bounded. Thus, the contraint c maps $\mathbb{H}^2(\Omega) \times \mathbb{C}(\Omega)$ to $\mathbb{L}^2(\Omega) \times \mathbb{L}^2(\partial\Omega)$, and the Lagrange multiplier y^* has two components $y \in \mathbb{L}^2(\Omega)$ and $z \in \mathbb{L}^2(\partial\Omega)$, respectively. The Lagrangian reads

$$L(m,u) = J(u) + \int_\Omega \left[-\nabla \cdot (e^m \nabla u)\right] y\, d\Omega + \int_{\partial\Omega} (u - g)\, z\, ds.$$

For arbitrary direction $h \in H^2(\Omega)$ and $n \in \mathbb{C}(\partial\Omega)$, the stationary condition (2.1) becomes

[6] For example $u \in \mathbb{H}^s(\Omega)$ with $s > d/2$, then the Sobolev embedding theorem says that u is in fact continuous, and hence pointwise evaluation is meaningful. Weaker assumption can be found in [15].

$$\sum_{i=1}^{N^{\text{obs}}} \int_\Omega \left[u(\mathbf{x}) - u^{obs}(\mathbf{x})\right] h(\mathbf{x}) \delta(\mathbf{x}-\mathbf{x}_i)\, d\Omega + \int_\Omega [-\nabla\cdot(e^m \nabla h)]\, y\, d\Omega + \int_{\partial\Omega} hz\, ds + \int_\Omega [-\nabla\cdot(e^m n \nabla u)]\, y\, d\Omega = 0, \quad \forall h, n.$$

We next restrict $y \in \mathbb{H}^2(\Omega)$ and integrate the second term by parts two times we arrive at: $\forall h, n$,

$$\sum_{i=1}^{N^{\text{obs}}} \int_\Omega \left[u(\mathbf{x}) - u^{obs}(\mathbf{x})\right] h(\mathbf{x}) \delta(\mathbf{x}-\mathbf{x}_i)\, d\Omega + \int_\Omega [-\nabla\cdot(e^m \nabla y)]\, h\, d\Omega + \int_{\partial\Omega} hz\, ds - \int_{\partial\Omega} e^m \nabla h \cdot n\, y\, ds + \int_{\partial\Omega} e^m \nabla y \cdot n\, h\, ds + \int_\Omega [-\nabla\cdot(e^m n \nabla u)]\, y\, d\Omega = 0.$$

Following a similar strategy as in Example 2.6 gives the adjoint equation

$$\begin{aligned} -\nabla\cdot(e^m \nabla y) &= -\sum_{i=1}^{N^{\text{obs}}} \left[u(\mathbf{x}) - u^{obs}(\mathbf{x})\right] \delta(\mathbf{x}-\mathbf{x}_i), \quad \text{in } \Omega, \\ y &= 0, \quad \text{in } \partial\Omega, \end{aligned}$$

and

$$z = -e^m \nabla y \cdot n.$$

Again, we see that the second component of the adjoint variable depends on the first, hence y is in fact the only adjoint variable. The control equation is given by

$$e^m \nabla u \cdot \nabla y = 0.$$

△

Exercise 2.26. Consider the following PDE-constrained optimization problem

$$\min_{m,u} J(u) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{i=1}^{N^{\text{obs}}} \left(u(\mathbf{x}_i) - u_i^{obs}\right)^2 = \frac{1}{2} \sum_{i=1}^{N^{\text{obs}}} \int_\Omega \left[u(\mathbf{x}) - u^{obs}(\mathbf{x})\right]^2 \delta(\mathbf{x}-\mathbf{x}_i)\, d\Omega$$

subject to

$$\begin{aligned} -\nabla\cdot(e^m \nabla u) &= 0, \quad \text{in } \Omega, \\ -e^m \nabla u \cdot n &= -1, \quad \text{in } \Gamma, \\ -e^m \nabla u \cdot n &= Bi\, u, \quad \text{in } \partial\Omega \setminus \Gamma, \end{aligned}$$

where Bi is the Biôt number, which is given. Derive the first order optimality condition (2.7) for this problem. •

Exercise 2.27. A function $\pi(\mathbf{x})$ on $\mathbb{R}^n$ is a probability distribution if

$$0 < \pi(\mathbf{x}), \quad \int_{\mathbb{R}^n} \pi(\mathbf{x})\, d\mathbf{x} = 1.$$

Let us define the *entropy* of a probability distribution $\pi(\mathbf{x})$ as

$$H(\pi(\mathbf{x})) \stackrel{\text{def}}{=} -\int_{\mathbb{R}^n} \pi(\mathbf{x}) \log(\pi(\mathbf{x}))\, d\mathbf{x}.$$

Show that a probability distribution $\pi(\mathbf{x})$ that maximizes the entropy subject to the mean μ, i.e.,

$$\int_{\mathbb{R}^n} \mathbf{x}\pi(\mathbf{x})\, d\mathbf{x} = \mu,$$

and the covariance Σ, i.e.,

$$\int_{\mathbb{R}^n} (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T \pi(\mathbf{x})\, d\mathbf{x} = \Sigma,$$

is the normal distribution with mean μ and covariance Σ.

•

Exercise 2.28. We are given a *prior distribution* $\pi_{\text{prior}}(\mathbf{x})$ (more later in Chapter 7). Consider the following optimization problem

$$\pi_{\text{post}}(\mathbf{x}) \stackrel{\text{def}}{=} \arg\min_{\pi(\mathbf{x})} \int_{\mathbb{R}^n} \pi(\mathbf{x}) \log\left(\frac{\pi(\mathbf{x})}{\pi_{\text{prior}}(\mathbf{x})}\right) d\mathbf{x} + \int_{\mathbb{R}^n} \|\mathbf{y} - \mathscr{G}(\mathbf{x})\|_{\Sigma}^2 \pi(\mathbf{x})\, d\mathbf{x},$$

where we have defined

$$\|\mathbf{y} - \mathscr{G}(\mathbf{x})\|_{\Sigma}^2 \stackrel{\text{def}}{=} [\mathbf{y} - \mathscr{G}(\mathbf{x})]^T \Sigma^{-1} [\mathbf{y} - \mathscr{G}(\mathbf{x})].$$

subject to

$$0 < \pi(\mathbf{x}), \quad \text{and} \int_{\mathbb{R}^n} \pi(\mathbf{x})\, d\mathbf{x} = 1.$$

Note that the optimization problem finds a probability density $\pi_{\text{post}}(\mathbf{x})$ function such that it is not very "far away" from $\pi_{\text{prior}}(\mathbf{x})$ and that the average mismatch between the observation $\mathbf{y}$ and the prediction $\mathscr{G}(\mathbf{x})$ is minimized. Find $\pi_{\text{post}}(\mathbf{x})$. **Hint: you can ignore the first constraint because it will be automatically satisfied as you will see.**

•

2.3 Appendix

Add a demonstration that the stationary condition of Lemma 2.3 is the same of the stationary condition of the Lagrangian in Theorem 2.5: finite dimensional problem to show that the gradient of the cost functional must be orthogonal to the nullspace of the derivative of the constraint. Since the gradient of the constraint is itself or-

thogonal to the nullspace. It must be true that the gradient of the cost functional must align with the gradient of the constraint and this is exactly the Lagrangian theorem.

A little theory on bounded linear operator: particular shows that boundedness is equivalent of the continuity. This is important because we use this fact everywhere in the note.

Also, density argument and denseness

Maybe some of the nullspace and rangespace theory is good here. The closedness of the nullspace.

Part II
Bayesian Inversion Framework (Finite Dimensions)

Chapter 3
Basics on probability

3.1 Introduction

Let us motivate you by considering the following particular inverse problem, namely, the deconvolution problem. Given the observation signal $g(s)$, we would like to reconstruct the input signal $f(t) : [0,1] \to \mathbb{R}$, where the observation and the input obey the following relation

$$g(s_j) = \int_0^1 a(s_j,t) f(t)\, dt, \quad 0 \leq j \leq n. \tag{3.1}$$

Here, $a : [0,1] \times [0,1] \to \mathbb{R}$ is known as the blurring kernel. So, in fact we don't know the output signal completely, but at a finite number of observation points. A straightforward approach you may think of is to apply some numerical quadrature on the right side of (3.1), and then recover $f(t)$ at the quadrature points by inverting the resulting matrix. If you do this, you realize that the matrix is ill-conditioned, and it is not a good idea to invert it. There are techniques to go around this issue, but let us not pursue them here. Instead, we recast the deconvolution task into an optimization problem such as

$$\min_{f(t)} \sum_{j=0}^{n} \left(g(s_j) - \int_0^1 a(s_j,t) f(t)\, dt \right)^2, \tag{3.2}$$

that is, we minimize the *misfit* between the mathematical model $\int_0^1 a(s_j,t) f(t)\, dt$ and the actual observations. However, the ill-conditioning nature of our inverse problem does not go away. Indeed, (3.2) may have multiple solutions and multiple minima. In addition, a solution to (3.2) may not depend continuously on $g(s_j), 0 \leq j \leq n$. So what is the point of recast? Clearly, if the cost function (also known as the data misfit) is a parabola, then the optimal solution is unique. This immediately suggests that one should add a quadratic term to the cost function to make it more like a parabola, and hence making the optimization problem easier.

This is essentially the idea behind the *Tikhonov regularization*, which proposes to solve the nearby problem

$$\min_{f(t)} \sum_{j=0}^{n} \left(g(s_j) - \int_0^1 a(s_j,t) f(t)\, dt \right)^2 + \frac{\kappa}{2} \left\| \mathscr{R}^{1/2} f \right\|^2,$$

where κ is known as the regularization parameter, and $\|\cdot\|$ is some appropriate norm. Perhaps, two popular choices for $R^{1/2}$ are ∇ and Δ, the gradient and Laplace operator, respectively, and we discuss them in details in the following.

Now, in practice, we are typically not able to observe $g(s_j)$ directly but its noise-corrupted values

$$g^{obs}(s_j) = g(s_j) + e_j, \quad 0 \le j \le n,$$

where e_j, $j = 0, \ldots, n$, are some random noise. You can think of the noise as the inaccuracy in observation/measurement devices. The question you may ask is how to incorporate this kind of randomness in the above deterministic solution methods. There are works in this direction, but let us introduce a statistical framework based on the Bayesian paradigm to you in this note. This approach is appealing since it can incorporate most, if not all, kinds of randomness in a systematic manner.

Some portion of this chapter follows the presentation of the two excellent books by Somersalo *et. al.* [10, 17]. The pace is necessary slow since we develop this note for readers with minimal knowledge in probability theory. The only requirement is to either be familiar with or adopt the conditional probability formula concept. This is the corner stone on which we build the rest of the theory. Clearly, the theory we present here is by no means complete since the subject is vast, and still under development.

3.2 Some concepts from probability theory

We begin with the definition of randomness.

Definition 3.1. An even is *d*eterministic if its outcome is completely predictable.

Definition 3.2. A *random event* is the complement of a deterministic event, that is, its outcome is not fully predictable.

Example 3.1. If today is Wednesday, then "tomorrow is Thursday" is deterministic, but whether it rains tomorrow is not fully predictable. △

As a result, randomness means lack of information and it is the direct consequence of our ignorance. To express our belief[1] on random events, we use probability; probability of uncertain events is always less than 1, an event that surely happens

[1] Different person has different belief which leads to different solution of the Bayesian inference problem. Specifically, one's belief is based on his known information (expressed in terms of σ-algebra) and "weights" on each information (expressed in terms of probability measure). That is, people working with different probability spaces have different solutions.

has probability 1, and an event that never happens has 0 probability. In particular, to reflect the subjective nature, we call it *subjective probability* or *Bayesian probability* since it represents belief, and hence depending upon one's experience/knowledge to decide what is reasonable to believe.

Example 3.2. Let us consider the event of tossing a coin. Clearly, this is a random event since we don't know whether head or tail will appear. Nevertheless, we believe that out of n tossing times, $n/2$ times is head and $n/2$ times is tail.[2] We express this belief in terms of probability as: the (subjective) probability of getting a head is $\frac{1}{2}$ and the (subjective) probability of getting a tail is $\frac{1}{2}$. △

We define $(\Omega, \mathscr{F}, \mathbb{P})$ as a *probability space*. One typically call Ω the *sample space*, $\mathscr{F}$ a σ-algebra containing all events $A \subset \Omega$, and $\mathbb{P}$ a probability measure defined on $\mathscr{F}$. We can think of an event A as information and the probability that A happens, i.e. $\mathbb{P}[A]$, is the weight assigned to that information. We require that

$$0 \leq \mathbb{P}[A] \leq 1, \quad \mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\Omega] = 1.$$

Example 3.3. Back to the tossing coin example, we trivially have $\Omega = \{head, tail\}$, $\mathscr{F} = \{\emptyset, \{head\}, \{tail\}, \Omega\}$. The weights are $\mathbb{P}[\emptyset] = 0, \mathbb{P}[\{tail\}] = \mathbb{P}[\{head\}] = \frac{1}{2}$, and $\mathbb{P}[\{head, tail\}] = 1$. △

Two events A and B are independent[3] if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B].$$

One of the central ideas in Bayesian probability is the *conditional probability*[4]. The conditional probability of A on/given B is defined as[5]

This is the corner stone formula to build most of results in this note, make sure that you feel comfortable with it.

$$\boxed{\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]},} \tag{3.3}$$

which can also be rephrased as the probability that A happens *provided* B has already happened. An intuitive way to understand this formula is to consider the probability measure as the standard area (or volume) measure. In this case, the conditional

[2] One can believe that out of n tossing times, $n/3$ times is head and $2n/3$ times is tail if he uses an *unfair* coin.

[3] Probability theory is often believed to be a part of measure theory, but independence is where it departs from the measure theory umbrella.

[4] A more general and rigorous tool is conditional expectation, a particular of which is conditional probability.

[5] This was initially introduced by Kolmogorov, a father of modern probability theory. An elegant derivation of this formula is based on the conditional expectation, but this would take us too deep into the probability theory [16].

measure is nothing more than the ratio of the area of the intersection $A \cap B$ and the area of B.

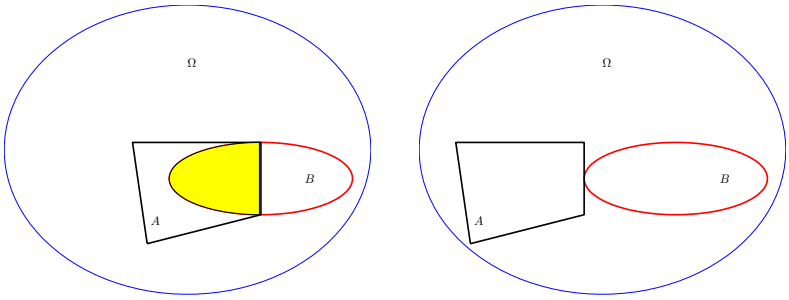
Example 3.4. Assume that we want to roll a dice. Denote B as the event of getting of face bigger than 4, and A the event of getting face 6. Using (3.3) we have

$$\mathbb{P}[A|B] = \frac{1/6}{1/3} = 1/2.$$

We can solve the problem using a more elementary argument. B happens when we either get face 5 or face 6. The probability of getting face 6 when B has already happened is clearly $\frac{1}{2}$. △

The conditional probability can also be understood as the probability when the sample space is restricted to B.

Exercise 3.1. Determine $\mathbb{P}[A|B]$ in Figure 3.1. •



(a) $\mathbb{P}[A|B] =$ (b) $\mathbb{P}[A|B] =$

Fig. 3.1 Demonstration of conditional probability.

Exercise 3.2. Show that the following Bayes formula for conditional probability holds

$$\boxed{\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.} \tag{3.4}$$

•

By inspection, if A and B are mutually independent, we have

$$\mathbb{P}[A|B] = \mathbb{P}[A], \quad \mathbb{P}[B|A] = \mathbb{P}[B].$$

The probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is an abstract object which is useful for theoretical developments, but far from practical considerations. In practice, it is usually cir-

cumvented by probability densities over the *state space*, which are easier to handle and have certain physical meanings. We shall come back to this point in a moment.

Definition 3.3. The state space S is the set containing all the possible outcomes.

3.3 Appendix

Definition 3.4 (σ-algebra). A family $\mathscr{F}$ of subsets of a non-empty set Ω is called a σ-algebra if

- $\emptyset \in \mathscr{F}$,
- $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$ (here A^c is the complement of A), and
- $A_n \in \mathscr{F}$ for all $n \in \mathbb{N}$ implies $\cup_n A_n \in \mathscr{F}$.

Clearly the power set of Ω is a σ-algebra. Let $\mathscr{A}$ be a family of subsets and the intersection of all σ-algebras containing $\mathscr{A}$, i.e. the smallest σ-algebra containing $\mathscr{A}$, is called the σ-algebra generated by $\mathscr{A}$.The σ-algebra generated by all open sets in Ω is called the Borel algebra.

Chapter 4
Random Variables and the Bayes formula

For the sake of clarity we consider the state space S (and also T) as the standard Euclidean space $\mathbb{R}^n$ for now. However, results developed below that do not involve probability densities (with respect to the Lebesgue measure) are also valid for general (e.g. infinite dimensional) state space S. We are in position to introduce the key player, the *random variable*.

Definition 4.1. A random variable m is a *measurable* map[1] from the sample space Ω to the state space S (with $\mathscr{S}$ as its σ-algebra)

$$m : \Omega \ni \omega \mapsto m(\omega) \in S.$$

We call $m(\omega)$ a random variable since we are uncertain about its outcome. In other words, we admit our ignorance about m by calling it a random variable. This ignorance is in turn a direct consequence of the uncertainty in the outcome of elementary event ω.

Definition 4.2. The *probability distribution* (or distribution or law for short) of a random variable m is defined as

$$\mu_m(A) \stackrel{\text{def}}{=} \mathbb{P}\left[m^{-1}(A)\right] = \mathbb{P}\left[\{m \in A\}\right], \quad \forall A \in \mathscr{S}, \tag{4.1}$$

where we have used the popular notation[2]

$$m^{-1}(A) \stackrel{\text{def}}{=} \{m \in A\} \stackrel{\text{def}}{=} \{\omega \in \Omega : m(\omega) \in A\}.$$

From the definition, we can see that the distribution is a probability measure[3] on S. In other words, the random variable m induces a probability measure, defined as μ_m, on the state space S. The key property of the induced probability measure μ_m is

Do you see why?

[1] A map is measurable if through its the inverse a measurable set in S is mapped into a measurable set in Ω.

[2] Rigorously, A must be a measurable subset of S.

[3] In fact, it is the push-forward measure by the random variable m.

the following. The probability for an event A in the state space to happen, denoted as $\mu_m(A)$, is defined as the probability for an event $B = m^{-1}(A)$ in the sample space to happen (see Figure 4.1 for an illustration). The distribution and the *probability density* (unless otherwise stated, density is understood with respect to the Lebesgue measure λ on S) π_m of m obey the following relation

$$\boxed{\mu_m(A) \overset{\text{def}}{=} \int_A \pi_m(m)\,\lambda(dm) \overset{\text{def}}{=} \mathbb{P}[\{m \in A\}], \quad \forall A \subset \mathscr{S}}. \tag{4.2}$$

Do you see this?

where the second equality of the definition is from (4.1). The meaning of random variable $m(\omega)$ can now be seen in Figure 4.1. It maps the event $B \in \Omega$ into the set $A = m(B)$ in the state space such that the "area" under the density function $\pi_M(m)$ and above A is exactly the probability that B happens.

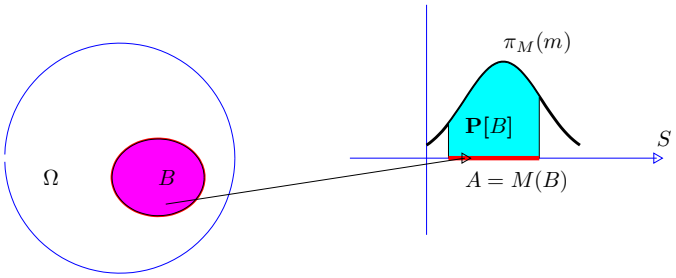


Fig. 4.1 Demonstration of random variable: $\mu_m(A) = \mathbb{P}[B]$.

We deduce the change of variable formula

$$\boxed{d\mu_m(m) \overset{\text{def}}{=} \mu_m(dm) = \pi(m)\,\lambda(dm) \overset{\text{def}}{=} \pi(m)\,d\lambda(m)},$$

where we write $\pi(m)$ instead of $\pi_m(m)$ if there is no ambiguity. *Note that $d\mu_m(m)$ is nothing more than the differential area under the curve $\pi(m)$ around m.* From an analogy to the differential calculus, we can formally rewrite the above formula as

$$\boxed{\frac{d\mu_m}{d\lambda}(m) = \pi(m)}, \tag{4.3}$$

which is an instance of the Radon-Nikodym derivative. In fact, one can show that ordinary derivative is a special case of the Radon-Nikodym derivative.

Exercise 4.1 (Ordinary derivative is a special case of Radon-Nikodym derivative). Take $f(x) = F'(x)$, where F is defined in Theorem 4.2. Show that

$$\frac{d\nu}{d\lambda}(x) = f(x) = F'(x),$$

where ν is the measure in Theorem 4.2, and λ is the standard Lesbegue measure on $\mathbb{R}$. •

Solution: We have

$$\nu((a,b]) = F(b) - F(a) = \int_a^b f(x)\, d\lambda(x),$$

which, by definition of the Radon-Nikodym derivative, implies

$$\frac{d\nu}{d\lambda}(x) = f(x) = F'(x),$$

and thus the Radon-Nikodym derivative reduces to the ordinary derivative in this case.

Thus, *when we say $\pi(m)$ is the density of the random variable m, unless otherwise stated, we implicity mean that the Radon-Nikodym derivative of its probability measure with respect to the Lebesgue measure is $\pi(m)$.*

Definition 4.9 shows that if μ_r and λ obeys (4.3), then μ_r is *absolutely continuous* with respect to λ. *Do you see this?*

Remark 4.1. In theory, we introduce the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ in order to compute the probability of a subset[4] A in the state space S, and this is essentially the meaning of (4.1). However, once we know the probability density function $\pi_m(m)$, we can operate directly on the state space S without the need for referring back to probability space $(\Omega, \mathscr{F}, \mathbb{P})$, as shown in definition (4.2). This is the key observation, a consequence of which is that we simply ignore the underlying probability space in practice, since we don't need them in computation of probability in the state space. However, to intuitively understand the source of randomness, we need to go back to the probability space where the outcome of all events, except Ω, is uncertain. As a result, the pair $(S, \pi_m(m))$ contains complete information describing our ignorance about the outcome of random variable m. To the rest of this note, we shall work directly on the state space.

Remark 4.2. At this point, one may wonder what is the point of introducing the abstract probability space $(\Omega, \mathscr{F}, \mathbb{P})$ because it seems to unnecessarily make life more complicated? Well, its introduction is two fold. First, as discussed above, the probability space not only shows the origin of randomness but also provides the

[4] Again, it needs to be measurable.

probability measure $\mathbb{P}$ for the computation of the randomness; it is also used to define random variables and furnishes a decent understanding about them. Second, the concepts of distribution and density in (4.1) and (4.2), which are introduced for random variable m, a measurable map from Ω to S, are valid for measurable maps[5] from an arbitrary space V to another space W. Here, W plays the role of S, and V the role of Ω on which we have a probability measure. For example, later in Section 5.1, we introduce the parameter-to-observable map $h(m) : S \to \mathbb{R}^r$, then S plays the role of Ω and $\mathbb{R}^r$ of S in (4.1) and (4.2).

Definition 4.3. The *expectation* or the *mean* of a random variable m is the quantity

$$\mathbb{E}[m] \stackrel{\text{def}}{=} \int_S m\pi(m)\, dm = \overline{m}, \tag{4.4}$$

and the *variance* is

$$\mathbb{V}ar[m] \stackrel{\text{def}}{=} \mathbb{E}\left[(m-\overline{m})^2\right] \stackrel{\text{def}}{=} \int_S (m-\overline{m})^2\, \pi(m)\, dm.$$

Example 4.1. Let $\Omega = [-2,2]$ and define

$$\mathbb{P}[A \in \Omega] = \int_A \frac{1}{4}\, d\Omega,$$

where $d\Omega$ is the standard Lesbegue measure. Define a random variable $m : \Omega \to \mathbb{R}$ as

$$m(\omega) = \begin{cases} 2 & \text{if } \omega \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

We can compute the mean of m as

$$\begin{aligned} \mathbb{E}[m] \stackrel{\text{def}}{=} \int_{\mathbb{R}} m\pi(m)\, dm = \int_{\mathbb{R}} m\, \mu_m &= \int_\Omega m(\omega)\, \mathbb{P}[d\Omega] \\ &= \int_\Omega \frac{1}{4} m(\omega)\, d\Omega = \int_0^2 \frac{1}{4} \times 2\, d\Omega = 1. \end{aligned}$$

△

This example shows that in general one *should not expect* the random variable under consideration can take its mean as a realization.

Exercise 4.2. A Gaussian random variable has the density given as

$$\pi(m) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(m-\overline{m})^2\right),$$

where $\overline{m}$ and σ are known as the mean and the deviation of the Gaussian distribution.

Show that $\mathbb{E}[m] = \overline{m}$ and $\mathbb{V}ar[m] = \sigma^2$. •

[5] Again, they must be measurable.

As we will see, the Bayes formula for probability densities is about the joint density of two or more random variables. So let us define the joint distribution and joint density of two random variables here.

Definition 4.4. Denote μ_{my} and π_{my} as the joint distribution and density, respectively, of two random variables m with values in S and y with values in T (with $\mathscr{T}$ as its σ-algebra) defined on the same probability space, then the joint distribution function and the joint probability density, in the light of (4.2), satisfy

$$\mu_{my}\left(\{m\in A\},\{y\in B\}\right)\overset{\text{def}}{=}\int_{A\times B}\pi_{my}(m,y)\,dmdy,\quad \forall A\times B\subset\mathscr{S}\times\mathscr{T}, \tag{4.5}$$

where the notation $A\times B\subset\mathscr{S}\times\mathscr{T}$ simply means that $A\in\mathscr{S}$ and $B\in\mathscr{T}$.

We say that m and y are independent if

$$\mu_{my}\left(\{m\in A\},\{y\in B\}\right)=\mu_m(A)\,\mu_y(B),\quad \forall A\times B\subset\mathscr{S}\times\mathscr{T},$$

or if

$$\pi_{my}(m,y)=\pi_m(m)\,\pi_y(y).$$

Definition 4.5. The *marginal* density of m is the probability density of m when y may take on any value, i.e.,

$$\pi_m(m)=\int_T\pi_{my}(m,y)\,dy.$$

Similarly, the marginal density of y is the density of y regardless m, namely,

$$\pi_y(y)=\int_S\pi_{my}(m,y)\,dm.$$

Before deriving the Bayes formula, we define conditional density $\pi(m|y)$ in the same spirit as (4.2) as

$$\mu_{m|y}\left(\{m\in A\}|y\right)=\int_A\pi(m|y)\,dm.$$

Let us prove the following important result.

Theorem 4.1. *The conditional density of m given $y=y$ is given by*

$$\boxed{\pi(m|y)=\frac{\pi(m,y)}{\pi(y)}}.$$

Proof. From the definition of conditional probability (3.3), we have

$$\begin{aligned}
&\mu_{m|y}\left(\{m\in A\}|y\right)=\mathbb{P}\left[\{m\in A\}|y=y\right] && \text{(definition (4.1))}\\
&=\lim_{\Delta y\to 0}\frac{\mathbb{P}\left[\{m\in A\},y\le y'\le y+\Delta y\right]}{\mathbb{P}\left[y\le y'\le y+\Delta y\right]} && \text{(definition (3.3))}\\
&=\lim_{\Delta y\to 0}\frac{\int_A \pi(m,y)\,dm\Delta y}{\pi(y)\,\Delta y} && \text{(definitions (4.2), (4.5))}\\
&=\int_A \frac{\pi(m,y)}{\pi(y)}\,dm,
\end{aligned}$$

which ends the proof.

By symmetry, we have

$$\pi(m,y)=\pi(m|y)\,\pi(y)=\pi(y|m)\,\pi(m),$$

from which the well-known Bayes formula for finite dimensional state spaces follows

$$\boxed{\pi(m|y)=\frac{\pi(y|m)\,\pi(m)}{\pi(y)}}. \tag{4.6}$$

Exercise 4.3. Prove directly the Bayes formula for conditional density (4.6) using the Bayes formula for conditional probability (3.4). •

Solution:

$$\begin{aligned}
&\mu_{m|y}\left(\{m\in A\}|y\right)=\mathbb{P}\left[\{m\in A\}|y=y\right] && \text{(definition (4.1))}\\
&=\lim_{\Delta y\to 0}\frac{\mathbb{P}\left[y\le y'\le y+\Delta y|\,\{m\in A\}\right]\mathbb{P}\left[\{m\in A\}\right]}{\mathbb{P}\left[y\le y'\le y+\Delta y\right]} && \text{(definition (3.4))}\\
&=\lim_{\Delta y\to 0}\frac{\int_A\left(\int_{\Delta y}\pi(y|m)\,dy\right)\pi(m)\,dm}{\pi(y)\,\Delta y} && \text{(definitions (4.2), (4.5))}\\
&=\lim_{\Delta y\to 0}\frac{\Delta y\int_A\pi(y|m)\,\pi(m)\,dm}{\pi(y)\,\Delta y}\\
&=\int_A\frac{\pi(y|m)\,\pi(m)}{\pi(y)}\,dm,
\end{aligned}$$

which ends the proof.

Definition 4.6 (Likelihood). We call $\pi(y|m)$ the likelihood. It is the probability density of y given $m=m$.

Definition 4.7 (Prior). We call $\pi(m)$ the prior. It is the probability density of m regardless y. The prior encodes, in the Bayesian framework, all information before any observations/data are made.

Definition 4.8 (Posterior). The density $\pi(m|y)$ is called the *posterior*, the distribution of parameter m given the measurement $y=y$, and it is the solution of the Bayesian inverse problem under consideration.

4.1 Appendix

Theorem 4.2. *Suppose $F : \mathbb{R} \to \mathbb{R}$ is: i) non-decreasing, and 2) right continuous. There is a unique measure ν on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that*

$$\nu((a,b]) = F(b) - F(a), \quad \forall a,b,$$

where $\mathscr{B}(\mathbb{R})$ is the Borel algebra on $\mathbb{R}$.

Definition 4.9 (Absolutely continuous of measures). We say ν is absolutely continuous with respect to μ, denoted by $\nu \ll \mu$, if $\mu(A) = 0$ implies $\nu(A) = 0$.

Chapter 5
Construction of the likelihood

Now if we define the the posterior and prior measures as

$$d\mu(m) \overset{\text{def}}{=} \pi(m|y)\, d\lambda(m), \quad d\nu(m) \overset{\text{def}}{=} \pi(m)\, d\lambda(m),$$

we can rewrite the Bayes formula as

$$\boxed{\frac{d\mu}{d\nu}(m) = \frac{\pi(y|m)}{\pi(y)} \propto \pi(y|m).} \tag{5.1}$$

Again, from our interpretation in Figure 4.1, $d\mu(m)$ and $d\nu(m)$ can be considered as the differential areas around m under the curves $\pi(m|y)$ and $\pi(m)$, respectively.

Note that we can ignore $\pi(y)$ in the last expression since $\pi(y)$ is independent of m, and hence can be considered as a proportional constant.

It is important to point out that, unlike the standard form (4.6) *that is only valid for finite dimensional cases, the Bayes formula in the form* (5.1) *is also valid for infinite dimensional problems. Of course, when both prior and posterior measures admit a density with respect to a reference measure, Lebesgue measure for example, then* (5.1) *reduces to* (4.6)*.*

Definition 5.1. The *conditional mean* is defined as

$$\mathbb{E}[m|y] = \int_S m\pi(m|y)\, dm.$$

Exercise 5.1. Show that

$$\mathbb{E}[m] = \int_T \mathbb{E}[m|y]\, \pi(y)\, dy.$$

•

Solution:

$$\mathbb{E}[m] = \int_S m\pi(m)\, dm = \int_S m \int_T \pi(m|y)\, \pi(y)\, dy\, dm.$$

where we have used the definition of marginal probability density.

5.1 Construction of likelihood

In this section, we present a popular approach to construct the likelihood. We begin with the additive noise case. The ideal deterministic model is given by

$$y = h(m),$$

where $y \in \mathbb{R}^r$. But due to random additive noise e, we have the following statistical model instead

$$y^{obs} = h(m) + e, \tag{5.2}$$

where y^{obs} is the actual observation rather than $y = f(m)$. Since the noise comes from external sources, in this note, it is assumed to be independent of m. In the likelihood modeling, we pretend to have realization(s) of m and the task is to construct the distribution of y^{obs}. From (5.2), one can see that the randomness in y^{obs} is the randomness in e shifted by an amount $h(m)$, see Figure 5.1, and hence

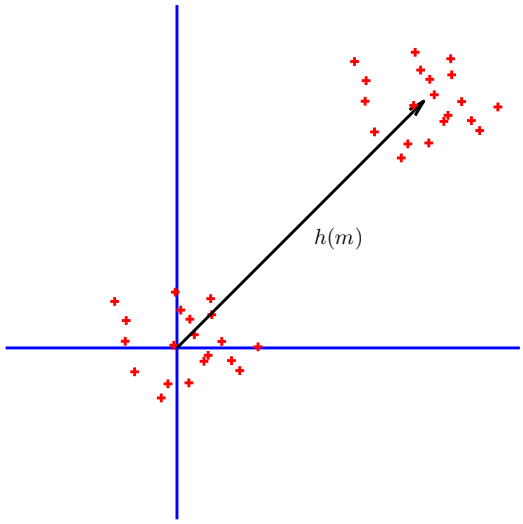


Fig. 5.1 Likelihood model with additive noise.

$\pi_{y^{obs}|m}\left(y^{obs}|m\right) = \pi_e\left(y^{obs} - h(m)\right)$. More rigorously, assume that both y^{obs} and e are random variables on a same probability space, we have

$$\int_A \pi_{y^{obs}|m}\left(y^{obs}|m\right) dy^{obs} \overset{\text{def}}{=} \mu_{y^{obs}|m}(A) \overset{(4.1)}{=} \mu_e(A - h(m))$$
$$= \int_{A-h(m)} \pi_e(e)\, de \overset{\text{change of variable}}{=} \int_A \pi_e\left(y^{obs} - h(m)\right) dy^{obs}, \quad \forall A \subset S,$$

which implies

$$\pi_{y^{obs}|m}\left(y^{obs}|m\right)=\pi_e\left(y^{obs}-h(m)\right).$$

Exercise 5.2. We consider the following multiplicative noise case

$$y^{obs}=eh(m). \tag{5.3}$$

Show that the likelihood for multiplicative noise model (5.3) has the following form

$$\pi_{y^{obs}|m}\left(y^{obs}|m\right)=\frac{\pi_e\left(y^{obs}/h(m)\right)}{h(m)},\quad h(m)\neq 0. \tag{5.4}$$

Now look at the following multiplicative noise model

$$y_i^{obs}=e_i\times m_i,\quad i=1,\ldots,n,$$

where each e_i is independent, identically distributed by the following log-normal distribution

$$W_i=\log(e_i)\sim\mathcal{N}\left(w_0,\sigma^2\right),\quad w_0=\log(\alpha_0).$$

Determine the likelihood

$$\pi_{\mathbf{y}^{obs}|\mathbf{m}}\left(\mathbf{y}^{obs}|\mathbf{m}\right),$$

where $\mathbf{y}^{obs}\overset{\text{def}}{=}\left[y_1^{obs},\ldots,y_n^{obs}\right]$ and $\mathbf{m}\overset{\text{def}}{=}[m_1,\ldots,m_n]$. •

Solution: Since e_i are independent, y_i^{obs} are also independent. Using (5.4) we have

$$\pi_{\mathbf{y}^{obs}|\mathbf{m}}\left(\mathbf{y}^{obs}|\mathbf{m}\right)\propto\Pi_{i=1}^n\frac{1}{m_i}\pi_e\left(y_i^{obs}/m_i\right).$$

We next determine π_e. From $dw=\dfrac{de}{e}$ and the independence we have

$$\pi_e(e)\propto\frac{1}{e}\exp\left(-\frac{1}{\sigma^2}\left(\log(e)-w_0\right)^2\right).$$

Exercise 5.3. Can you generalize the result for the noise model $e=g\left(y^{obs},h(x)\right)$? •

For concreteness, let us consider the following one dimensional deblurring (deconvolution) problem

$$g(s_j)=\int_0^1 a(s_j,t)f(t)\,dt+e(s_j),\quad 0\leq j\leq n,$$

where $a(s,t)=\dfrac{1}{\sqrt{2\pi\beta^2}}\exp(-\dfrac{1}{2\beta^2}(t-s)^2)$ is a given kernel, and $s_j=j/n, j=0,\ldots,n$ the mesh points. Our task is to reconstruct $f(t):[0,1]\to\mathbb{R}$ from the noisy

observations $g(s_j)$, $j = 0, \ldots, n$. To cast the function reconstruction problem, which is in infinite dimensional space, into a reconstruction problem in $\mathbb{R}^n$, we discretize $f(t)$ on the same mesh and use simple rectangle method for the integral. Let us define $y^{obs} = [g(s_0), \ldots, g(s_n)]^T$, $m = (f(s_0), \ldots, f(s_n))^T$, and $\mathscr{A}_{i,j} = a(s_i, s_j)/n$, then the discrete deconvolution problem reads

$$y^{obs} = \mathscr{A} m + e.$$

Here, we assume $e \sim \mathscr{N}\left(0, \sigma^2 I\right)$, where I is the identity matrix in $\mathbb{R}^{(n+1)\times(n+1)}$. Since Section 5.1 suggests the likelihood of the form

$$\pi\left(y^{obs}|m\right) = \mathscr{N}\left(\mathscr{A} m, \sigma^2 I\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(y^{obs} - \mathscr{A} m\right)^T \left(y^{obs} - \mathscr{A} m\right)\right),$$

the Bayesian solution to our inverse problem is, by virtue of the Bayes formula (4.6), given by

$$\boxed{\pi_{\text{post}}\left(m|y^{obs}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(y^{obs} - \mathscr{A} m\right)^T \left(y^{obs} - \mathscr{A} m\right)\right) \times \pi_{\text{prior}}(m)}, \quad (5.5)$$

where we have ignored the denominator $\pi\left(y^{obs}\right)$ since it does not depend on the parameter of interest m. Thus, the posterior is "completely determined" once the prior is given and this is the subject of the next section.

Chapter 6
Prior Elicitation

As discussed previously, the prior belief depends on a person's knowledge and experience. In order to obtain a good prior, one sometimes needs to perform some expert elicitation. Nevertheless, there is no universal rule and one has to be careful in constructing a prior. In fact, prior construction is a subject of current research, and it is problem-dependent.

6.1 Smooth priors

In this section, we believe that the unknown function $f(t)$ is smooth, which can be translated into, among other possibilities, the following simplest requirement on the pointwise values $f(s_i)$, and hence m_i,

$$m_i = \frac{1}{2}\left(m_{i-1} + m_{i+1}\right), \tag{6.1}$$

that is, the value of $f(s)$ at a point is more or less the same of its neighbor. But, this is by no means the correct behavior of the unknown function $f(s)$. We therefore admit some uncertainty in our belief (6.1) by adding an *innovative* term W_j such that

$$m_i = \frac{1}{2}\left(m_{i-1} + m_{i+1}\right) + W_j,$$

where $W \sim \mathcal{N}\left(0, \gamma^2 I\right)$. The standard deviation γ determines how much the reconstructed function $f(t)$ departs from the smoothness model (6.1). In terms of matrices, we obtain

$$Lm = W,$$

where L is given by

$$L = \frac{1}{2}\begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n+1)},$$

which is the second order finite difference matrix approximating the Laplacian Δf. Indeed,

$$\Delta f(s_j) \approx n^2 \left(Lm\right)_j. \tag{6.2}$$

Suppose $\det(L) \neq 0$, the prior distribution is therefore given by (using the technique in Section 5.1)

$$\pi_{\text{pre}} \propto \exp\left(-\frac{1}{\gamma^2}\|Lm\|^2\right). \tag{6.3}$$

But L has rank of $n-1$, and hence π_{pre} is a degenerate Gaussian density in $\mathbb{R}^{n+1}$. In fact, (6.3) is not valid since $\det(L) = 0$. The reason is that we have not specified the smoothness of $f(s)$ at the boundary points. In other words, we have not specified any boundary conditions for the Laplacian $\Delta f(s)$. This is a crucial point in prior elicitation via differential operators. One needs to make sure that the operator is positive definite by incorporating some well-posed boundary conditions. Throughout the lecture notes, unless otherwise stated, $\|\cdot\|$ denotes the usual Euclidean norm.[1]

Let us first consider the case with zero Dirichlet boundary condition, that is, we believe that $f(s)$ is smooth and (close to) zero at the boundaries, then

$$\begin{aligned} m_0 &= \frac{1}{2}\left(m_{-1} + m_1\right) + W_0 = \frac{1}{2}m_1 + W_0, \quad W_0 \sim \mathcal{N}\left(0, \gamma^2\right) \\ m_n &= \frac{1}{2}\left(m_{n-1} + m_{n+1}\right) + W_n = \frac{1}{2}m_{n-1} + W_n, \quad W_n \sim \mathcal{N}\left(0, \gamma^2\right). \end{aligned}$$

Note that we have extended $f(s)$ by zero outside the domain $[0,1]$ since we "know" that it is smooth. Consequently, we have $L_D m = W$ with

$$L_D = \frac{1}{2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{6.4}$$

which is the second order finite difference matrix corresponding to zero Dirichlet boundary conditions. The prior density in this case reads

$$\pi^D_{\text{prior}}(m) \propto \exp\left(-\frac{1}{2\gamma^2}\|L_D m\|^2\right). \tag{6.5}$$

[1] The ℓ^2-norm if you wish

It is instructive to draw some random realizations from π^D_{prior} (we are ahead of ourselves here since sampling will be discussed in Chapter **??**), and we show five of them in Figure 6.1 together with the prior standard deviation curve. As can be seen, all the draws are almost zero at the boundary and the prior variance (uncertainty) is close to zero as well. This is not surprising since our prior belief says so. How do we compute the standard deviation curve? Well, it is straightforward. We first compute the pointwise variance as

Why are they not exactly zero?

$$\mathbb{V}ar[m_j] \stackrel{\text{def}}{=} \mathbb{E}[m_j^2] = e_j^T \left(\int_{\mathbb{R}^{n+1}} mm^T \pi^D_{\text{prior}} dm \right) e_j \stackrel{\text{def}}{=} \gamma^2 e_j^T \left(L_D^T L_D \right)^{-1} e_j,$$

where e_j is the jth canonical basis vector in $\mathbb{R}^{n+1}$, and we have used the fact that the prior is Gaussian in the last equality. So we in fact plot the square root of the diagonal of $\gamma^2 \left(L_D^T L_D \right)^{-1}$, the covariance matrix, as the standard deviation curve. One can see that the uncertainty is largest in the middle of the domain since it is farthest from the constrained boundary. The points closer to the boundaries have smaller variance, that is, they are more correlated to the "known" boundary data, and hence less uncertain.

Do we really have the complete continuous curve?

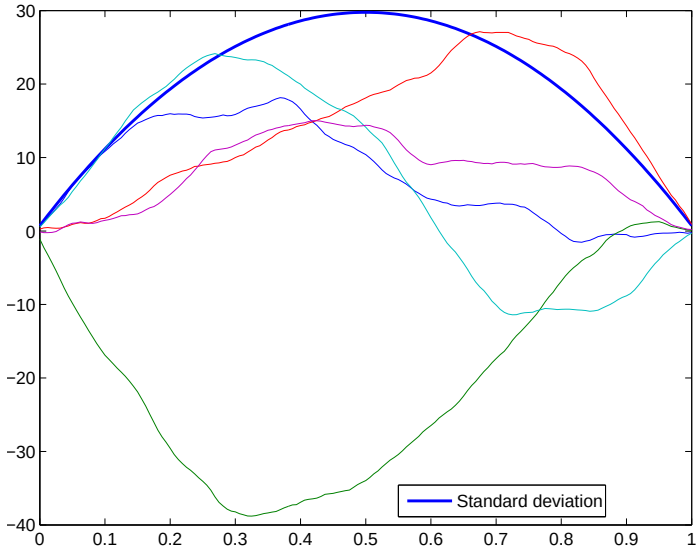


Fig. 6.1 Prior random draws from π^D_{prior} together with the standard deviation curve.

Now, you may ask why $f(s)$ must be zero at the boundary, and you are right! There is no reason to believe that must be the case. However, we don't know the exact values of $f(s)$ at the boundary either, even though we believe that we may have non-zero Dirichlet boundary condition. If this is the case, we have to admit our

ignorance and let the data from the likelihood correct us in the posterior. To be consistent with the Bayesian philosophy, if we do not know anything about boundary conditions, let them be, for convenience, Gaussian random variables such as

$$m_0 \sim \mathcal{N}\left(0, \frac{\gamma^2}{\delta_0^2}\right), \quad m_n \sim \mathcal{N}\left(0, \frac{\gamma^2}{\delta_n^2}\right).$$

Hence, the prior can now be written as

$$\pi^R_{\text{prior}}(m) \propto \exp\left(-\frac{1}{2\gamma^2}\|L_R m\|^2\right), \tag{6.6}$$

where

$$L_R = \frac{1}{2}\begin{bmatrix} 2\delta_0 & 0 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & 0 & 2\delta_n \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

A question that immediately arises is how to determine δ_0 and δ_n. Since the boundary values are now independent random variables, we are less certain about them compared to the previous case. But to which uncertain level we want them to be? Well, let's make every values equally uncertain, meaning we have the same ignorance about the values at these points, that is, we would like to have the same variances everywhere. To approximately accomplish this, we require

$$\mathbb{V}ar[m_0] = \frac{\gamma^2}{\delta_0^2} = \mathbb{V}ar[m_n] = \frac{\gamma^2}{\delta_n^2} = \mathbb{V}ar\left[m_{[n/2]}\right] = \gamma^2 e_{[n/2]}^T \left(L_R^T L_R\right)^{-1} e_{[n/2]},$$

where $[n/2]$ denotes the largest integer smaller than $n/2$. It follows that

$$\delta_0^2 = \delta_n^2 = \frac{1}{e_{[n/2]}^T \left(L_R^T L_R\right)^{-1} e_{[n/2]}}.$$

Is it sensible to do so?

However, this requires to solve a nonlinear equation for $\delta_0 = \delta_n$, since L_R depends on them. To simplify the computation, we use the following approximation

$$\delta_0^2 = \delta_n^2 = \frac{1}{e_{[n/2]}^T \left(L_D^T L_D\right)^{-1} e_{[n/2]}}.$$

Again, we draw five random realizations from π^R_{prior} and put them together with the standard deviation curve in Figure 6.2. As can be observed, the uncertainty is more or less the same at every point and prior realizations are no longer constrained to have zero boundary conditions.

Exercise 6.1. Consider the following general scheme

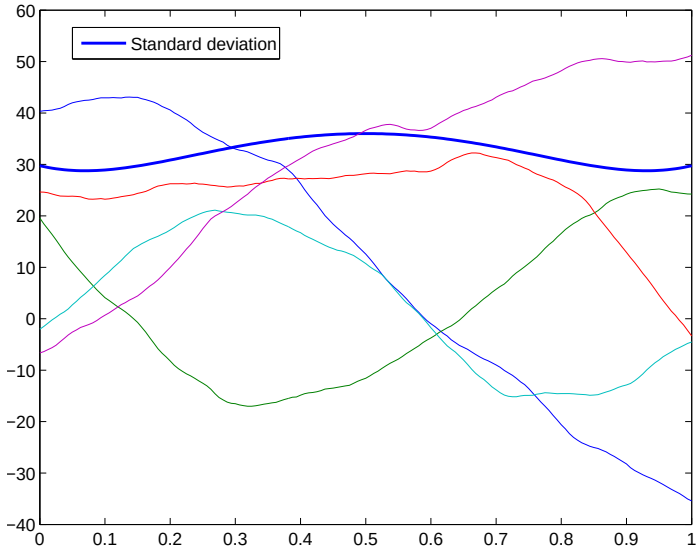


Fig. 6.2 Prior random draws from π^R_{prior} together with the standard deviation curve.

$$m_i = \lambda_i m_{i-1} + (1-\lambda_i) m_{i+1} + e_i, \quad 0 \le \lambda_i \le 1.$$

Convince yourself that by choosing a particular set of λ_i, you can recover all the above prior models. Replace `BayesianPriorElicitation.m` by a generic code with input parameters λ_i. Experience new prior models by using different values of λ_i (those that don't reproduce priors presented in the text). •

Exercise 6.2. Construct a prior with a non-zero Dirichlet boundary condition at $s = 0$ and zero Neumann boundary condition at $s = 1$. Draw a few samples together with the variance curve to see whether your prior model indeed conveys your belief. •

6.2 "Non-smooth" priors

We first consider the case in which we believe that $f(s)$ is still smooth but may have discontinuities at known locations on the mesh. Can we design a prior to convey this belief? A natural approach is to require that m_j is equal to m_{j-1} plus a random jump, i.e.,

$$m_j = m_{j-1} + e_j,$$

where $e_j \sim \mathcal{N}\left(0, \gamma^2\right)$, and for simplicity, let us assume that $m_0 = 0$. The prior density in this case would be

$$\pi_{\text{pren}}(m) \propto \exp\left(-\frac{1}{2\gamma^2}\|L_N m\|^2\right), \tag{6.7}$$

where

$$L_N = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

But, if we think that there is a particular big jump, relative to others, from m_{j-1} to m_j, then the mathematical translation of this belief is $e_j \sim \mathcal{N}\left(0, \frac{\gamma^2}{\theta^2}\right)$ with $\theta < 1$. The corresponding prior in this case reads

$$\pi^O_{\text{prior}}(m) \propto \exp\left(-\frac{1}{2\gamma^2}\|JL_N m\|^2\right), \tag{6.8}$$

with

$$J = \text{diag}\left(1,\ldots,1,\underbrace{\theta}_{j\text{th index}},1,\ldots,1\right).$$

Let's draw some random realizations from $\pi^O_{\text{prior}}(m)$ in Figure 6.3 with $n = 160$, $j = 80$, $\gamma = 1$, and $\theta = 0.01$. As desired, all realizations have a sudden jump at $j = 80$, and the standard deviation of the jump is $1/\theta = 100$. In addition, compared to priors in Figure 6.1 and 6.2, the realizations from $\pi^O_{\text{prior}}(m)$ are less smooth, which confirms that our belief is indeed conveyed.

Exercise 6.3. Use `BayesianPriorElicitation.m` to construct examples with 2 or more sudden jumps and plot a few random realizations to see whether your belief is conveyed. •

A more interesting and more practical situation is the one in which we don't know how many jump discontinuities and their locations. A natural prior in this situation is a generalized version of (6.8), e.g.,

$$\pi^M_{\text{prior}}(m) \propto C(M)\exp\left(-\frac{1}{2\gamma^2}\|ML_N m\|^2\right), \tag{6.9}$$

with

$$M = \text{diag}(\theta_1,\ldots,\theta_n),$$

where θ_i, $i = 1,\ldots,n$, are unknown. In fact, these are called *hyper-parameters* and one can determine them using information from the likelihood; the readers are referred to [10] for the details.

What is wrong with this?

Exercise 6.4. Modify the scheme in Exercise 6.1 to include priors with sudden jumps. •

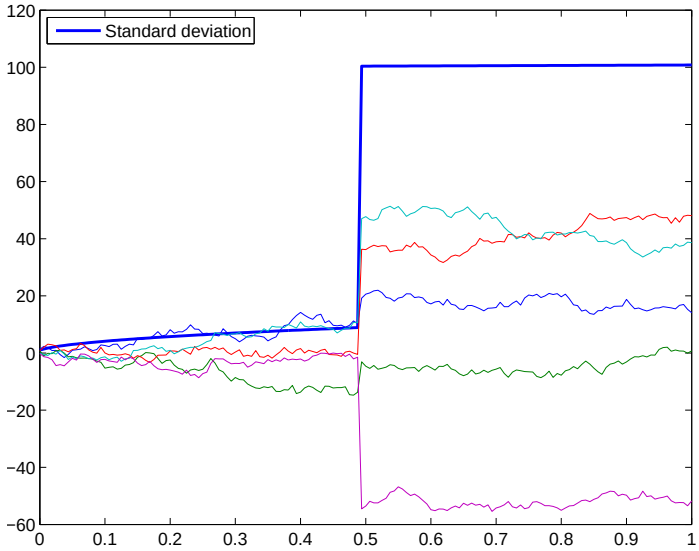


Fig. 6.3 Prior random draws from π^{O}_{prior} together with the standard deviation curve.

Chapter 7
Bayesian inverse solution versus deterministic inverse solution

Abstract In this note, we are interested in solving inverse problems using statistical techniques.

7.1 Posterior as the solution to Bayesian inverse problems

In this section, we explore the posterior (5.5), the solution of our Bayesian problem, given the likelihood in Section 5.1 and priors in Chapter 6.

To derive results that are valid for all priors discussed so far, we work with the following generic prior

$$\pi_{\text{prior}}(m) \propto \exp\left(-\frac{1}{2\gamma^2}\left\|\Gamma^{-\frac{1}{2}}m\right\|^2\right),$$

where $\Gamma^{-\frac{1}{2}} \in \{L_D, L_A, HL_N\}$, each of which again presents a different belief. The Bayesian solution (5.5) can be now written as

$$\pi_{\text{post}}\left(m|y^{obs}\right) \propto \exp\left(-\underbrace{\left[\frac{1}{2\sigma^2}\left\|y^{obs}-\mathscr{A}m\right\|^2+\frac{1}{2\gamma^2}\left\|\Gamma^{-\frac{1}{2}}m\right\|^2\right]}_{T(m)}\right),$$

where $T(m)$ is the familiar (to you I hope) *Tikhonov functional*; it is sometimes called the *potential*. We re-emphasize here that the Bayesian solution is the posterior probability density, and if we draw samples from it, we want to know what the most likely function m is going to be. In other words, we ask for the most probable point m in the posterior distribution. This point is known as the *Maximum A Posteriori (MAP)* estimator/point, namely, the point at which the posterior density is maximized. Let us denote this point as m_{MAP}, and we have

$$m_{MAP} \overset{\text{def}}{=} \arg\max_m \pi_{\text{post}}\left(m|y^{obs}\right) = \arg\min_m T(m).$$

Hence, the MAP point is exactly the deterministic solution of the Tikhonov functional!

This is fundamental. If you have not seen this, prove it!

Since both likelihood and prior are Gaussian, the posterior is also a Gaussian. For our case, the resulting posterior Gaussian reads

$$\begin{aligned}\pi_{\text{post}}\left(m|y^{obs}\right) &\propto \exp\left(-\frac{1}{2}\left\|m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}\right\|_H^2\right)\\ &= \exp\left(-\frac{1}{2}\left(m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}, H\left(m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}\right)\right)\right)\\ &\overset{\text{def}}{=} \exp\left(-\frac{1}{2}\left(m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}, \Gamma_{\text{post}}^{-1}\left(m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}\right)\right)\right)\end{aligned}$$

where

$$\boxed{H = \frac{1}{\sigma^2}\mathscr{A}^T\mathscr{A} + \frac{1}{\gamma^2}\Gamma^{-1}},$$

is the Hessian of the Tikhonov functional (aka the regularized misfit), and we have used the weighted norm $\|\cdot\|_H^2 = \left\|H^{\frac{1}{2}}\cdot\right\|^2$.

Exercise 7.1. Show that the posterior is indeed a Gaussian, i.e.,

$$\pi_{\text{post}}\left(m|y^{obs}\right) \propto \exp\left(-\frac{1}{2}\left\|m-\frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs}\right\|_H^2\right).$$

•

The other important point is that the posterior covariance matrix is precisely the inverse of the Hessian of the regularized misfit, i.e.,

$$\boxed{\Gamma_{\text{post}} = H^{-1}}.$$

Last, but not least, we have showed that the MAP point is given by

$$m_{MAP} = \frac{1}{\sigma^2}H^{-1}\mathscr{A}^T y^{obs} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathscr{A}^T\mathscr{A} + \frac{1}{\gamma^2}\Gamma^{-1}\right)^{-1}\mathscr{A}^T y^{obs},$$

which is, again, exactly the solution of the Tikhonov functional for linear inverse problem.

Exercise 7.2. Show that m_{MAP} is also the least squares solution of the following over-determined system

$$\begin{bmatrix}\frac{1}{\sigma}\mathscr{A}\\ \frac{1}{\gamma}\Gamma^{-\frac{1}{2}}\end{bmatrix} m = \begin{bmatrix}\frac{1}{\sigma}y^{obs}\\ 0\end{bmatrix}$$

•

Exercise 7.3. Show that the posterior mean, which is in fact the conditional mean, is precisely the MAP point. •

Since the covariance matrix, generalization of the variance in multi-dimensional spaces, represents the uncertainty, quantifying the uncertainty in the MAP estimator is ready by simply computing the inverse the Hessian matrix. Let's us now numerically explore the Bayesian posterior solution.

We choose $\beta = 0.05$, $n = 100$, and $\gamma = 1/n$. The truth underlying function that we would like to invert for is given by

$$f(t) = 10(t-0.5)\exp\left(-50(t-0.5)^2\right) - 0.8 + 1.6t.$$

The noise level is taken to be the 5% of the maximum value of $f(s)$, i.e. $\sigma = 0.05 \max_{s\in[0,1]} |f(s)|$.

We first consider the belief described by π^D_{prior} in which we think that $f(s)$ is zero at the boundaries. Figures 7.1 plots the MAP estimator, the truth function $f(s)$, and the predicted uncertainty. As can be observed, the MAP is in good agreement with the truth function inside the interval $[0,1]$, though it is far from recovering $f(s)$ at the boundaries. This is the price we have to pay for not admitting our ignorance about the boundary values of $f(s)$. The likelihood in fact sees this discrepancy in the prior knowledge and tries to make correction by lifting the MAP away from 0, but not enough to be a good reconstruction. The reason is that our incorrect prior is strong enough such that the information from the data y^{obs} cannot help much.

Exercise 7.4. Can you make the prior less strong? *Change some parameter to make prior contribution less!* Use `BayesianPosterior.m` to test your answer. Is the prediction better in terms of satisfying the boundary conditions? Is the uncertainty smaller? If not, why? •

On the other hand, if we admit this ignorance and use the corresponding prior π^D_{prior}, we see much better reconstruction in Figure 7.2. In this case, we in fact let the information from the data y^{obs} determine the appropriate values for the Dirichlet boundary conditions rather than setting them to zero. By doing this, we allow the likelihood and the prior to be well-balanced leading to good reconstruction and uncertainty quantification.

Exercise 7.5. Play with `BayesianPosterior.m` by varying γ, the data misfit (or the likelihood) contribution, and σ, the regularization (or the prior) contribution. •

Exercise 7.6. Use your favorite deterministic inversion approach to solve the above deconvolution problem and then compare it with the solution in Figure 7.2. •

Now consider the case in which the truth function has a jump discontinuity at $j = 70$. Assume we also know that the magnitude of the jump is 10. In particular, we take the truth function $f(s)$ as the following step function

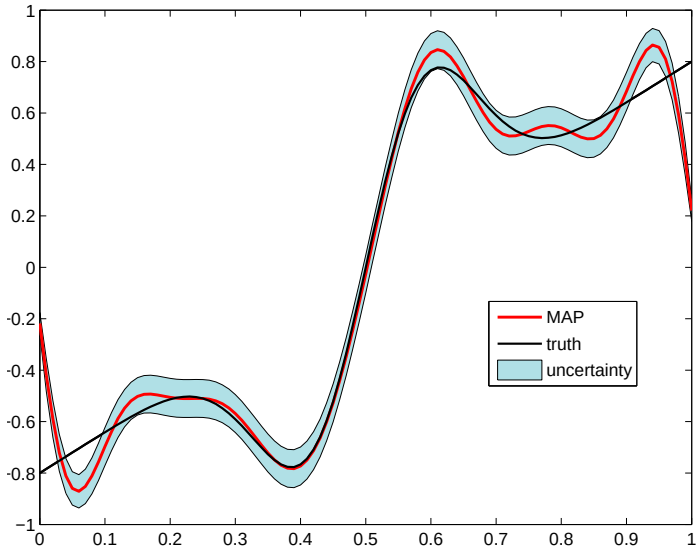


Fig. 7.1 The MAP estimator, the truth function, and the predicted uncertainty (95% credibility region) using π^D_{prior}.

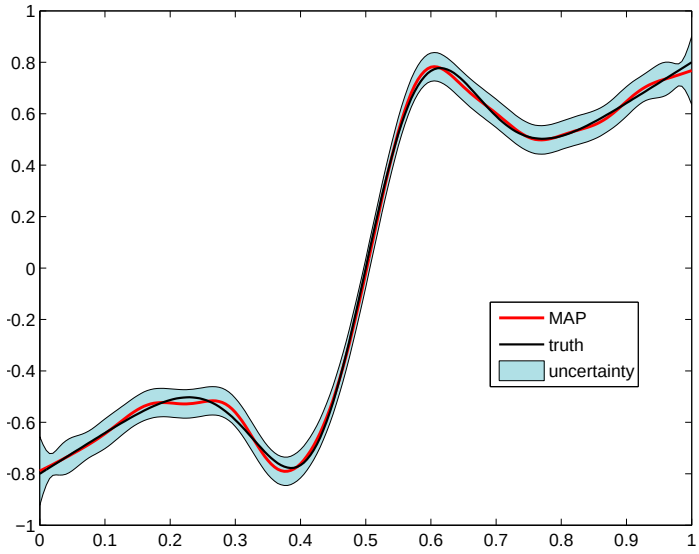


Fig. 7.2 The MAP estimator, the truth function, and the predicted uncertainty (95% credibility region) using π^A_{prior}.

$$f(s) = \begin{cases} 0 & \text{if } s \le 0.7 \\ 10 & \text{otherwise} \end{cases}.$$

Since we otherwise have no further information about $f(s)$, let us be more conservative by choosing $\gamma = 1$ and $\theta = 0.1$ at $j = 70$ in π^O_{prior} as we discussed in (6.8). Figure 7.3 shows that we are doing pretty well in recovering the jump and other parts of the truth function.

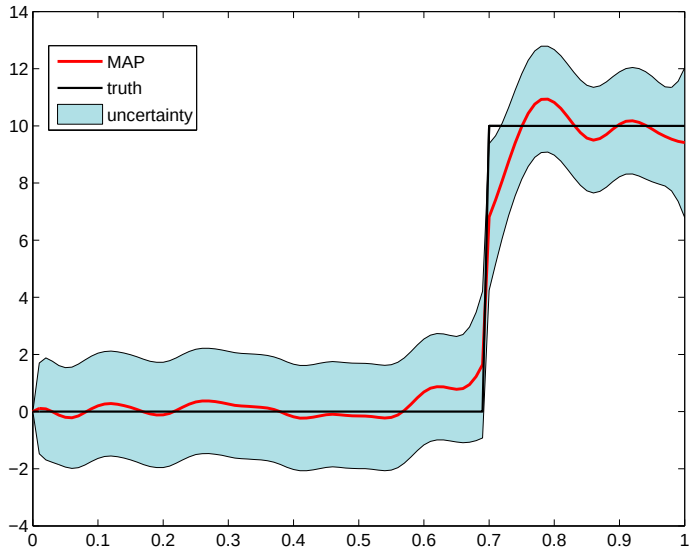


Fig. 7.3 The MAP estimator, the truth function, and the predicted uncertainty (95% credibility region) using π^O_{prior}.

A question you may ask is whether we can do better? The answer is yes by taking smaller γ if the truth function does not vary much everywhere except at the jump discontinuity. We take this prior information into account by taking $\gamma = 1.e-8$, for example, then our reconstruction is almost perfect in Figure 7.4. *Why?*

Exercise 7.7. Try `BayesianPosteriorJump.m` with γ deceasing from 1 to $1.e-8$ to see the improvement in quality of the reconstruction. •

Exercise 7.8. Use your favorite deterministic inversion approach to solve the above deconvolution problem with discontinuity and then compare it with the solution in Figure 7.4. •

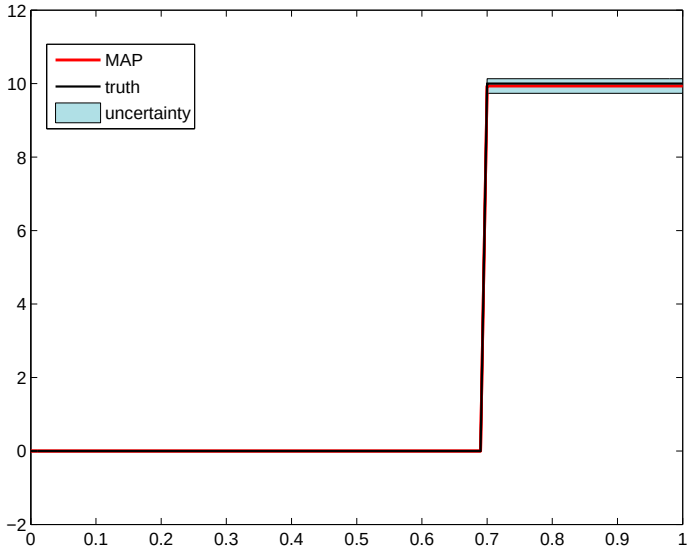


Fig. 7.4 The MAP estimator, the truth function, and the predicted uncertainty (95% credibility region) using π^O_{prior}.

7.2 Connection between Bayesian inverse problems and deterministic inverse problems

We have touched upon the relationship between Bayesian inverse problem and deterministic inverse problem in Section 7.1 by pointing out that the potential of the posterior density is precisely the Tikhonov functional up to a constant. We also point out that the MAP estimator is exactly the solution of the deterministic inverse problem. Note that we derive this relation for a linear likelihood model, but it is in fact true for nonlinear ones (e.g. nonlinear parameter-to-observable map $\mathscr{A}m$).

Can you confirm this?

Up to this point, you may realize that the Bayesian solution contains much more information than its deterministic counterpart. Instead of having a point estimate, the MAP point, we have a complete posterior distribution to explore. In particular, we can talk about a simple uncertainty quantification by examining the diagonal of the posterior covariance matrix. We can even discuss about the posterior correlation structure by looking at the off diagonal elements, though we are not going to do it here in this lecture note. Since, again, both likelihood and prior are Gaussian, the posterior is a Gaussian distribution, and hence the MAP point (the first order moment) and the covariance matrix (the second order moment) are the complete description of the posterior. If, however, the likelihood is not Gaussian, say when the $\mathscr{A}m$ is nonlinear, then one can explore higher moments.

We hope the arguments above convince you that the Bayesian solution provide information far beyond the deterministic counterpart. In the remainder of this sec-

tion, let us dig into details the connection between the MAP point and the deterministic solution, particularly in the context of the deconvolution problem. Recall the definition of the MAP point

$$\begin{aligned} m_{MAP} &\overset{\text{def}}{=} \arg\min_m T(m) = \sigma^2 \left(\frac{1}{2} \left\| y^{obs} - \mathscr{A}m \right\|^2 + \frac{1}{2} \frac{\sigma^2}{\gamma^2} \left\| \Gamma^{-\frac{1}{2}} m \right\|^2 \right) \\ &= \arg\min_m T(m) = \sigma^2 \left(\frac{1}{2} \left\| y^{obs} - y \right\|^2 + \frac{1}{2} \kappa \left\| R^{\frac{1}{2}} m \right\|^2 \right), \end{aligned}$$

where we have defined $\kappa = \sigma^2/\gamma^2$, $R^{\frac{1}{2}} = \Gamma^{-\frac{1}{2}}$, and $y = \mathscr{A}m$.

We begin our discussion with zero Dirichlet boundary condition prior $\pi^D_{\text{prior}}(m)$ in (6.5). Recall in (6.2) and (6.4) that $L_D m$ is proportional to a discretization of the Laplacian operator with zero boundary conditions using second order finite difference method. Therefore, our Tikhonov functional is in fact a discretization, up to a constant, of the following potential in the infinite dimensional setting

$$T_\infty(f) = \frac{1}{2} \left\| y - y^{obs} \right\|^2 + \frac{1}{2} \kappa \|\Delta f\|^2_{L^2(0,1)},$$

where $\|\cdot\|^2_{L^2(0,1)} \overset{\text{def}}{=} \int_0^1 (\cdot)^2 \, ds$. Rewrite the preceding equation informally as

$$T_\infty(f) = \frac{1}{2} \left\| y - y^{obs} \right\|^2 + \frac{1}{2} \kappa \left(f, \Delta^2 f \right)_{L^2(0,1)},$$

and we immediately realize that the potential in our prior description, namely $\|L_D m\|^2$, is in fact a discretization of Tikhonov regularization using the biharmonic operator. This is another explanation for the smoothness of the prior realizations and the name smooth prior, since biharmonic regularization is very smooth. [1]

The power of the statistical approach lies in the construction of prior $\pi^R_{\text{prior}}(m)$. Here, the interpretation of rows corresponding to interior nodes s_j is still the discretization of the biharmonic regularization, but the design of those corresponding to the boundary points is purely statistics, for which we have no corresponding deterministic counterpart (or at least it is not clear how to construct it from a purely deterministic point of view). As the results in Section 7.1 showed, $\pi^R_{\text{prior}}(m)$ provided much more satisfactory results both in the prediction and in uncertainty quantification.

As for the "non-smooth"priors in Section 6.2, a simple inspection shows that $L_N m$ is, up to a constant, a discretization of ∇f. Similar to the above discussion, the potential in our prior description, namely $\|L_D m\|^2$, is now in fact a discretization

[1] From a functional analysis point of view, $\|\Delta f\|^2_{L^2(0,1)}$ is finite if $f \in H^2(0,1)$, and by Sobolev imbedding theorem we know that in fact $f \in C^{1,1/2-\varepsilon}$, the space of continuous differential functions whose first derivative is in the Hölder space of continuous function $C^{1/2-\varepsilon}$, for any $0 < \varepsilon < \frac{1}{2}$. So indeed f is more than continuously differentiable.

of Tikhonov regularization using the Laplacian operator.[2] As a result, the current prior is less smooth than the previous one with harmonic operator. Nevertheless, all the prior realizations corresponding to $\pi_{\text{pren}}(m)$ are at least continuous, though may have steep gradient at s_j as shown in Figures 7.3 and 7.4. The rigorous arguments for the prior smoothness require the Sobolev embedding theorem, but we avoid the details.

For those who have not seen the Sobolev embedding theorem, you only loose the insight on why $\pi^O_{\text{prior}}(m)$ could give very steep gradient realizations (which is the prior belief we start with). Nevertheless, you still can see that $\pi^O_{\text{prior}}(m)$ gives less smooth realizations than $\pi^D_{\text{prior}}(m)$ does, since, at least, the MAP point corresponding to $\pi^O_{\text{prior}}(m)$ only requires finite first derivative of f while second derivative of f needs to be finite at the MAP point if $\pi^D_{\text{prior}}(m)$ is used.

[2] Again, Sobolev embedding theorem shows that $f \in C^{1/2-\varepsilon}$ for $\|\nabla f\|^2_{L^2(0,1)}$ to be finite. Hence, all prior realizations corresponding to $\pi_{\text{pren}}(m)$ are at least continuous. The prior $\pi^O_{\text{prior}}(m)$ is different, due to the scaling matrix J. As long as θ stays away from zero, prior realizations are still in $H^1(0,1)$, and hence continuous though having steep gradient at s_j as shown in Figures 7.3 and 7.4. But as θ approaches zero, prior realizations are leaving $H^1(0,1)$, and therefore may be no longer continuous. Note that in one dimension, $H^{\frac{1}{2}+\varepsilon}$ is enough to be embedded in the space of C^{ε}-Hölder continuous functions. If you like to know a bit about the Sobolev embedding theorem, see [3].

Chapter 8
Independent and identically distributed random draws

Sampling methods discussed in this note are based on two fundamental iid random generators that are available as built-in functions in Matlab. The first one is `rand.m` function which can draw iid random numbers (vectors) from the uniform distribution in $[0,1]$, denoted as $U[0,1]$, and the second one is `randn.m` function that generates iid numbers (vectors) from standard normal distribution $\mathcal{N}(0,I)$, where I is the identity matrix of appropriate size.

The most trivial task is how to draw iid samples $\{m_1, m_2, \ldots, m_N\}$ from a multivariate Gaussian $\mathcal{N}(\overline{m}, \Gamma)$. This can be done through a so-called *whitening* process. The first step is to carry out the following decomposition

$$\Gamma = RR^T,$$

which can be done, for example, using Cholesky factorization. The second step is to define a new random variable as

$$Z = R^{-1}(m - \overline{m}),$$

then Z is a standard multivariate Gaussian, i.e. its density is $\mathcal{N}(0,I)$, for which `randn.m` can be used to generate iid samples

Show that Z is a standard multivariate Gaussian.

$$\{Z_1, Z_2, \ldots, Z_N\} = \texttt{randn(n,N)}.$$

We now generate iid samples m_i via

$$m_i = \overline{m} + RZ_i.$$

Exercise 8.1. Look at `BayesianPriorElicitation.m` to see how we apply the above whitening process to generate multivariate Gaussian prior random realizations. •

You may ask what if the distribution under consideration is not Gaussian, which is true for most practical applications. Well, if the target density $\pi(m)$ is one di-

mensional or multivariate with independent components (in this case, we can draw samples from individual components separately), then we still can draw iid samples from $\pi(m)$, but this time via the standard uniform distribution $U[0,1]$. $U[0,1]$ has 1 as its density function, i.e.,

$$\mu_U(A) = \int_A ds, \quad \forall A \subset [0,1]. \tag{8.1}$$

Now suppose that we would like to draw iid samples from a one dimensional ($S = \mathbb{R}$) distribution with density $\pi(m) > 0$. We still allow $\pi(m)$ to be zero, but only at isolated points on $\mathbb{R}$, and the reason will be clear in a moment. Define the cumulative distribution function (CDF) as

$$\Phi(w) \stackrel{\text{def}}{=} \mathbb{P}[m < w] = \int_{-\infty}^{w} \pi(m)\, dm, \tag{8.2}$$

Why?

then it is clearly that $\Phi(w)$ is non-decreasing and $0 \leq \Phi(w) \leq 1$. Let us define a new random variable Z as

$$Z = \Phi(m). \tag{8.3}$$

Our next step is to show that Z is actually a standard uniform random variable, i.e. $Z \sim U[0,1]$, and then show how to draw m via Z. We begin by the following observation

$$\mathbb{P}[Z < a] = \mathbb{P}[\Phi(m) < a] = \mathbb{P}\left[m < \Phi^{-1}(a)\right] = \int_{-\infty}^{\Phi^{-1}(a)} \pi(m)\, dm, \tag{8.4}$$

where we have used (8.3) in the first equality, the monotonicity of $\Phi(m)$ in the second equality, and the definition of CDF (8.2) in the last equality. Now, we can view (8.3) as the change of variable formula $z = \Phi(m)$, then combining this fact with (8.2) to have

This already shows that the density of z is 1!

$$dz = d\Phi(m) = \Phi'(m)\, dm = \pi(m)\, dm, \text{ and } z = a \text{ when } x = \Phi^{-1}(a).$$

Do you see the second equality?

Consequently, (8.4) becomes

$$\mathbb{P}[Z < a] = \int_0^a dz = \mu_Z(Z < a),$$

which says that the density of Z is 1, and hence Z must be a standard uniform random variable. In terms of our language at the end of Section 9.1, we can define $m = g(Z) = \Phi^{-1}(Z)$, then drawing iid samples for m is simple by first drawing iid samples from Z, then mapping them through g. Let us summarize the idea in Algorithm 1.

The above method works perfectly if one can compute the analytical inverse of the CDF easily and efficiently; it is particularly efficient for discrete random variables, as we shall show. Of course we can always compute the inverse CDF numerically. However, note that the CDF is an integral operation, and hence its inverse is

Algorithm 1 CDF-based sampling algorithm

1. Draw $z \sim U[0,1]$,
2. Compute the inverse of the CDF to draw m, i.e. $m = \Phi^{-1}(z)$. Go back to Step 1.

some kind of differentiation. As shown before, numerical differentiation could be an ill-posed problem and we do not want to add extra ill-posedness on top of the original ill-posed inverse problem that we started with. Instead, let us introduce a simpler but more robust algorithm that works for multivariate distribution without requiring the independence of individual components. We shall first discuss the algorithm and then analyze it to show why it works.

Suppose that we want to draw iid samples from a target density $\pi(m)$, but we only know it up to a constant $C > 0$, i.e., $C\pi(m)$. (This is perfect for our Bayesian inversion framework since we typically know the posterior up to a constant as in (5.5).) Assume that we have a *proposal distribution* $q(m)$ at hand, for which we know how to sample easily and efficiently. This is not a limitation since we can always take either the standard normal distribution or uniform distribution as the proposal distribution. We further assume that there exists $D > 0$ such that

$$C\pi(m) \leq Dq(m), \tag{8.5}$$

then we can draw a sample from $\pi(m)$ by the rejection-acceptance sampling Algorithm 2.

Algorithm 2 Rejection-Acceptance sampling algorithm

1. Draw m from the proposal $q(m)$,
2. Compute the *acceptance probability*

$$\alpha = \frac{C\pi(m)}{Dq(m)},$$

3. Accept m with probability α or reject it with probability $1-\alpha$. Go back to Step 1.

In practice, we carry out Step 3 of Algorithm 2 by flipping an "α-coin". In particular, we draw u from $U[0,1]$, then accept m if $\alpha > u$. It may seem to be magic why Algorithm 2 provides random samples from $\pi(m)$. Let us confirm this using the Bayes formula (4.6).

Proposition 8.1. *Accepted m in Algorithm 2 is distributed by the target density* π.

Make sure you understand this proof since we will reuse most of it for the Metropolis-Hastings algorithm!

Proof. Denote B as the event of accepting a draw q (or the acceptance event). Algorithm 2 tells us that the probability of B given m, which is precisely the acceptance probability, is

$$\mathbb{P}[B|m] = \alpha = \frac{C\pi(m)}{Dq(m)}. \tag{8.6}$$

On the other hand, the prior probability of m in the incremental event $dA = \left[m', m' + dm\right]$ in Step 1 is $q(m)\,dm$. Applying the Bayes formula for conditional probability (3.4) yields the distribution of a draw m provided that it has been already accepted

$$\mathbb{P}[m \in dA|B] = \frac{\mathbb{P}[B|m]\,q(m)\,dm}{\mathbb{P}[B]} = \pi(m)\,dm,$$

where we have used (8.6) and $\mathbb{P}[B]$, the probability of accepting a draw from q, is the following marginal probability

$$\mathbb{P}[B] = \int_S \mathbb{P}[B|m]\,q(m)\,dm = \frac{C}{D}\int_S \pi(m)\,dm = \frac{C}{D}.$$

Note that

$$\mathbb{P}[B, m \in dm] = \pi(B,m)\,dm = \mathbb{P}[B|m]\,\pi_{\text{prior}}(m) = \mathbb{P}[B|m]\,q(m)\,dm,$$

an application of (3.3), is the probability of the joint event of drawing an m from $q(m)$ and accept it. The probability of B, the acceptance event, is the total of accepting probability, which is exactly the marginal probability. As a result, we have

$$\mathbb{P}[m \in A|B] = \int_A \pi(m)\,dm,$$

which, by definition (4.2), says that the accepted m in Step 3 of Algorithm 2 is distributed by $\pi(m)$, and this is the desired result. □

Algorithm 2 is typically slow in practice in the sense that a large portion of samples is rejected, particularly for high dimensional problem, though it provides iid samples from the true underlying density. Another problem with this algorithm is the computation of D. Clearly, we can take very large D and the condition (8.5) would be satisfied. However, the larger D is the smaller the acceptance probability α, making Algorithm 2 inefficient since most of draws from $q(m)$ will be rejected. As a result, we need to minimize D and this could be nontrivial depending the complexity of the target density.

Exercise 8.2. You are given the following target density

$$\pi(m) = \frac{g(m)}{C}\exp\left(-\frac{m^2}{2}\right),$$

where C is some constant independent of m, and

$$g(m) = \begin{cases} 1 \text{ if } x > a \\ 0 \text{ otherwise} \end{cases}, \quad a \in \mathbb{R}.$$

Take the proposal density as $q(m) = \mathcal{N}(0,1)$.

1. Find the smallest D that satisfies condition (8.5).

2. Implement the rejection-acceptance sampling Algorithm 2 in Matlab and draw 10000 samples, by taking $a = 1$. Use Matlab `hist.m` to plot the histogram. Does its shape resemble the exact density shape?
3. Increase a as much as you can, is there any problem with Algorithm 2? Can you explain why?

•

Exercise 8.3. You are given the following target density

$$\pi(m) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sqrt{m_1^2+m_2^2}-1\right)^2-\frac{1}{2\delta^2}(m_2-1)^2\right),$$

where $\sigma = 0.1$ and $\delta = 1$. Take the proposal density as $q(m) = \mathcal{N}(0, I_2)$, where I_2 is the 2×2 identity matrix.

1. Find a reasonable D, using any means you like, that satisfies condition (8.5).
2. Implement the rejection-acceptance sampling Algorithm 2 in Matlab and draw 10000 samples. Plot a contour plot for the target density, and you should see the horse-shoe shape, then plot all the samples as dots on top of the contour. Do most of the samples sit on the horse-shoe?

•

Chapter 9
Classical Limit Theorems

In the last section, we have shown that if the paramter-to-observable map is linear, i.e. $h(m) = \mathscr{A}m$, and both the noise and the prior models are Gaussian, then the MAP point and the posterior covariance matrix are exactly the solution and the inverse of the Hessian of the Tikhonov functional, respectively. Moreover, since the posterior is Gaussian, the MAP point is identically the mean, and hence the posterior distribution is completely characterized. In practice, $h(m)$ is typically nonlinear. Consequently, the posterior distribution is no longer Gaussian. Nevertheless, the MAP point is still the solution of the Tikhonov functional, though the mean and the convariance matrix are to be determined. The question is how to estimate the mean and the covariance matrix of a non-Gaussian density.

We begin by recalling the definition the mean

$$\overline{m} = \mathbb{E}[m],$$

and a natural idea is to approximate the integral by some numerical integration. For example, suppose $S = [0,1]$ and then we can divide S into N intervals, each of which has length of $1/N$. Using a rectangle rule gives

$$\overline{m} \approx \frac{(m_1 + \ldots + m_N)}{N}. \tag{9.1}$$

But this kind of method cannot be extended to $S = \mathbb{R}^n$. This is where the central limit theorem and law of large numbers come to rescue. They say that the simple formula (9.1) is still valid with a simple error estimation expression.

9.1 Some classical limit theorems

Theorem 9.1 (The central limit theorem (CLT)). *Assume that real valued random variables $m_1, \ldots$ are independent and identically distributed (iid), each with expectation $\overline{m}$ and variance σ^2. Then*

$$Z_N = \frac{1}{\sigma\sqrt{N}}(m_1 + m_2 + \cdots + m_N) - \frac{\overline{m}}{\sigma}\sqrt{N}$$

converges, in distribution[1]*, to a standard normal random variable. In particular,*

$$\lim_{N\to\infty} \mathbb{P}[Z_N \le m] = \frac{1}{2\pi}\int_{-\infty}^{m} \exp\left(-\frac{t^2}{2}\right) dt \tag{9.2}$$

Proof. The proof is elementary, though technical, using the concept of characteristic function (Fourier transform of a random variable). A complete proof can be consulted from [16].

Theorem 9.2 (Strong law of large numbers (LLN)). *Assume random variables $m_1, \ldots$ are independent and identically distributed (iid), each with finite expectation $\overline{m}$ and finite variance σ^2. Then*

$$\lim_{N\to\infty} S_N \overset{def}{=} \frac{1}{N}(m_1 + m_2 + \cdots + m_N) = \overline{m} \tag{9.3}$$

almost surely[2]*.*

Proof. A beautiful, though not classical, proof of this theorem is based on backward martingale, tail σ-algebra, and uniform integrability. Let's accept it in this note and see [16] for a complete proof.

Remark 9.1. The central limit theorem says that no matter what the underlying common distribution looks like, the sum of iid random variables, when properly scaled and centralized, converges in distribution to a standard normal distribution. The strong law of large numbers, on the other hand, states that the average of the sum is, as expected in the limit, precisely the mean of the common distribution with probability one.

Exercise 9.1 (Numerical verification of CLT and LLN). In Matlab, draw N iid samples from the standard uniform distribution, for which we know that the mean is 0.5.

1. Plot the histogram of Z_N as a function of N and observe the convergence of the histogram to the standard normal distribution. Estimate the mean and variance of Z_N as a function of N and see whether they converege to 0 and 1.
2. Plot S_N as a function of N and see whether it converges to 0.5.

•

[1] Convergence in distribution is also known as weak convergence and it is beyond the scope of this introductory note. You can think of the distribution of Z_n is more and more like the standard normal distribution as $n \to \infty$, and it is precisely (9.2).

[2] Almost sure convergence is the same as convergence with probability one, that is, the event on which the convergence (9.3) does not happen has zero probability.

Both the central limit theorem (CLT) and the strong law of large numbers (LLN) are useful, particularly LLN, and we use them routinely. For example, if we are given an iid sample $\{m_1, m_2, \cdots, m_N\}$ from a common distribution $\pi(m)$, the first thing we should do is perhaps to compute the the sample mean S_N to estimate the actual mean $\overline{m}$. From LLN we know that the sample mean can be as close as desired if N is sufficiently large. A question immediately arises is whether we can estimate the error between the sample mean and the truth mean, given a finite N. Let us first give an answer based on a simple application of the CLT. Since the sample $\{m_1, m_2, \cdots, m_N\}$ satisfies the condition of the CLT, we know that Z_N converges to $\mathcal{N}(0,1)$. It follows that, at least for sufficiently large N, the mean squared error between z_N and 0 can be estimated as

$$1 \approx \mathbb{V}ar[Z_N] \overset{\text{def}}{=} \left\| Z_N - \overline{Z_N} \right\|^2_{L^2(S,\mathbb{P})} \approx \|Z_N - 0\|^2_{L^2(S,\mathbb{P})} \overset{\text{def}}{=} \mathbb{E}\left[(Z_N - 0)^2 \right],$$

which, after some simple algebra manipulations, can be rewritten as

$$\|S_N - \overline{m}\|^2_{L^2(S,\mathbb{P})} \approx \frac{\sigma^2}{N} \tag{9.4}$$

Exercise 9.2. Show that (9.4) holds. •

The result (9.4) shows that the error of the sample mean S_N in the $L^2(S,\mathbb{P})$-norm goes to zero like $1/\sqrt{N}$. One should be aware of the popular statement that the error goes to zero like $1/\sqrt{N}$ independent of dimension is not entirely correct because the variance σ^2, and hence the standard deviation σ, of the underlying distribution $\pi(m)$ may depend on the dimension n.

If we are a little bit delicate, we may not feel completely comfortable with the error estimate (9.4) since we can rewrite it as

$$\|S_N - \overline{m}\|_{L^2(S,\mathbb{P})} = C\frac{\sigma}{\sqrt{N}},$$

and we are not sure how big C is and the dependence of C on N. Let us attempt to determine C. We have

$$\begin{aligned} \|S_N - \overline{m}\|^2_{L^2(S,\mathbb{P})} &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N (m_i - \overline{m}) \right) \left(\sum_{j=1}^N (m_j - \overline{m}) \right) \right] \\ &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N (m_i - \overline{m})^2 \right) \right] = \frac{1}{N^2}\sum_{i=1}^N \sigma^2 = \frac{\sigma^2}{N}, \end{aligned}$$

where we have used $\overline{m} = \dfrac{1}{N}\displaystyle\sum_{i=1}^N \overline{m}$ in the first equality, $\mathbb{E}\left[(m_i - \overline{m})(m_j - \overline{m})\right] = 0$ if $i \neq j$ in the second equality since m_i, $i = 1, \ldots, N$ are iid random variables, and the definition of variance in the third equality. So in fact $C = 1$.

In practice, we rarely work with m directly but indirectly via some mapping $g : S \to T$. We have that $g(m_i)$, $i = 1,\ldots,N$ are iid[3] if m_i, $i = 1,\ldots,N$ are iid.

Exercise 9.3. Suppose the density of m is $\pi(m)$ and $z = g(m)$ is differentially invertible, i.e. $m = g^{-1}(z)$ exists and differentiable, what is the density of $g(m)$? •

Perhaps, one of the most popular and practical problems is to evaluate the mean of g, i.e.,

$$I \overset{\text{def}}{=} \mathbb{E}[G(m)] = \int_S g(m)\,\pi(m)\,dm, \tag{9.5}$$

which is an integral in $\mathbb{R}^n$.

Exercise 9.4. Define $z = g(m) \in T$, the definition of the mean in (4.4) gives

$$\mathbb{E}[G(m)] \equiv \mathbb{E}[Z] \overset{\text{def}}{=} \int_T z\pi_Z(z)\,dz.$$

Derive formula (9.5). •

Again, we emphasize that using any numerical integration methods that you know of for integral (9.5) is either infeasible or prohibitedly expensive when the dimension n is large, and hence not scalable. The LLN provides a reasonable answer if we can draw iid samples $\{g(m_1),\ldots,g(m_N)\}$ since we know that

$$\lim_{N\to\infty} \underbrace{\frac{1}{N}\left(g(m_1)+\ldots+g(m_N)\right)}_{I_N} = I$$

Do you trivially see this?

with probability 1. Moreover, as shown above, the mean squared error is given by

$$\|I_N - I\|^2_{L^2(T,\mathbb{P})} = \mathbb{E}\left[(I_N - I)^2\right] = \frac{\mathbb{V}ar[G(m)]}{N}.$$

Again, the error decreases to zero like $1/\sqrt{N}$ "independent" of the dimension of T, but we need to be careful with such a statement unless $\mathbb{V}ar[G(m)]$ DOES NOT depend on the dimension.

A particular function g of interest is the following

$$g(m) = (m - \overline{m})(m - \overline{m})^T,$$

whose expectation is precisely the covariance matrix

$$\Gamma = cov(m) = \mathbb{E}\left[(m - \overline{m})(m - \overline{m})^T\right] = \mathbb{E}[G].$$

The average I_N in this case is known as the sample (aka empirical) covariance matrix. Denote

[3] We avoid technicalities here, but g needs to be a Borel function for the statement to be true.

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \overline{m})(m_i - \overline{m})^T$$

as the sample covariance matrix. Clearly, $\hat{\Gamma}$ converges almost surely to Γ by LLN. Note that $\overline{m}$ is typically not available in practice, and we have to resort to a computable approximation

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T, \tag{9.6}$$

with

$$\hat{m} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} m_i$$

denoting the sample mean. One can show that $\hat{\Gamma}$ converges to Γ but it needs more techinical argument on convergence in probability[4], and hence is omitted. The easier question to answer is whether $\hat{\Gamma}$ is an *unbiased estimator* of Γ, and the detail is in Exercise 9.5.

Exercise 9.5. It turns out that (9.6) is a biased estimator for Γ. We define $\tilde{m}$ an unbiased estimator for m if $\mathbb{E}[\tilde{m}] = m$. For example, $\hat{m}$ is an unbiased estimator for $\overline{m}$. Indeed

$$\mathbb{E}[\hat{m}] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[m_i] = \overline{m},$$

since, again, m_i are iid with mean $\overline{m}$.

1. Show that
$$\mathbb{E}[(\hat{m} - \overline{m})(\hat{m} - \overline{m})] = \frac{\Gamma}{N}.$$
2. Start from
$$J = \sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T = \sum_{i=1}^{N}(m_i - \overline{m} + \overline{m} - \hat{m})(m_i - \overline{m} + \overline{m} - \hat{m})^T$$
to show that
$$\mathbb{E}[J] = (N-1)\Gamma,$$
and then conclude that
$$\frac{J}{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T$$
is an unbiased estimator for Γ. So the "correct scaling" in $\hat{\Gamma}$ should be $N-1$ instead of N. For sufficient large N, the difference is however negligible.

•

[4] The key is the Slutsky's theorem on convergence in probability of sequence of random variables.

9.2 Appendix

Definition 9.1 (Convergence in distribution). A sequence of random variables m_N, $N \in \mathbb{N}$, converges in distribution to m if μ_{m_N} converges weakly to μ_m, i.e.,

$$\int f \, d\mu_{m_N} \overset{N\to\infty}{\to} \int f \, d\mu_m$$

for any bounded and continuous function f.

Definition 9.2 (Almost surely convergence). A sequence of random variables m_N, $N \in \mathbb{N}$, converges almost surely to m if

$$\mathbb{P}[\{m_N - m \neq 0\}] = 0,$$

or equivalently

$$\mathbb{P}[\{m_N - m = 0\}] = 1.$$

Chapter 10
Markov chain Monte Carlo I

Chapter 8 presents methods to draw i.i.d. samples from an arbitrary distribution and Chapter 9 shows that i.i.d. samples can be used to estimate the moments (the mean and the covariance, in particular). The most robust i.i.d sampling method that works in any dimension is the rejection-acceptance sampling algorithm though it may be slow in practice. In this chapter, we introduce the Markov chain Monte Carlo (MCMC) method which is the most popular sampling approach. It is in general more effective than any methods discussed so far, particularly for complex target density in high dimensions, though it has its own problems. One of them is that we no longer have i.i.d. samples but correlated ones. Next, let us introduce some notations to study MCMC methods.

Definition 10.1. A collection $\{m_0, m_1, \ldots, m_N, \ldots\}$ is called *Markov chain* if the distribution of m_k depends only on the immediate previous state m_{k-1}.

Definition 10.2. We call the probability of m_k in A starting from m_{k-1} as the *transition probability* and denote it as $P(m_{k-1}, A)$. With an abuse of notation, we introduce the *transition kernel* $P(m_{k-1}, m)$ such that

What does $P(m_{k-1}, dm)$ mean?

$$P(m_{k-1}, A) \stackrel{\text{def}}{=} \int_A P(m_{k-1}, m)\, dm = \int_A P(m_{k-1}, dm)\,.$$

Clearly $P(m, S) = \int_S P(m, p)\, dp = 1$.

Example 10.1. Assume that we have a set of Internet websites that may be linked to the others. We represent these sites as nodes and mutual linkings by directed arrows connecting nodes such as in Figure 10.1. We assign the network of nodes a *transition matrix* P as

$$P = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}. \tag{10.1}$$

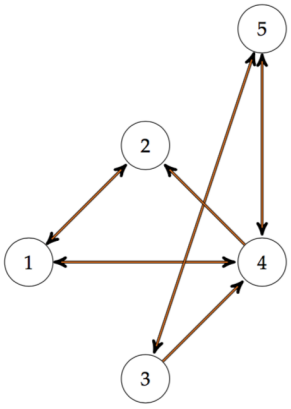


Fig. 10.1 Five internet websites and their connections.

The jth row of P is the probability of moving from the jth node to the rest. For example, the first row says that if we start from node 1, we can move to either node 2 or node 4, each with probability $\frac{1}{2}$. Note that we have treated all the nodes equally, that is, the transition probability from one node to other linked nodes is the same (a node is not linked to itself in this model).

Let $m_{k-1} = 4$, then the probability kernel $P(m_{k-1} = 4, m)$ is exactly the fourth row of P, i.e.,

$$P(m_{k-1} = 4, m) = [1/3, 1/3, 0, 0, 1/3].$$

The transition probability $P(m_{k-1} = 4, m_k = 1)$ is thus given by

$$\begin{aligned} P(m_{k-1} = 4, m_k = 1) &= \int_S P(m_{k-1} = 4, m)\, \delta(m-1)\, dm \\ &= \sum_{k=1}^{5} P(4,k)\, \delta(k-1) = P(4,1) = 1/3. \end{aligned}$$

△

Definition 10.3. We call $\mu(dm) = \pi(m)\, dm$ the invariant distribution and $\pi(m)$ invariant density of the transition probability $P(m_{k-1}, dm)$ if

$$\mu(dm) = \pi(m)\, dm = \int_S P(p, dm)\, \pi(p)\, dp. \tag{10.2}$$

Example 10.2. The discrete version of (10.2), applying to our website Example 10.1 and denoting the invariant measure as π_∞, reads

$$\pi_\infty(j) = \sum_{k=1}^{5} P(k,j)\, \pi_\infty(k) = \pi_\infty P(:, j), \quad \forall j = 1, \ldots, 5,$$

which shows that the invariant measure π_∞ is the left eignvector of the transition matrix P corresponding to the unity eigenvalue. $\triangle$

Exercise 10.1. Assume we are initially at node 4, and we represent the initial probability density as

$$\pi_0 = \big[0,0,0,1,0\big],$$

that is, we are initially at node 4 with certainty. In order to know the next node to visit, we first compute the probability density of the next state by

$$\pi_1 = \pi_0 P,$$

then randomly move to a node by drawing a sample from the (discrete) probability density π_1. In general, the probability density after k steps is given by

$$\pi_k = \pi_{k-1} P = \ldots = \pi_0 P^k. \tag{10.3}$$

Observing (10.3) you may wonder what happens if k approaches infinity. Assume, on credit, the limit probability density π_∞ exists, then it ought to satisfy

$$\pi_\infty = \pi_\infty P. \tag{10.4}$$

It follows that π_∞ is, if exists, nothing more than the invariant measure!

Figure 10.2 shows the visiting frequency (blue) for each node after $N = 1500$ moves. Here, visiting frequency of a node is the number of visits to that node divided by N. We expect that numerical visiting frequencies approximate the visiting probabilities in the limit. We confirm this expectation by also plotting the components of π_∞ (red) in Figure 10.2. By the way, π_{1500} is equal to π_∞ up to machine zero, meaning that a draw from π_N, $N \geq 1500$, is distributed by the limit distribution π_∞.

Suppose you are seeking sites that contains a keyword of interest for which all the nodes, and hence websites, contain. A good search engine will show you all these websites. The question is now which website should be ranked first, second, and so on? You may guess that node 4 should be the first one in the list. However, Figure 10.2 shows that node 1 is the most visited one, and hence should appear at the top of the website list coming from the search engine!

Use the CDF-based sampling algorithm, namely Algorithm 1, to reproduce Figure 10.2. Compare the probability density π_{1500} with the limit density, are they the same? Generate 5 figures corresponding to starting nodes $1,\ldots,5$, what do you observe? •

Exercise 10.2. Using the above probabilistic method to determine the probability that the economy, as shown in Figure 10.3, is in recession. •

Definition 10.4. A Markov chain $\mathscr{M} = \{m_0, m_1, \ldots, m_N, \ldots\}$ is *reversible* with respect to $\pi(\cdot)$ if

$$\pi(m)P(m,p) = \pi(p)P(p,m), \quad \forall m,p \in \mathscr{M}. \tag{10.5}$$

What do you conclude if you integrate both side of (10.5) *with respect to p (or m)?*

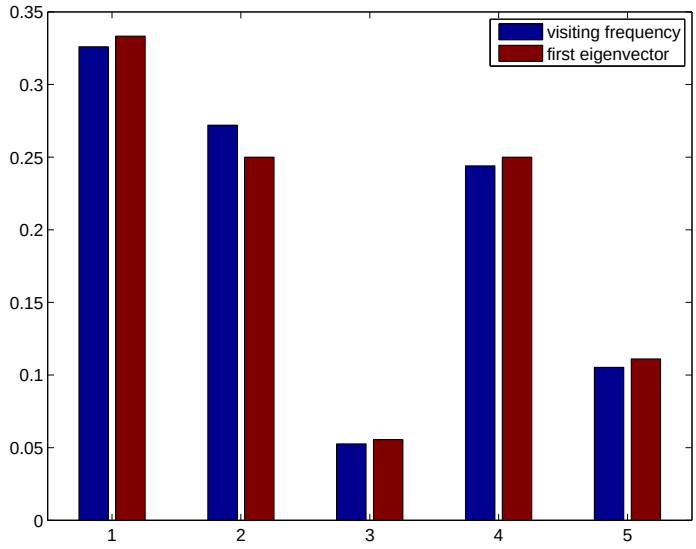


Fig. 10.2 Visiting frequency after $N = 1500$ moves and the first eigenvector.

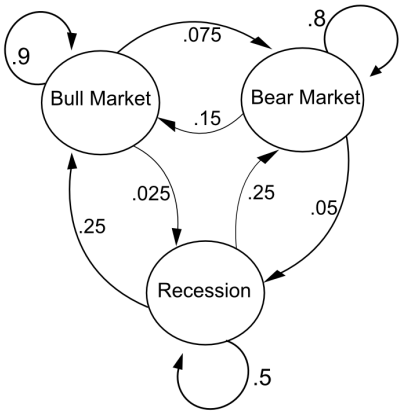


Fig. 10.3 An example of economy states (source *Wikipedia*).

The reversibility relation (10.5) is also known as *detailed balanced equation*. You can think of the reversibility saying that the likelihood of moving from m to p is equal to the likelihood of moving from p to m.

Exercise 10.3. What is the discrete version of (10.5)? Does the transition matrix in the website ranking problem satisfy the reversibility? If not, why? How about the transition matrix in Exercise 10.2? •

Solution:

$$\pi(i)P(i,j) = \pi(j)P(j,i).$$

The reversibility of a Markov chain is useful since we can immediately conclude that $\pi(m)$ is an invariant density.

Proposition 10.1. *If the Markov chain $\mathscr{M} = \{m_0, m_1, \ldots, m_N, \ldots\}$ is reversible with respect to $\pi(m)$, then $\pi(m)$ is an invariant density.*

Proof. We need to prove (10.2), but it is straightforward since

$$\int_S \pi(p)\,P(p,dm)\,dp \overset{\text{reversibility}}{=} \pi(m)\,dm \int_S P(m,p)\,dp = \pi(m)\,dm.$$

□

The above discussion seems to indicate that if a Markov chain is reversible then eventually the states in the chain are distributed by the underlying invariant distribution. *It is important to point out that this is only a sufficient for the Markov chain to converges to a desired stationary distribution*. Indeed, one can construct non-reversible chains that converge (sometimes faster than the reversible counterparts) a distribution, but this is beyond the scope of this book. A question you may ask is how to construct a transition kernel such that reversibility holds. This is exactly the question Markov chain Monte Carlo methods are designed to answer. Let us now present the Metropolis-Hastings MCMC method in Algorithm 3.

Algorithm 3 Metropolis-Hastings MCMC Algorithm

Choose initial m_0
for $k = 0, \ldots, N$ **do**

1. Draw a sample p from the proposal density $q(m_k, p)$
2. Compute $\pi(p)$, $q(m_k, p)$, and $q(p, m_k)$
3. Compute the acceptance probability

$$\alpha(m_k, p) = \min\left\{1, \frac{\pi(p)q(p,m_k)}{\pi(m_k)q(m_k,p)}\right\}$$

4. **Accept** and set $m_{k+1} = p$ with probability $\alpha(m_k, p)$. Otherwise, **reject** and set $m_{k+1} = m_k$

end for

The idea behind the Metropolis-Hastings Algorithm 3 is very similar to that of rejection-acceptance sampling algorithm. That is, we first draw a sample from an "easy" distribution $q(m_k, p)$, then make correction so that it is distributed more like the target density $\pi(p)$. However, there are two main differences. First, the proposal

distribution $q(m_k, p)$ is a function of the last state m_k. Second, the acceptance probability involves both the last state m_k and the proposal move p. As a result, a chain generated from Algorithm 3 is in fact a Markov chain.

What remains is to show that the transition kernel of Algorithm 3 indeed satisfies the reversibility condition (10.5). This is the focus of the next proposition.

Proposition 10.2. *Markov chains generated by Algorithm 3 are reversible.*

Proof. We proceed in two steps. In the first step, we consider the case in which the proposal p is accepted. Denote B as the event of accepting a draw q (or the acceptance event). Following the same proof of Proposition 8.1, we have

$$\mathbb{P}[B|p] = \alpha(m_k, p),$$

Make sure you see this!

leading to

$$\pi(B, p) = \mathbb{P}[B|p]\, \pi_{\text{prior}}(p) = \alpha(m_k, p) q(m_k, p),$$

which is exactly $P(m_k, p)$, the probability density of the joint event of drawing p from $q(m_k, p)$ and accept it, starting from m_k. It follows that the reversibility holds since

$$\begin{aligned}
\pi(m_k) P(m_k, p) &= \pi(m_k)\, q(m_k, p) \min\left\{1, \frac{\pi(p) q(p, m_k)}{\pi(m_k) q(m_k, p)}\right\} \\
&= \min\{\pi(m_k) q(m_k, p), \pi(p) q(p, m_k)\} \\
&= \min\left\{\frac{\pi(m_k) q(m_k, p)}{\pi(p) q(p, m_k)}, 1\right\} \pi(p) q(p, m_k) \\
&= \pi(p) P(p, m_k).
\end{aligned}$$

In the second step, we remain at m_k, i.e., $m_{k+1} = m_k$, then the reversibility is trivially satisfied no matter what the transition kernel $P(m_k, p)$ is. This is the end of the proof. □

Explain why

Exercise 10.4. What is the probability of staying put at m_k?

Chapter 11
Markov chain Monte Carlo II

The next question question we need to discuss is whether the Markov chain indeed converges to the desired stationary distribution and how fast the convergence is. The detailed answers to these questions are technical and beyond the scope of this book. Nevertheless we attempt to give an intuitive answers here.

Irreducibility. It turns out that even a Markov chain has $\pi(m)$ as its stationary distribution, it may fail to converge to stationarity. That is, the stationary distribution may not be unique or the chain may never get from one state to another. To avoid this issue, the chain is required to be irreducible[1] [21, 24]. In other words, any state has positive probability of eventually being reached from an arbitrary state. For our setting in this chapter where $\mathbb{X} = \mathbb{R}^n$, we need to mild condidtions for the Markov chain from Algorithm 3 to be irreducible: 1) the desired density $\pi(m)$ is finite everywhere; and 2) the proposal $q(\cdot,\cdot)$ is positive and continuous.

Aperiodicity. The fact that even irreducible chain may not converge in distribution is due to periodicity problem, i.e., the chain may oscillate between states and hence is not convergent. Fortunately, for our setting the chain is automatically aperiodic [24].

Theorem 11.1. *If the Markov chain has $\pi(m)$ as its stationary distribution, and it is both irreducible and aperiodic, then for* almost everywhere *$m \in \mathbb{X}$[2], then the Markov chain is eventually distributed as $\pi(m)$, i.e.,*

$$\lim_{N\to\infty} P^N(m,A) = \mu(A), \quad \forall A \in \mathscr{S},$$

where $\mu(A)$ is the distribution function (the law), and $P^N(m,A) \overset{def}{=} \mathbb{P}[m_N \in A | m_0 = m]$ denotes N-step transition law of the Markov chain. Furthermore, the LLN holds:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} g(m_i) = \int_{\mathbb{X}} g(m)\, d\pi(m),$$

[1] In fact, only ϕ-irreducibility, a weaker condition, is needed.

[2] The correct requirement for $\mathbb{X}$ is that it has countably generated σ-algebra [24].

for all g such that $\int_{\mathbb{X}} |g(m)|\, d\pi(m) < \infty$.

Theorem 11.1 however has a little caveat, that is, it is true for almost every $m \in \mathbb{X}$ except for the zero-probability exceptional set. For example, if we are unlucky to start the chain in this exceptional set, the chain does not converge! To overcome this issue, the chain needs to be *Harris recurrent*, which means that $\forall A : \mu(A) > 0$, and for all m, the chain eventually reaches A from m with probability 1. For our setting, if the proposal $q(m,\cdot)$ is absolutely continuous with respect to $\pi(\cdot)$, i.e.,

$$q(m,A) = \int_A d\pi(x), \quad \forall A \in \mathscr{S},$$

then the Markov chain is *Harris recurrent*.

Up to this point, we know that under mild conditions, the Markov chain from the above Metropolis-Hastings algorithm converges, but we haven't discussed the convergence rate. The reason is that it is much more technical involving the notion of *ergodicity*. The readers are referred to [21, 24] for the details.

As you can see the Metropolis-Hastings algorithm is simple and elegant, but provides us a reversible transition kernel, which is exactly what we are looking for. The keys behind this are Steps 3 and 4 in Algorithm 3, known as Metropolized steps. At this point, we should be able to implement the algorithm except for one small detail: what should we choose for the proposal density? Let us choose the following Gaussian kernel

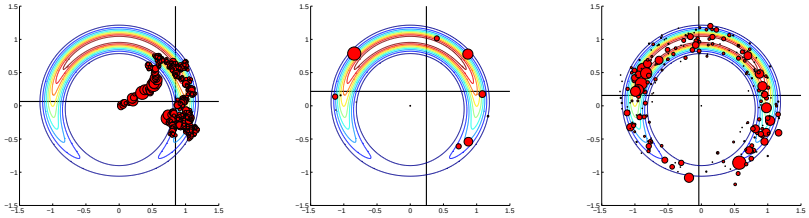
$$q(m,p) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{1}{2\gamma^2}\|m-p\|^2\right),$$

which is the most popular choice. Metropolis-Hastings algorithm with above isotropic Gaussian proposal is known as *Random Walk Metropolis-Hastings* (RWMH) algorithm. For this particular method, the acceptance probability is very simple, i.e.,

$$\alpha = \min\left\{1, \frac{\pi(p)}{\pi(m)}\right\}.$$

We are now in the position to implement the method. For concreteness, we apply the RWMH algorithm on the horse-shoe shape in Exercise 8.3. We take the origin as the starting point m_0. Let us first be conservative by choosing a small proposal variance $\gamma^2 = 0.02^2$ so that the proposal p is very close to the current state m_k. In order to see how the MCMC chain evolves, we plot each state m_k as a circle (red) centered at m_k with radius proportional to the number of staying-puts. Figure 11.1(a) shows the results for $N = 1000$. We observe that the chain takes about 200 MCMC simulations to enter the high probability density region. This is known as *burn-in* time in MCMC literature, which tells us how long a MCMC chain takes to start exploring the density. In other words, after the burn-in time, a MCMC begins to distribute like the target density. As can be seen, the chain corresponding to small proposal variance γ^2 explores the target density very slowly. If we approximate the

average acceptance rate by taking the ratio of the number of accepted proposal over N, it is 0.905 for this case. That is, almost all the proposals p are accepted, but exploring a very small region of high probability density.



(a) $\gamma = 0.02$ (b) $\gamma = 5$ (c) $\gamma = 0.5$

Fig. 11.1 RWMH with different proposal variance γ^2.

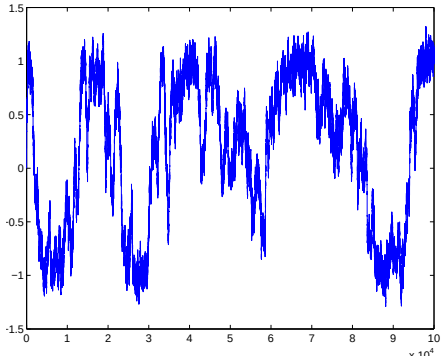
Let us now increase the proposal stepsize γ to 5, and we show the corresponding chain in Figure 11.1(b). This time, the chain immediately explores the target density without any burn-in time. However, it does so in an extremely slow manner. Most of the time the proposal p is rejected, resulting in a few big circles in Figure 11.1(b). The average acceptance rate in this case is 0.014, which shows that most of the time we reject proposals p.

The results in Figures 11.1(a) and 11.1(b) are two extreme cases, both of which explore the target density in a very lazy manner since the chain either accepts all the small moves with very high acceptance rate or rejects big moves with very low acceptance rate. This suggests that there must be an *optimal* acceptance rate for the RWMH algorithm. Indeed, one can show that the optimal acceptance rate is 0.234 [23]. For our horse-shoe target, it turns out that the corresponding optimal stepsize is approximately $\gamma = 0.5$. To confirm this, we generate a new chain with this stepsize, again with $N = 1000$, and show the result in Figure 11.1(c). As can be seen, the samples spread out nicely over the horse-shoe.

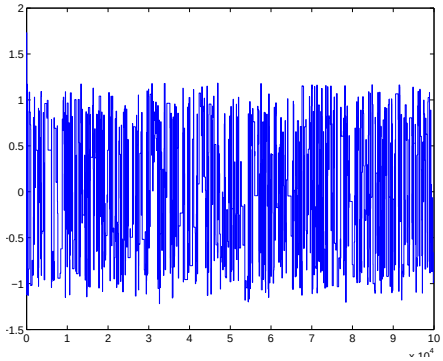
We have judged the quality and convergence of a MCMC chain by looking at the scatter plot of the samples. Another simple approach is to look at the trace plot of components of m. For example, we show the trace plot of the first component in Figure 11.2 for the above three stepsizes. The rule of thumb is that a Markov chain is considered to be good if its trace plot is close to a white noise one, a "fuzzy worm", in which all the samples are completely uncorrelated. Based on this criteria, we again conclude that $\gamma = 0.5$ is the best compared to the other two extreme cases.

Plot a trace plot for a one dimensional Gaussian white noise to see how it looks like!

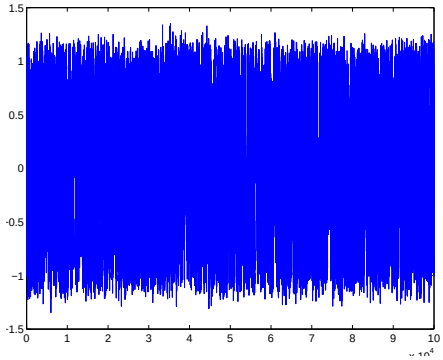
Nevertheless, the above two simple criteria are neither rigorous nor possible in high dimensions. This observation immediately reminds us the strong law of large number in computing the mean and its dimension-independent error analysis using the central limit theorem. Since the target is symmetric about the vertical axis, the first component of the mean must be zero. Let us use the strong law of large number



(a) $\gamma = 0.02$



(b) $\gamma = 5$



(c) $\gamma = 0.5$

Fig. 11.2 Trace plots of the first component of m with different γ^2.

to estimate the means for the above three stepsizes and show them as cross signs in Figures 11.1(a), 11.1(b), and 11.1(c). As can be seen and expected, the sample mean

for the optimal stepsize is the most accurate, though it is not exactly on the vertical axis since $\overline{m}_1 = -0.038$. This implies that the sample size of $N = 1000$ is small. If we take $N = 10000$, the sample mean gives $\overline{m}_1 = 0.003$, signifying the convergence when N increases.

Take $N = 100000$ for the optimal stepsize case, and again compute the sample mean using `BayesianMCMC.m`. Is the sample mean better? If not, why?

However, the application of LLN and CLT is very limited for Markov chains since they don't provide iid samples. Indeed, as in the above Markov chain theory, the states of the chain eventually identically distributed by $\pi(m)$, but they are always correlated instead of independent since any state in the chain depends on the previous one. What we could hope for is that the current state is effectively independent from its kth previous state. In that case, the effective number of iid samples is N/k, and the mean square error, by the central limit theorem, decays as $\sqrt{k/N}$. As the result, if k is large, the decay rate is very slow. How to estimate k is the goal of the *autocorrelation* study, as we now discuss.

We shall compute the autocorrelation for each component of m separately, therefore, without loss of generality, assume that $m \in \mathbb{R}^1$ and that the Markov chain $\{m_j\}_{j=0}^N$ has zero mean. Consider the following discrete convolution quantities

$$c_k = \sum_{j=0}^{N-k} m_{j+k} m_j, \quad ,k = 0,\ldots,N-1,$$

and define the autocorrelation of m with *lag* k as

$$\hat{c}_k = \frac{c_k}{c_0}, \quad k = 0,\ldots,N-1.$$

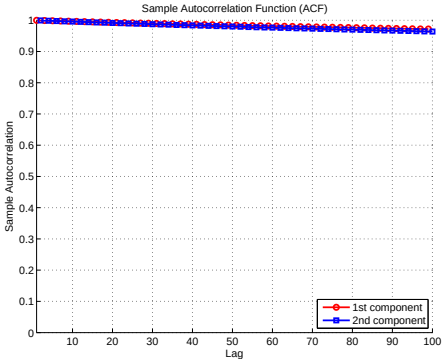
If $\hat{c}_k$ is zero, then we say that the correlation length of the Markov chain is approximately k, that is, any state m_j is considered to be insignificantly correlated to m_{j-k} (and hence any state before m_{j-k}), and to m_{j+k} (and hence any state after m_{j+k}). In other words, every kth sample point can be considered to be approximately independent. Note that this is simply a heuristic and one should be aware that independece implies un-correlation but not vice versa.

Let us now approximately compute the correlation length for three Markov chains corresponding to $\gamma = 0.02$, $\gamma = 0.5$, and $\gamma = 5$, respectively, with $N = 100000$. We first subtract away the sample mean as

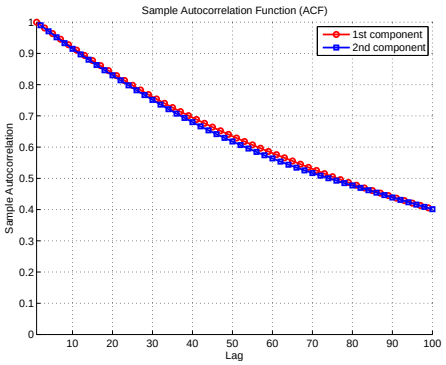
$$z_j = m_j - \frac{1}{N+1} \sum_{i=0}^{N} m_i.$$

Then, we plot the autocorrelation functions $\hat{c}_k$ for each component of the zero mean sample $\{z_j\}_{j=0}^N$ in Figure 11.3. As can be observed, the autocorrelation length for the chain with optimal stepsize $\gamma = 0.5$ is about $k = 100$, while the others are much larger (not shown here). That is, every 100th sample point can be considered to be independent for $\gamma = 0.5$. The case with $\gamma = 0.02$ is the worst, indicating slow move around the target density. The stepsize of $\gamma = 5$ is better, but so big that the chain

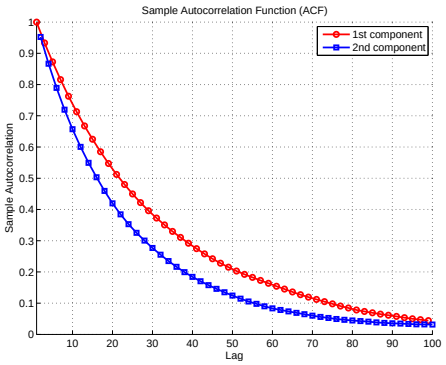
remains at each state for a long period of time, and hence autocorrelation length is still significant relatively to that of $\gamma = 0.5$.



(a) $\gamma = 0.02$



(b) $\gamma = 5$



(c) $\gamma = 0.5$

Fig. 11.3 Autocorrelation function plot for both components of m with different γ^2.

Extensive MCMC methods including improvements on the standard RWMH algorithm can be found in [**?**]. Let us introduce two simple modifications through the following two exercises.

Exercise 11.1. Consider the following target density

$$\pi(m) \propto \exp\left(-\frac{1}{2\delta^2}\|m\|^2 - \frac{1}{2\sigma^2}\|y - h(m)\|^2\right), \tag{11.1}$$

where

$$h(m) = \begin{bmatrix} m_1^2 - m_2 \\ m_2/5 \end{bmatrix}, \quad y = \begin{bmatrix} -0.2 \\ 0.1 \end{bmatrix}.$$

Take $\delta = 1$ and $\sigma = 0.1$.

1. Modify `BayesianMCMC.m` to simulate the target density in (11.1) with $N = 5000$.
2. Tune the proposal stepsize γ so that the average acceptance probability is about 0.234. Show the scatter, trace, and autocorrelation plots for the optimal stepsize.

Exercise 11.2. So far the proposal density $q(m,p)$ is isotropic and independent of the target density $\pi(m)$. For anisotropic target density, isotropic proposal is not a good idea, intuitively. The reason is that the proposal is distributed equally in all directions, whereas it is not in the target density. A natural idea is to shape the proposal density to make it locally resemble the target density. A simple idea in this direction is to linearize $h(m)$, and then define the proposal density as

$$q(m_k,p) \propto \exp\left(-\frac{1}{2\delta^2}\|p\|^2 - \frac{1}{2\sigma^2}\|y - h(m_k) - \nabla h(m_k)(p - m_k)\|^2\right),$$

1. Determine $H(m_k)$ such that $q(m_k,p) = \mathcal{N}\left(m_k, H(m_k)^{-1}\right)$, by keeping only the quadratic term in $p - m_k$.
2. Modify `BayesianMCMC.m` to simulate the target density in (11.1) using the proposal density $q(m_k,p) = \mathcal{N}\left(m_k, H(m_k)^{-1}\right)$. Show the scatter, trace, and autocorrelation plots. Is it better than the isotropic proposal density?

Exercise 11.3. Another idea to improve the standard RWMH algorithm is by adaptation. Let's investigate a simple adaptation strategy. Use the resulting sample in Exercise 11.1 to compute the empirical covariance $\hat{\Gamma}$, then use it to construct the proposal density $q(m,p) = \mathcal{N}\left(m, \hat{\Gamma}\right)$. Show the scatter, trace, and autocorrelation plots. Is it better than the isotropic proposal density?

Part III
Computational Methods for Large-Scale PDE-Constrained Bayesian Inversions

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Part IV
Introduction to concentration of measures

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Chapter 12
Concentration of Gaussian random variables

12.1 Important mathematical preliminaries

Lemma 12.1 (Markov inequality). *Let m be a non-negative random variable. There holds:*

$$\mathbb{P}[m \geq t] \leq \frac{\mathbb{E}[m]}{t}, \quad \forall t > 0.$$

Proof. We have

$$\mathbb{E}[m] = \mathbb{E}\left[m\mathbb{1}_{\{m\geq t\}}\right] + \mathbb{E}\left[m\mathbb{1}_{\{m<t\}}\right] \geq t\mathbb{E}\left[\mathbb{1}_{\{m\geq t\}}\right] \stackrel{\text{def}}{=} t\mathbb{P}[m \geq t],$$

where

$$\mathbb{1}_{\{m\geq t\}} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } m \geq t \\ 0 & \text{otherwise} \end{cases}.$$

□

Exercise 12.1 (Chebyshev and Chernoff inequalities). Using Markov inequality to show the following Chebyshev inequality:

$$\mathbb{P}[|m| \geq t] \leq \frac{\sigma^2}{t^2}, \quad \forall t > 0,$$

for any zero mean random variable m with variance σ^2.

Can you derive a Chebyshev inequality when $\mathbb{E}[m] = \overline{m}$?

We can see that the Chebyshev inequality improves the Markov inequality using the second moment. In fact we can use all the moments to drastically sharpen the inquality, and this due to Chernoff.

Show that, by using the monoticity of the exponential funciton (instead of the square function), there holds

$$\mathbb{P}[m \geq t] \leq \min_{\lambda>0} \frac{\mathbb{E}\left[e^{\lambda m}\right]}{e^{\lambda t}}, \quad \forall t > 0, \tag{12.1}$$

for any non-negative random variable m. That is, assume the *moment generating function (MGF)* $\mathbb{E}\left[e^{\lambda m}\right]$ is bounded, the tail probability of m decays exponentially. This "concentration phenomenon" will be made precise in this chapter for a large class of sub-gaussian random variables with bounded MGF. Exercise 12.2 provides a concentration result for Gaussian variables. •

Exercise 12.2 (A tail bound of normal distribution). We now apply the Chernoff inequality (12.1) to normal random variables, i.e. $m \sim \mathcal{N}\left(\overline{m}, \sigma^2\right)$.

1. Show that the MGF of m is given by

$$\mathbb{E}\left[e^{\lambda m}\right] = e^{\lambda \overline{m} + \frac{\lambda^2 \sigma^2}{2}}, \quad \forall t \in \mathbb{R}.$$

2. Solve the optimization on the right hand side of (12.1) to conclude that

$$\mathbb{P}\left[m - \overline{m} \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall t \geq 0. \tag{12.2}$$

•

Proposition 12.1 (Connection between tail bound and expectation). *For any random variable m, there holds:*

$$\mathbb{E}[m] = \int_0^\infty \mathbb{P}[m > t]\, dt - \int_{-\infty}^0 \mathbb{P}[m < t]\, dt.$$

Proof. Assume that $m \geq 0$, we have

$$m = \int_0^m dt = \int_0^\infty \mathbb{1}_{\{m > t\}}\, dt.$$

Taking expectation both sides and with the help with Fubini theorem we conclude

$$\mathbb{E}[m] = \int_0^\infty \mathbb{P}[m > t]\, dt.$$

Do you see this? The proof for the general case is similar by noting that

$$m = m\mathbb{1}_{m \geq 0} + m\mathbb{1}_{m < 0}.$$

□

Lemma 12.2 (Hutchinson 1989). *Let* $\mathbf{m}$ *be an n-dimensional random vector with* $\mathbb{E}[\mathbf{m}] = \mathbf{0}$ *and* $\mathbb{V}ar[\mathbf{m}] = \mathbb{E}\left[\mathbf{m}\mathbf{m}^T\right] = \mathbf{I}$. *Then:*

$$Tr(\mathscr{A}) = \mathbb{E}\left[\mathbf{m}^T \mathscr{A} \mathbf{m}\right],$$

for any matrix $\mathscr{A} \in \mathbb{R}^{n \times n}$.

Do you see this? *Proof.* We have

$$Tr(\mathscr{A}) = Tr(\mathscr{A}\mathbb{V}ar[\mathbf{m}]) = \mathbb{E}\left[Tr\left(\mathscr{A}\mathbf{m}\mathbf{m}^T\right)\right] = \mathbb{E}\left[Tr\left(\mathbf{m}^T\mathscr{A}\mathbf{m}\right)\right] = \mathbb{E}\left[\mathbf{m}^T\mathscr{A}\mathbf{m}\right].$$

□

A useful bound that we use repeatedly is the *union bound*.

Lemma 12.3 (Union bound).

$$\mathbb{P}[A_1 \cup A_2 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \mathbb{P}[A_i], \quad \forall n \in \mathbb{N}.$$

12.2 Concentration of sum of scalar Gaussian random variables

Let us start with an interesting application of the Chebyshev inequality, namely, a version of the weak law of large numbers. Let $m_i \sim \mathscr{N}\left(\overline{m}, \sigma^2\right)$, and $S_N \stackrel{\text{def}}{=} \frac{1}{N}(m_1 + m_2 + \cdots + m_N)$. We can show *Can you easily show this?*

$$S_N \sim \mathscr{N}\left(\overline{m}, \sigma^2/N\right),$$

which together with the Chebyshev inequality leads to

$$\mathbb{P}[|S_N - \overline{m}| \geq t] \leq \frac{1}{N}\frac{\sigma^2}{t^2},$$

which in turns yields

$$\lim_{N\to\infty} \mathbb{P}[|S_N - \overline{m}| \geq t] = 0, \quad \forall t > 0,$$

that is, the probability that S_N deviates from its mean approaches 0 as the number of i.i.d random summands increases. This is of course not surprising since we already know from the LLN that S_N indeed converges to $\overline{m}$ almost surely. What is more interesting is whether the quadratic decay of the deviation is sharp. The answer is that it is very conservative in this case. To see this, we apply (12.2) to $S_N - \overline{m}$ to obtain

$$\mathbb{P}[|S_N - \overline{m}| \geq t] \leq 2e^{-N\frac{t^2}{2\sigma^2}},$$

which shows that *a t-deviation of S_N from its mean decays exponentially in both t and N*.

In this chapter by concentration we refer to the phenomenon that a random variable concentrates around its mean. We quantify this phenomenon via concentration inequality of the form

$$\mathbb{P}[|m - \overline{m}| \geq t] \leq \text{ small quantity}, \tag{12.3}$$

i.e. the *tail bound*.

Remark 12.1. It is important to see that while results from CLT and LLN are asymptotic (valid for $N \to \infty$), concentration inequalities are non-asymptotic (valid for finite N).

We have shown that normal random variables have exponential tail bounds which decay rapidly (the best possible we argue). It turns out that this behavior is shared by a large class of *sub-gaussian random variables*.

Chapter 13
Concentration of sub-Gaussian random variables

Definition 13.1 (Sub-gaussian random variables[1]). A random variable m is called sub-gaussian if its MGF is dominated[2] by a mean zero normal random variable with variance σ^2, i.e.,

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{13.1}$$

m is sometimes called a σ-sub-gaussian or a sub-gaussian with *proxy* σ^2. A direct consequence of the definition 1 is that a sub-gaussian random variable has zero mean and its variance is bounded above by σ^2.

Proposition 13.1. *If m is a σ-sub-gaussian, then $\mathbb{E}[m] = 0$ and $\mathbb{V}ar[m] \leq \sigma^2$.*

Proof. For any λ, using Taylor expansion for both sides of (13.1) we have *Why the first equality is true?*

$$\sum_{n=0}^{\infty} \frac{\lambda^n}{n!}\mathbb{E}[m^n] = \mathbb{E}\left[e^{\lambda m}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} = \sum_{n=0}^{\infty} \frac{\lambda^{2n}\sigma^{2n}}{2^n n!}.$$

Now dividing both sides by $\lambda > 0$ and taking the limit $\lambda \to 0$, we conclude that $\mathbb{E}[m] = 0$. Next, dividing both sides by λ^2 and again taking the limit $\lambda \to 0$ we can conclude that $\mathbb{V}ar[m] \leq \sigma^2$. □

The following theorem characterizes sub-gaussian random variables.

Theorem 13.1 (Sub-gaussian properties). *Let m be a random variable. Then the following are equivalent:*

i) There exists a constant c_1 such that the tail of m satisfies

$$\mathbb{P}[|m| \geq t] \leq 2e^{-\frac{t^2}{2c_1^2}}, \quad \forall t \geq 0. \tag{13.2}$$

[1] A weaker definition is based on the tail bound, i.e., m is a sub-gaussian random variable if

$$\mathbb{P}[|m| \geq t] \leq 2e^{-t^2/c^2},$$

for some constant c.

[2] That is, the Laplace transform of m is dominated by the Laplace transform of $\mathcal{N}(0, \sigma^2)$.

ii) There exists a constant c_2 such that the moments of m satisfy

$$\|m\|_p \overset{def}{=} (\mathbb{E}[|m|^p])^{1/p} \leq c_2\sqrt{p}, \quad \forall p \geq 1.$$

iii)There exists a constant c_3 such that the MGF of m^2 satisfies

$$\mathbb{E}\left[e^{\lambda^2 m^2}\right] \leq e^{c_3^2\lambda^2}, \quad \forall |\lambda| \leq \frac{1}{c_3}.$$

Moreover, if $\mathbb{E}[m] = 0$, then all the above are equivalent to

iv) There exists a constant c_4 such that the MGF of m satisfies

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Proof. We prove *i*) $\Rightarrow$ *ii*), *ii*) $\Rightarrow$ *iii*), *iii*) $\Rightarrow$ *i*), *iv*) $\Rightarrow$ *i*), and *iii*) $\Rightarrow$ *iv*)

- *i*) $\Rightarrow$ *ii*): Using Proposition 12.1 we have

$$\mathbb{E}[|m|^p] = \int_0^\infty \mathbb{P}[|m|^p > t]\, dt = \int_0^\infty \mathbb{P}[|m| > u]\, pu^{p-1}\, du \leq 2\int_0^\infty pu^{p-1} e^{-\frac{u^2}{2c_1^2}}\, du$$
$$= \left(\sqrt{2}c_1\right)^p 2\int_0^\infty pt^{p-1}e^{-t^2}\, dt \leq \left(\sqrt{2}c_1\right)^p p\Gamma(p/2) \leq \left(\sqrt{2}c_1\right)^p p\,(p/2)^{p/2},$$

  where we have used a change of variable $t = u^p$ in the second equality, i) in the first inequality, definition of the Gamma function in the third equality, and a Stirling's approximation in the second inequality. Taking the *pth*-root both side and using the fact that $p^{1/p} < 2$ ends the proof with $c_2 = 2c_1$.
- *ii*) $\Rightarrow$ *iii*): By Taylor expansion and the monotone convergence theorem we have

$$\mathbb{E}\left[e^{\lambda^2 m^2}\right] = 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{k!}\mathbb{E}\left[m^{2k}\right] \leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{(k/e)^k}\left(8c_1^2\right)^k k^k = \frac{1}{1 - 8c_1^2 e\lambda^2},$$

  where we have used *ii*) and a Stirling approximation $k! \geq (k/e)^k$ in the first inequality. The last equality holds provided that $8c_1^2 e\lambda^2 \leq 1$. We can further bound the right hand side if we take λ such that $8c_1^2 e\lambda^2 \leq 1/2$ and use the inequality $(1-x)^{-1} \leq e^{2x}$ for $x \in [0, 1/2]$, i.e.,

$$\mathbb{E}\left[e^{\lambda^2 m^2}\right] \leq e^{16c_1^2 e\lambda^2},$$

  which ends the proof by taking $c_3^2 = 16c_1^2 e$.
- *iii*) $\Rightarrow$ *i*): we have

$$\mathbb{P}[|m| \geq t] = \mathbb{P}\left[e^{m^2} \geq e^{t^2}\right] \overset{Markov}{\leq} \frac{\mathbb{E}\left[e^{m^2}\right]}{e^{t^2}},$$

which ends the proof by applying *iii*) with $\lambda = c_3 = 1$.

- *iv*) $\Rightarrow$ *i*): Using Chernoff inequality (12.1) we have

$$\mathbb{P}[m \geq t] \leq \min_{\lambda>0} \frac{\mathbb{E}\left[e^{\lambda m}\right]}{e^{\lambda t}} \leq \min_{\lambda>0} e^{\sigma^2\lambda^2/2-\lambda t} = e^{-\frac{t^2}{2\sigma^2}},$$

where we have used *iv*) in the second inequality.

- *iii*) $\Rightarrow$ *iv*): Since Note that

$$e^m \leq m + e^{m^2}, \quad \forall m \in \mathbb{R}$$

and thus

$$\mathbb{E}\left[e^{\lambda m}\right] \leq \mathbb{E}[\lambda m] + \mathbb{E}\left[e^{\lambda^2 m^2}\right] \leq e^{\lambda^2},$$

for $|\lambda| \leq 1$, where we have used the fact that $\mathbb{E}[m] = 0$ and *iii*) (by choosing $c_3 = 1$) in the last inequality. Now for $|\lambda| > 1$, starting from

$$\lambda m \leq \frac{1}{2}\left(\lambda^2 + m^2\right),$$

we have

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\lambda^2/2} e^{m^2/2} \leq e^{\lambda^2/2} e^{1/2} \leq e^{\lambda^2},$$

where we have used *iii*) (by choosing $c_3 = 1$) in the second inequality and $|\lambda| > 1$ in the last inequality. We thus obtain

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\lambda^2}, \quad \forall \lambda \in \mathbb{R},$$

and this concludes the proof.

□

Remark 13.1. Note that the first three assertions are equivalent for any random variable. The equivalence says that if a random variable has a exponential decaying tail bound of the form (13.2), not only its expectation (see Proposition 12.1) is bounded, but its L^p-norm grows like $\mathcal{O}(\sqrt{p})$. It also says that the exponential decaying tail bound (13.2) is necessary and sufficient for the integrability of fast growing function $e^{\lambda^2 m^2}$. Assertion *iv*) tells us that sub-gaussian random variables have all these properties, which is not surprising since a Gaussian random varialbe is also a sub-gaussian random variable.

Similar to Gaussian distributions, a finite sum of sub-gaussian random variables is sub-gaussian.

Proposition 13.2 (Sum of independent sub-gaussians). *Assume $m_1, \ldots, m_N$ are independent, sub-gaussian random variables with proxy c_i^2, then $\sum_{i=1}^{N} \mathbf{a}_i m_i$ is also a sub-gaussian random variable with proxy $\sum_{i=1}^{N} c_i^2 \mathbf{a}_i^2$.*

Proof. The proof is straightforward:

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{N} \mathbf{a}_i m_i}\right] = \Pi_{i=1}^{N} \mathbb{E}\left[e^{\lambda \mathbf{a}_i m_i}\right] \leq e^{\lambda^2 \sum_{i=1}^{N} c_i^2 \mathbf{a}_i^2}.$$

□

Chapter 14
Basic concentration Inequalities

We are now in the position to prove a concentration result for a sum of sub-gaussian random variables.

Theorem 14.1 (General Hoeffding inequality). *Assume $m_1, \dots, m_N$ are independent, sub-gaussian random variables, then there exists a constant c such that the following concentration inequality holds:*

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} \mathbf{a}_i m_i\right| > t\right] \leq 2e^{-\frac{t^2}{c^2\|\mathbf{a}\|^2}}, \quad \forall t \geq 0.$$

Proof. Using the Chernoff inequality we have

$$\mathbb{P}\left[\sum_{i=1}^{N} \mathbf{a}_i m_i \geq t\right] \leq \min_{\lambda>0} \frac{\mathbb{E}\left[e^{\lambda \sum_{i=1}^{N} \mathbf{a}_i m_i}\right]}{e^{\lambda t}} \leq \min_{\lambda>0} \frac{e^{\lambda^2 \sum_{i=1}^{N} c_i^2 \mathbf{a}_i^2}}{e^{\lambda t}} = e^{-\frac{t^2}{2\sum_{i=1}^{N} c_i^2 \mathbf{a}_i^2}},$$

where we have used Proposition 13.2 in the third inequality. Similarly, we can show that

$$\mathbb{P}\left[\sum_{i=1}^{N} \mathbf{a}_i m_i \leq -t\right] \leq e^{-\frac{t^2}{2\sum_{i=1}^{N} c_i^2 \mathbf{a}_i^2}},$$

which together with the previous inequality concludes the proof. □

Corollary 14.1. *Show that there exists a constant c such that*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} m_i\right| > t\right] \leq 2e^{-N\frac{t^2}{c^2}}, \quad \forall t \geq 0.$$

The results of Corollary 14.1 is a direct extension of concentration phenomenon observed from Gaussian random variables that we discuss at the beginning of this section. This, together with the LLN, says that the sample mean not only converges almost surely to the mean but also concentrates at the mean. The Hoeffding inequality quantifies this concentration via the exponential decaying tail probability.

Exercise 14.1. Let m be a Rademacher (aka symmetric Bernoulli) random variables , i.e., probability of m being either 1 or -1 is 1/2. Show that *Rademacher distribution is also a sub-gaussian*, i.e.,

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\frac{\lambda^2}{2}}. \tag{14.1}$$

•

Exercise 14.2 (Hoeffding Lemma). Let m be a random variable such that $\mathbb{E}[m] = 0$ and $m \in [a,b]$ almost surely. Show that

$$\mathbb{E}\left[e^{\lambda m}\right] \leq e^{\lambda^2 \frac{(b-a)^2}{8}}.$$

In particular, *mean zero bounded random variables are subgaussian*.

Hint. With the help of Jensen inequality, show that

$$\mathbb{E}_m\left[e^{\lambda m}\right] = \mathbb{E}_m\left[e^{\lambda \mathbb{E}_{m'}[m-m']}\right] \leq \mathbb{E}_{m,m'}\left[e^{\lambda(m-m')}\right],$$

and then show that

$$\mathbb{E}_{m,m'}\left[e^{\lambda(m-m')}\right] = \mathbb{E}_{m,m'}\left[\mathbb{E}_y\left[e^{\lambda y(m-m')}\right]\right],$$

where y is a Rademacher random variable. Then finish the work with the help of (14.1). •

Exercise 14.3 (Hoeffding inequality for sum of bounded random variables). Let $m_i, i = 1,\ldots,N$ be independent bounded random variables such that $m_i \in [a_i,b_i]$ almost surely. Show that

$$\mathbb{P}\left[|S_N - \mathbb{E}[S_N]| > t\right] \leq 2e^{-2N^2 \frac{t^2}{\sum_{i=1}^N (b_i-a_i)^2}}, \quad \forall t \geq 0,$$

where $S_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^N m_i$.

Deduce a concentration inequality for $\displaystyle\sum_{i=1}^N m_i$.

Application. Toss a fair coin N time, what is the probability of getting at least $3N/4$ heads? •

Definition 14.1 (ℓ-percent sparse random variables [20, 18]).

Let $s = \dfrac{1}{1-\ell}$ where $\ell \in [0,1)$ is the level of sparsity desired. Then

$$\zeta = \sqrt{s}\begin{cases} +1 & \text{with probability } \dfrac{1}{2s}, \\ 0 & \text{with probability } \ell = 1 - \dfrac{1}{s}, \\ -1 & \text{with probability } \dfrac{1}{2s} \end{cases} \tag{14.2}$$

is a ℓ-*percent sparse distribution.*

Note that for $\ell = 0$, ζ corresponds to a Rademacher distribution, and that $\ell = 2/3$ corresponds to the *Achlioptas distribution* [1]. By inspection we have that $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}\left[\zeta^2\right] = 1$. As shall be shown, this class of random variable is important for large-scale application involving random vectors/matrices, since vectors/matrices generated from this class of randomn variables can be very sparse depending on ℓ.

Exercise 14.4. Show that $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}\left[\zeta^2\right] = 1$. Morever, show that ℓ-*percent sparse random variables are sub-gaussian* and determine their proxy. •

In practice, random variables do not always have zero mean. In that case we can center a random variable by substracting out its mean. We can then define a random variable m with mean $\overline{m}$ as a sub-gaussian with proxy σ^2 by saying that $m - \overline{m}$ is a sub-gaussian, i.e.,

$$\mathbb{E}\left[e^{\lambda(m-\overline{m})}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

All the above results are valid for $m - \overline{m}$.

Exercise 14.5. Prove a concentration result similar to that of Corollary 14.1 for the case when $m_i - \overline{m}_i$ are sub-gaussian with proxy σ_i^2. •

Chapter 15
Some applications of concentration inequalities

15.1 Some large-scale matrix computation with randomization

Next is an application of the concentration inequality to a few large-scale computational problems in scientific computing. For concreteness, let us consider the problem of estimating the trace of a large matrix $Tr(\mathscr{A})$ (other interesting problems can be found in [4]). From Lemma 12.2 we know that

$$Tr(\mathscr{A}) = \mathbb{E}\left[\mathbf{m}^T \mathscr{A} \mathbf{m}\right] \overset{\text{Monte Carlo}}{\approx} \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}_i^T \mathscr{A} \mathbf{m}_i =: S_N,$$

where $\mathbf{m}$ is an arbitrary random vector with zero mean and $\mathbb{E}\left[\mathbf{m}\mathbf{m}^T\right] = I$. Clearly, S_N is an unbiased estimator for $Tr(\mathscr{A})$ and it converges almost surely to $Tr(\mathscr{A})$. The question that we are interested here is how many samples is "enough" for such an estimator. Addressing this question is of practical important since we want to minimize the cost, i.e. choosing as small ensemble size N as we can, while having an accurate estimator. In this section we consider symmetric positive definite matrix $\mathscr{A}$ which admits computable lower and upper bounds [4] for $\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i$, i.e.,

$$L_i \le \mathbf{m}_i^T \mathscr{A} \mathbf{m}_i \le U_i.$$

Thus each $\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i$ is a bounded random variable with mean $\mathbb{E}\left[\mathbf{m}_i^T \mathscr{A} \mathbf{m}_i\right] = Tr(\mathscr{A})$. Applying the Hoeffding inequality from Exercise 14.3 we have

$$\mathbb{P}\left[|S_N - Tr(\mathscr{A})| > t\right] \le 2e^{-2N^2 \frac{t^2}{\sum_{i=1}^{N}(U_i - L_i)^2}}, \quad \forall t \ge 0.$$

In other words,

$$-t \le S_N - Tr(\mathscr{A}) \le t \tag{15.1}$$

holds with probability

$$1 - 2e^{-2N^2 \frac{t^2}{\sum_{i=1}^{N}(U_i - L_i)^2}}.$$

If we now can pick a tolerance t and a success probability β, we can solve

$$1-2e^{-2N^2\frac{t^2}{\sum_{i=1}^N (U_i-L_i)^2}}=\beta$$

for the ensemble size N as

$$N=\sqrt{\frac{\sum_{i=1}^N (U_i-L_i)^2}{2t^2}\ln\left(\frac{2}{1-\beta}\right)},$$

which is the desire ensemble size to obtain the error bound (15.1) with probability β.

Exercise 15.1. Your task is to estimate the trace of the matrix in the first numerical example of [4]. Pick a few pairs (t,β) (some with big β, say, 0.5) and report your observations. •

15.2 Dimension reduction with random projection

We now show that concentration inequalities are the key to the success of many randomized methods for dimension reduction of big data. To begin, assume $\mathbf{x}\in\mathbb{R}^N$ and consider a random matrix $\mathscr{A}\in\mathbb{R}^{n\times N}$ whose entries, $\mathscr{A}_{ij}$ *are i.i.d random variables with zero mean and unit variance*. Let us define the following random "projection" $\mathscr{P}$

$$\mathbf{z}=\mathscr{P}\mathbf{x}:=\frac{1}{\sqrt{n}}\mathscr{A}\mathbf{x},$$

and we are going to show that with high probability the *random projection* $\mathscr{P}$ preserves length.

Exercise 15.2. Show that components of $\mathbf{z}$, i.e. $\mathbf{z}_i=\mathscr{A}(i,:)\mathbf{x}, i=1\ldots,n$ are i.i.d. random variables with

$$\mathbb{E}[\mathbf{z}_i]=0,\quad \text{and } \mathbb{V}ar[\mathbf{z}_i]=\frac{\|\mathbf{x}\|^2}{n},$$

In particular, show that $\mathbb{E}\left[\|\mathbf{z}\|^2\right]=\|\mathbf{x}\|^2$. •

Exercise 15.2 shows a remarkable fact that, on average, mapping via random matrix with i.i.d. random entries having mean zero and unit variance preserves the length. Clearly, we are more interested in the performance of an actual realization of $\mathbf{z}$, i.e., how far $\mathbf{z}^2$ is from its mean. We address this question probabilistically. To that end we observe that $\mathbf{z}_i^2$ are i.i.d. random variables and the expectation of their sum is $\|\mathbf{x}\|^2$, and we shall show that the sum actually concentrates around the mean. By Chernoff inequality we have

$$\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon \|\mathbf{x}\|^2\right] = \mathbb{P}\left[n\|\mathbf{z}\|^2 \geq n(1+\varepsilon)\|\mathbf{x}\|^2\right]$$
$$\leq \min_{\lambda} e^{-n\lambda(1+\varepsilon)\|\mathbf{x}\|^2} \Pi_{i=1}^{n} \mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right],$$

and similar to other concentration results that we have seen so far, the task at hand is to bound the MGF $\mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right]$. Before considering general sub-gaussian random variable, let us first study the case when $\mathscr{A}_{ij}$ are standard normal random variables, for which we can evaluate the MGF exactly.

Exercise 15.3. Suppose $\zeta \sim \mathcal{N}\left(0, \sigma^2\right)$ and $t \leq 1/\left(2\sigma^2\right)$. Show that

$$\mathbb{E}\left[e^{t\zeta^2}\right] = \frac{1}{\sqrt{1-2t\sigma^2}}.$$

Then deduce that $\mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right] = \dfrac{1}{\sqrt{1-2\lambda\|\mathbf{x}\|^2}}$ for $\lambda \leq 1/\left(2\|\mathbf{x}\|^2\right)$.

Hint. Direct evaluation of the integral and the fact that $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} e^{-\frac{\theta^2}{2}}\, d\theta = 1$.

•

The result of Exercise 15.3 together with the fact that $\mathbf{z}_i$ are i.i.d. yields *Check it*

$$\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon \|\mathbf{x}\|^2\right] \leq \min_{\lambda} e^{\frac{n}{2} f(\lambda)},$$

where we have defined $f(\lambda) := -2\lambda(1+\varepsilon)\|\mathbf{x}\|^2 - \ln\left(1-2\lambda\|\mathbf{x}\|^2\right)$. It is easy to see that, for $0 \leq \lambda \leq 1/\left(2\|\mathbf{x}\|^2\right)$, $f(\lambda)$ attains its minimum at $\lambda^* = \dfrac{\varepsilon}{2(1+\varepsilon)\|\mathbf{x}\|^2}$. *Show it*

Thus, we have

$$\mathbb{P}\left[\|\mathbf{z}\|^2 \geq (1+\varepsilon)\|\mathbf{x}\|^2\right] \leq e^{\frac{n}{2}[\ln(1+\varepsilon)-\varepsilon]} \leq e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)},$$

where we have used the fact that, for $\varepsilon \in [0,1]$, $\ln(1+\varepsilon) - \varepsilon \leq -\dfrac{\varepsilon^2}{2} + \dfrac{\varepsilon^3}{3}$. *Check it*

Exercise 15.4. Show that

$$\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\|\mathbf{x}\|^2\right] \leq e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.$$

•

Together with the union bound we obtain the following concentration inequality

$$\boxed{\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\|\mathbf{x}\|^2 \text{ or } \|\mathbf{z}\|^2 \geq (1+\varepsilon)\|\mathbf{x}\|^2\right] \leq 2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.} \qquad (15.2)$$

That is, ε-distortion in length via the random projection $\mathscr{P}$ is less than $2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}$. In other words, with high probability, i.e., $1-2e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}$, the random projection $\mathscr{P}$ preserves the length of $\mathbf{x}$.

Exercise 15.5 (Concentration inequality for χ^2 distribution). We have essentially proved a concentration inequality for Chi-square distribution. To see this, let us recall a Chi-square distribution with n degrees of freedom, typically denoted as χ^2_n, is the sum of square of n i.i.d. standard normal random variables. That is, if $m \sim \chi^2_n$, then X can be represented as $m = \sum_{i=1}^{n} \xi_k^2$ where $\xi_k \sim \mathscr{N}(0,1)$. Show that m concentrates around its mean with the tail bound given by (15.2).

Hint. Rescale the random projection matrix $\mathscr{P}$ appropriately and judiciously pick a vector $\mathbf{x}$ so that $\|\mathbf{z}\|^2 \sim \chi^2_n$.

•

Suppose now we have m vectors $\mathbf{x}_i \in \mathbb{R}^N, i = 1,\ldots,m$, where $N \gg 1$ and we are interested in reducing the dimension of these vectors while (approximately) preserving their geometry. We are going to show that this task can be acommplished by the random projection $\mathbf{y}_i = \mathscr{P}\mathbf{x}_i \in \mathbb{R}^n$. Here, by preserving geometry we mean

$$\|\mathbf{y}_i\| \approx \|\mathbf{x}_i\|, \quad \text{and } \|\mathbf{y}_i - \mathbf{y}_j\| \approx \|\mathbf{x}_i - \mathbf{x}_j\|,$$

that is we like to reduce the dimension of the data vectors, but we desire to preserve their norm and their mutual distances. Since $\mathscr{P}$ is linear, it is sufficient to show that the latter holds. Since we have m vectors, we have $m(m-1)/2$ distinct difference vectors $\mathbf{x}_i - \mathbf{x}_j$ (called "pairs"). If we project each "pair" $\mathbf{x}_i - \mathbf{x}_j$ to obtain the corresponding vector $\mathbf{y}_i - \mathbf{y}_j$, the union bound and the concentration inequality (15.2) give

$$\begin{aligned}\mathbb{P}[\text{Some pair has } \varepsilon\text{-distortion}] &\leq \sum_{i=1}^{m(m-1)/2} \mathbb{P}[\text{Pair } i \text{ has } \varepsilon\text{-distortion}] \\ &= \frac{m(m-1)}{2}\mathbb{P}[\text{Pair 1 has } \varepsilon\text{-distortion}] \leq m(m-1)\, e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}.\end{aligned}$$

Thus, if desire

$$\mathbb{P}[\text{Some pair has } \varepsilon\text{-distortion}] \leq \frac{1}{m^\beta},$$

we can enforce

$$m(m-1)\, e^{\frac{n}{2}\left(-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)} \leq \frac{1}{m^\beta},$$

i.e.,

$$\boxed{n \geq \frac{2\beta\log(m) + 2\log(m(m-1))}{\varepsilon^2/2 - \varepsilon^3/3}.} \tag{15.3}$$

We have derived a version of the Johnson-Lindenstrauss lemma.

Lemma 15.1 (Johnson-Lindenstrauss Lemma). *Consider m vectors $\mathbf{x}_i \in \mathbb{R}^N, i = 1,\dots,m$, where $N \gg 1$. Define a random projection matrix $\mathscr{P} = \mathscr{A}/\sqrt{n} \in \mathbb{R}^{n\times N}$, where each component of $\mathscr{A}$ is i.i.d. standard normal random variable. For any $\beta > 0$, if we choose a reduced dimension n satisfying* (15.3)*, then with probability at least $1 - m^{-\beta}$ we have*

$$(1-\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \le \left\|\mathscr{P}\mathbf{x}_i - \mathscr{P}\mathbf{x}_j\right\|^2 \le (1+\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2, \quad \forall i,j = 1,\dots,m.$$

Remark 15.1. Lemma 15.1 shows that i.i.d. Gaussian random matrices can be used to reduce the dimension of high-dimensional data sets while almost preserving the geometry with high probability. The beauty here is that the reduced dimension is independent of the original data dimension. Unlike other dimensional reduction methods, random projections do not require the low dimensionality of the original data set.

Exercise 15.6. In this exercise, we are going to numerically verify the Johnson-Lindenstrauss lemma. To that end, pick a few dimensions for the original *big-data*, say $N = \{100, 1000, 3000, 6000\}$. For each N, generate $m = 3000$ uniform (or Gaussian) random vectors of dimensional N. Pick a distortion value, say $\varepsilon = 0.25$, and β so that the successful probability is 0.75. Now pick the minimum reduced dimension n using (15.3) and then compute the actual distortions for m data vectors and plot the histogram of the actual distortions for each N. Plot the mean, the min, and the max actual distortions for each N and, on the same figure for each case, plot the predicted distortion lines $1+\varepsilon$ and $1-\varepsilon$. Discuss on the validity of the Johnson-Lindenstrauss lemma. You may want check various values of β ranging from low to high successful probability. Also, pick a few values of ε and report your findings. •

Let us now extend the result to α-sub-gaussian distributions, i.e., $\mathscr{A}_{ij}$ are i.i.d. sub-gaussian random variables with zero mean and unit variance. All we need is to bound the MGF $\mathbb{E}\left[e^{\lambda n \mathbf{z}_i^2}\right]$ which, unlike the Gaussian case, does not admit a closed form expression. Nevertheless, by Proposition 13.2, $\mathbf{z}_i$ is a sub-gaussian with proxy $\alpha^2 \|\mathbf{x}\|^2/n$ and $n\mathbf{z}_i^2 - \left\|\mathbf{x}^2\right\|$ is a zero-mean random variable. We will use this fact to bound the tail and in turn the MGF of $n\mathbf{z}_i^2 - \|\mathbf{x}\|^2$. We start with the following general result.

Lemma 15.2. *Let m be a zero-mean random variable with the tail bound*

$$\mathbb{P}[|m| \ge t] \le 2e^{-2t/\beta} \tag{15.4}$$

for some $\beta > 0$. Then the MGF of m satisfies

$$\mathbb{E}[e^{sm}] \le e^{2s^2\beta^2}, \quad \forall |s| \le \frac{1}{2\beta}. \tag{15.5}$$

Proof. We observe that the statement is similar to the equivalent between *i*) and *iv*) in Theorem 13.1. Indeed, similar to the proof of *ii*) in Theorem 13.1 we have

$$\mathbb{E}\left[|m|^p\right] = 2\left(\beta/2\right)^p p\Gamma\left(p\right) \leq \beta^p p!$$

which, together with the fact that $\mathbb{E}\left[m\right] = 0$, gives

$$\mathbb{E}\left[e^{sm}\right] = 1 + \sum_{p=2}^{\infty} \frac{s^p \mathbb{E}\left[m^p\right]}{p!} \leq 1 + \sum_{p=2}^{\infty} \left(s\beta\right)^p = 1 + \frac{s^2\beta^2}{1-s\beta},$$

for $|s\beta| < 1$, and for $|s|\,\beta \leq 1/2$ we arrive at

$$\mathbb{E}\left[e^{sm}\right] \leq 1 + 2s^2\beta^2 \leq e^{s^2\beta^2},$$

which ends the proof.

To apply Lemma 15.2 to the random variable $n\mathbf{z}_i^2 - \|\mathbf{x}\|^2$ we just need to bound its tail. We have

$$\mathbb{P}\left[n\mathbf{z}_i^2 - \|\mathbf{x}\|^2 > \varepsilon\,\|\mathbf{x}\|^2\right] = \mathbb{P}\left[|\mathbf{z}_i| > \sqrt{\frac{1+\varepsilon}{n}}\,\|\mathbf{x}\|\right] \leq 2e^{-\frac{(1+\varepsilon)\|\mathbf{x}\|^2}{2\alpha^2\|\mathbf{x}\|^2}} \leq 2e^{-\frac{\varepsilon\|\mathbf{x}\|^2}{2\alpha^2\|\mathbf{x}\|^2}},$$

where have used the tail bound of $\mathbf{z}_i$ in the first inequality. Now applying Lemma 15.2 with $t = \varepsilon\,\|\mathbf{x}\|^2$ and $\beta = 4\alpha^2\,\|\mathbf{x}\|^2$ we have

$$\mathbb{E}\left[e^{\lambda\left(n\mathbf{z}_i^2 - \|\mathbf{x}\|^2\right)}\right] \leq e^{32\alpha^4\|\mathbf{x}\|^4\lambda^2}.$$

Consequently, for $\lambda < \dfrac{1}{8\alpha^2\,\|\mathbf{x}\|^2}$, we have

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{z}\|^2 - \|\mathbf{x}\|^2 \geq \varepsilon\,\|\mathbf{x}\|^2\right] &\leq \min_{\lambda} e^{-n\varepsilon\lambda\|\mathbf{x}\|^2}\Pi_{i=1}^n \mathbb{E}\left[e^{\lambda\left(n\mathbf{z}_i^2 - \|\mathbf{x}\|^2\right)}\right] \\ &\leq \min_{\lambda} e^{32\alpha^4\|\mathbf{x}\|^4\lambda^2 - \varepsilon\lambda\|\mathbf{x}\|^2} = e^{-n\frac{\varepsilon^2}{128\alpha^4}},\end{aligned}$$

and together with the union bound we obtain the concentration inequality

$$\boxed{\mathbb{P}\left[\|\mathbf{z}\|^2 \leq (1-\varepsilon)\,\|\mathbf{x}\|^2 \text{ or } \|\mathbf{z}\|^2 \geq (1+\varepsilon)\,\|\mathbf{x}\|^2\right] \leq 2e^{-n\frac{\varepsilon^2}{128\alpha^4}}.}$$

Similar to the Gaussian case, we conclude that with high probability, i.e., $1 - 2e^{-n\frac{\varepsilon^2}{128\alpha^4}}$, the random projection $\mathscr{P}$ with i.i.d α-sub-gaussian random entries preserves the length of $\mathbf{x}$. This shows that the Johnson-Lindenstrass Lemma is also valid for sub-gausian random projection.

Note that random variables with MGF satisfying (15.5) is called *sub-exponential*. The name is due to the exponential tail (15.4), which is heavier than Gaussian tail of sub-gaussian random variables. As a by-product of proving the Johnson-Lindenstrauss lemma for sub-gaussian distrubtions, we have shown that square

of sub-gaussian random variables is sub-exponential[1]. More importantly, we have shown that sub-exponential distributions also obey concentration inequalities. A more complete concentration inequality for sub-exponential random variables is given by Bernstein's inequality.

Theorem 15.1 (Bernstein's inequality). *Let $m_i, i = 1,\dots,N$ be zero mean sub-exponential random variables, i.e., m_i satisfies the tail bound*

$$\mathbb{P}[|m_i| \geq t] \leq 2e^{-2t/\beta_i}.$$

Then, for any $t \geq 0$, we have

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} m_i\right| \geq t\right] \leq 2\min\left\{e^{-\frac{t^2}{8\sum_{i=1}^{N}\beta_i^2}}, ce^{-\frac{t}{2\max_i \beta_i}}\right\},$$

for some constant $c = e^{\frac{\sum_{i=1}^{N}\beta_i^2}{2\max_i \beta_i^2}}$.

Proof. Again, Chernoff inequality is the key of the proof. We have

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq t\right] \leq \min_{\lambda>0} \frac{\Pi_{i=1}^{N}\mathbb{E}\left[e^{\lambda m_i}\right]}{e^{\lambda t}} \leq \min_{\lambda \leq 1/(2\max_i \beta_i)} e^{2\lambda^2\sum_{i=1}^{N}\beta_i^2 - \lambda t},$$

where we have used (15.5). Now optimizing λ, i.e. $\lambda^* = \min\left\{\frac{t}{4\sum_{i=1}^{N}\beta_i^2}, \frac{1}{2\max_i \beta_i}\right\}$, we obtain

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq t\right] \leq \min\left\{e^{-\frac{t^2}{8\sum_{i=1}^{N}\beta_i^2}}, ce^{-\frac{t}{2\max_i \beta_i}}\right\},$$

and this concludes the proof.

Compared to Proposition 13.2, the tail bound for sum of independent sub-exponential random variables is only Gaussian for small deviation t. For large deviation, the tail is exponential, and hence is heavier than Gaussian tails. Analogous results to Theorem 14.1 and Corollary 14.1 are given in the following exercises.

Exercise 15.7. Let $m_i, i = 1,\dots,N$ be sub-exponential random variables as in Theorem 15.1 and $\mathbf{a} \in \mathbb{R}^N$. Show that

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} \mathbf{a}_i m_i\right| \geq t\right] \leq 2\min\left\{e^{-\frac{t^2}{c_1}}, c_2 e^{-\frac{t}{c_3}}\right\},$$

and determine c_1, c_2, c_3. Then deduce the following upper bound

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} m_i\right| \geq t\right] \leq 2\min\left\{e^{-N\frac{t^2}{c_1}}, c_2 e^{-N\frac{t}{c_3}}\right\}, \tag{15.6}$$

[1] In fact product of two sub-gaussian distributions is sub-exponential.

•

Part V
Statistical Machine Learning

This part of the book presents a mathematical introduction to statistical learning theory. We restricted ourselves to supervised learning since it is closely related to the Bayesian inverse problem. This allows us to draw connection between the two disciplines and to unify the theme of the book.

Chapter 16
Statistical machine learning

16.1 What is machine learning in this book?

Given a set of data $S := \left\{\left(\mathbf{x}^i, y^i\right)\right\}_{i=1}^N$, where $y^i = f^*\left(\mathbf{x}^i\right)$ for some unknown function f^*, *in this book machine learning is the task of learning the function (or the map) f^* from the incomplete information described by the training set $\left\{\left(\mathbf{x}^i, y^i\right)\right\}_{i=1}^N$.* Note that we have assume that $y(\mathbf{x})$ is a scalar-valued function, and the extension to vector-valued function is straightforward.

Let X be a closed and bounded (and hence compact) subset of $\mathbb{R}^k$, and $y \in \mathbb{R}$. Let h be any approximation of f^* and we are interested in measuring the error that we commit in approximating f^* using h, the error is also known as the "risk" in the machine learning literature. To measure the risk we let π be a (Borel) probability measure on the on product space $Z := X \times y$, $\pi(y|\mathbf{x})$ be the conditional probability measure of y given $\mathbf{x}$, and $\pi(\mathbf{x}) := \int_y d\pi(\mathbf{x}, y)$ be the marginal probability measure on X. By Bayes' theorem, we know that $\pi(\mathbf{x}, y) = \pi(y|\mathbf{x}) \times \pi(\mathbf{x})$. One of the popular risk functions is the *least squares error* with respect to the product measure π:

Note the similarity between the risk and the misfit in Chapter 7.

$$\mathscr{R}(h) := \int_Z (h(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x}, y).$$

Cleary, the best $h^*(\mathbf{x}) \in \mathbb{L}^2(X, \pi(\mathbf{x}))$, i.e. $\int_X \|h^*\|_{\mathbb{R}^d}^2 \, d\pi(\mathbf{x}) < \infty$, is the one which minimizes the risk. Chapter 2 shows that the first variation of the risk $\mathscr{R}$ at h^* in any direction $\tilde{h}(\mathbf{x}) \in \mathbb{L}^2(X, \pi(\mathbf{x}))$ must vanish, i.e.,

$$\mathscr{D}\mathscr{R}\left(h^*; \tilde{h}\right) = 2\int_Z (h^*(\mathbf{x}) - y)\,\tilde{h}(\mathbf{x}) \, d\pi(\mathbf{x}, y) = 0, \quad \forall \tilde{h}(\mathbf{x}) \in \mathbb{L}^2(X, \pi(\mathbf{x})),$$

which, after using the Bayes' theorem, becomes

Do you see it? Note that h^* and $\tilde{h}$ are functions of only $\mathbf{x}$.

$$\int_X \left(h^*(\mathbf{x}) - \int_y y \, d\pi(y|\mathbf{x})\right) \tilde{h}(\mathbf{x}) \, d\pi(\mathbf{x}) = 0, \quad \forall \tilde{h}(\mathbf{x}) \in \mathbb{L}^2(X, \pi(\mathbf{x})),$$

See Chapter 2.

which, by Riesz representation theorem, implies that

$$\boxed{h^*(\mathbf{x}) = \int_y y\, d\pi(y|\mathbf{x}).} \tag{16.1}$$

Note that $h^*(\mathbf{x})$ is known as the *regresstion function of* $\pi(\mathbf{x},y)$.

Exercise 16.1. Show that the risk for any function h is given as

$$\mathscr{R}(h) = \int_X (h(\mathbf{x}) - h^*(\mathbf{x}))^2 \, d\pi(\mathbf{x}) + \sigma_\pi^2,$$

where we have defined

$$\sigma_\pi^2 := \int_Z (h^*(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x},y)$$

•

Solution: We have

$$\begin{aligned}
\mathscr{R}(h) &= \int_Z (h(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x},y), \\
&= \int_Z (h(\mathbf{x}) - h^*(\mathbf{x}) + h^*(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x},y) \\
&= \int_X (h(\mathbf{x}) - h^*(\mathbf{x}))^2 \, d\pi(\mathbf{x},y) + 2\int_Z (h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)\, d\pi(\mathbf{x},y) + \sigma_\pi^2
\end{aligned}$$

Since

$$\begin{aligned}
&\int_Z (h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)\, d\pi(\mathbf{x},y) \\
&= \int_X (h(\mathbf{x}) - h^*(\mathbf{x})) \left(h^*(\mathbf{x}) - \underbrace{\int_y y\, d\pi(y|\mathbf{x})}_{h^*} \right) d\pi(\mathbf{x}) = 0
\end{aligned}$$

Note that since σ_π^2 is independent of h, minimizing the risk is the same as minimizing distance between h and h^*. In particular, h^* is the minimizer of the risk.

16.2 Empirical Risk Minimization (ERM)

Assumption 16.1 (i.i.d. assumption of the training set S). The training set $S = \left\{ \left(\mathbf{x}^i, y^i\right) \right\}_{i=1}^N$ are i.i.d. draws from $\pi(\mathbf{x},y)$ on the product space $Z = X \times y$. Note that

$S = \left\{ \left(\mathbf{x}^i, y^i\right) \right\}_{i=1}^N$ can be equivalently considered as a random draw from on Z^N with the N-fold product measure $\pi(\mathbf{x}, y) \times \pi(\mathbf{x}, y) \times \ldots \times \pi(\mathbf{x}, y)$.

Unfortunately the regression function (16.1) is not computable, and hence the risk, since it would require the joint distribution $\pi(\mathbf{x}, y)$, which is unknown. What available about $\pi(\mathbf{x}, y)$ is the limited information given by the training set S. *We assuming that the training set is i.i.d. in the sense of Assumption 16.1*, then by the law of large numbers (see Chapter **??**) we have

$$\mathscr{R}_N(h) := \frac{1}{N} \sum_{i=1}^N \left(h\left(\mathbf{x}^i\right) - y^i\right)^2 \xrightarrow{a.s.} \mathscr{R}(h),$$

where $\mathscr{R}_N(h)$ *is known as the empirical risk*. Thus we have to resort to minimizing the empirical risk, i.e.,

$$\boxed{h_N(\mathbf{x}) \in \arg\min_h \mathscr{R}_N(h).} \tag{16.2}$$

This is the well-known *empirical risk minimization* (ERM) problem, and here h_N is a solution to the ERM problem.

Note that since the training set contain random realizations of $\pi(\mathbf{x}, y)$, $h_N(\mathbf{x})$ is a random function.

16.3 Overfitting and No-Free-Lunch theorem

It turns out that there we can always find a *learner*, i.e. a learning algorithm, $h_N(\mathbf{x})$ that makes that empirical risk vanish, i.e. $\mathscr{R}_N(h_N) = 0$ while having (significant) actual risk $\mathscr{R}(h_N) > 0$. This is known as overfitting: the learner does not *generalize* beyond the data, that is, the training error vanishes but the generalized error can be significant. Since the actual risk is our primary object of interest, overfitting is undersirable. You may hope to find a learner that avoid overfitting. Unfortunately this is impossible as pointed out by the following No-Free-Lunch theorem.

Theorem 16.1 (No-Free-Lunch). *For any $N \geq 1$, and any learner $h_N(\mathbf{x})$ built from the training data $S = \left\{ \left(\mathbf{x}^i, y^i\right) \right\}_{i=1}^N$, where $y \in \{0, 1\}$, there exists a joint distribution $\pi(\mathbf{x}, y)$ such that*

- $\mathscr{R}_N(h_N) = 0$ *and*
- $\mathbb{E}_\pi[\mathscr{R}(h_N)] \geq 1/8$.

Proof. See [26]. □

In other words, there is no universal learner—no learner can succeed in all learning tasks—unless further hypothesis or prior structure is present at the beginning of the learning process. This is, in the language of inverse problem, the ill-posedness nature of the learning problem in which we need to make inference on the high (possibly infinite) dimensional learner with limited information from data. We have

seen in Chapter 1 that this problem can be overcome by using regularization, and restricting the learner having a prior structure is in fact a regularization technique.

In this book, the prior structure takes the form of a class of functions, namely, the *hypothesis space* in which we find a learner $h_N(\mathbf{x})$ that approximates the regression function (16.1) as well as it can while avoiding the overfitting problem. We shall also determine the compromise between the sample size N and the hypothesis space in order to minimize the *generalized error*, i.e. the actual risk.

Chapter 17
Bias-variance tradeoff I

17.1 Hypothesis space $\mathscr{H}$

In this book we choose the hypothesis space $\mathscr{H}$ as a compact subset of the space of continuous functions $\mathbb{C}(X)$ equipped with the standard norm

$$\|f\|_{\mathbb{C}(X)} := \|f\|_\infty := \sup_X |f(\mathbf{x})|, \quad \forall f \in \mathbb{C}(X).$$

While this choice does not cover all the practical problems of interest, it provides rich mathematical structures to study machine learning methods.

17.2 Empirical target function

Recall from Chapter 16 that the best target function is the regression function h^* which is, unfortunately, not computable. In general, h^* does not reside in $\mathscr{H}$, and thus the best that we hope for is to find a *target function* $\hat{h}$ in $\mathscr{H}$ that is closest, e.g. in the least squares sense, to h^*:

$$\hat{h} := \arg\min_{\mathscr{H}} \int_X (f(\mathbf{x}) - h^*(\mathbf{x}))^2 \, d\pi(\mathbf{x}) \tag{17.1}$$

Exercise 17.1. Show that $\hat{h}$ is also closest to $\mathbf{y}$, that is, $\hat{h}$ is a minimizer of the following least squares problem

$$\min_{\mathscr{H}} \int_Z (f(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x}, \mathbf{y})$$

•

Solution: This is clear from the Exercise 16.1. Exercise 17.1 implies

$$\hat{h} := \arg\min_{\mathscr{H}} \mathscr{R}(f). \tag{17.2}$$

Compared to (17.1), for which we have no access to h^*, we partially know y in (17.2) through the training set $S := \left\{\left(\mathbf{x}^i, y^i\right)\right\}_{i=1}^N$. Therefore, instead of seeking $\hat{h}$, which is not possible, we resort to a minimizer, the *empirical target function*, $\hat{h}_N$ of the empirical risk minimization problem defined as

$$\boxed{\hat{h}_N := \arg\min_{\mathscr{H}} \mathscr{R}_N(f) = \frac{1}{N}\sum_i^N \left(f\left(\mathbf{x}^i\right) - y^i\right)^2.} \tag{17.3}$$

The question that needs to be addressed is if the target and the empirical target functions exist. To that end, by Theorem 17.1, we need to show that the risk function and the empircal risk function are continuous.

Assumption 17.1 (M-Boundedness of the misfit on $\mathscr{H}$). For any $h \in \mathscr{H}$ and almost everywhere (a.e. or aka a.s.) in X, there holds

$$|h(\mathbf{x}) - y| \leq M. \tag{17.4}$$

Proposition 17.1. *Suppose $\mathscr{H}$ is bounded in the sense of Assumption 17.1. Then $\mathscr{R}, \mathscr{R}_N : \mathscr{H} \to \mathbb{R}$ are Lipschitz continuous.*

Proof. The proof is straightforward using (17.4). Indeed, we have

$$|\mathscr{R}(h_1) - \mathscr{R}(h_2)| = \left|\int_Z (h_1 + h_2 - 2y)(h_1 - h_2)\, d\pi(\mathbf{x}, \mathbf{y})\right| \leq 2M \|h_1 - h2\|_\infty.$$

Similarly,

$$\begin{aligned}|\mathscr{R}_N(h_1) - \mathscr{R}_N(h_2)| &= \left|\frac{1}{N}\sum_i^N \left(h_1\left(\mathbf{x}^i\right) + h_2\left(\mathbf{x}^i\right) - 2y^i\right)\left(h_1\left(\mathbf{x}^i\right) - h_2\left(\mathbf{x}^i\right)\right)\right| \\ &\leq \frac{1}{N}\sum_i^N \left|h_1\left(\mathbf{x}^i\right) + h_2\left(\mathbf{x}^i\right) - 2y^i\right| \|h_1 - h_2\|_\infty \leq 2M \|h_1 - h_2\|_\infty.\end{aligned}$$

□

Let us define the *sampling error* in computing the risk for any h as

$$\mathscr{E}(h) := \mathscr{R}(h) - \mathscr{R}_N(h). \tag{17.5}$$

As will be shown in the next section this sampling error plays the key role in estimating the sample/estimation error.

Exercise 17.2 (Continuity of Sampling Error). Show that $\mathscr{E} : \mathscr{H} \to \mathbb{R}$ is Lipschitz continuous. •

Solution: Using Proposition 17.1 we have

$$|\mathscr{E}(h_1) - \mathscr{E}(h_2)| \leq |\mathscr{R}(h_1) - \mathscr{R}(h_2)| + |\mathscr{R}_N(h_1) - \mathscr{R}_N(h_2)| \leq 4M \|h_1 - h_2\|_\infty .$$

Corollary 17.1 (Existence of $\hat{h}$ and $\hat{h}_N$). *Suppose $\mathscr{H}$ is bounded in the sense of Assumption 17.1. Then $\hat{h}$ and $\hat{h}_N$ exist.*

Proof. The assertion is obvious due to Proposition 17.1 and Theorem 17.1. □

17.3 Bias-Variance Tradeoff

Suppose the empircal target $\hat{h}_N$ is already computed/estimated (to be discussed in details later), the question of interest is how to estimate the actual risk $\mathscr{R}\left(\hat{h}_N\right)$. This section decomposes this actual risk into two parts: the *sample or estimation* error and the approximation error. The compromise between these two errors is known as the *bias-variance tradeoff*. We begin with the following decomposition

$$\mathscr{R}\left(\hat{h}_N\right) = \underbrace{\mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}\left(\hat{h}\right)}_{\mathscr{S}\left(\hat{h}_N\right):=} + \underbrace{\mathscr{R}\left(\hat{h}\right)}_{\mathscr{B}\left(\hat{h}\right):=} .$$

Let us first consider $\mathscr{S}\left(\hat{h}_N\right)$. We have

$$\begin{aligned} \mathscr{S}\left(\hat{h}_N\right) &= \underbrace{\mathscr{R}_N\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}\right)}_{\leq 0} + \mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}_N\right) + \mathscr{R}_N\left(\hat{h}\right) - \mathscr{R}\left(\hat{h}\right) \\ &\leq \left|\mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}_N\right)\right| + \left|\mathscr{R}_N\left(\hat{h}\right) - \mathscr{R}\left(\hat{h}\right)\right|, \end{aligned} \tag{17.6}$$

where the negativeness of the first difference is due to the fact that $\hat{h}_N$ is a minimizer of $\mathscr{R}_N$. Note that the last two terms are the sampling errors incurred by approximating $\mathscr{R}\left(\hat{h}_N\right)$ and $\mathscr{R}\left(\hat{h}\right)$ using the training set S. Since $\mathscr{S}\left(\hat{h}_N\right)$ is bounded by the sampling errors, it is conventionally named as the *sample/estimation* error or the *variance*.

Now due to Exercise 16.1 we can rewrite $\mathscr{B}\left(\hat{h}\right)$ as

$$\mathscr{B}\left(\hat{h}\right) = \int_X \left(\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right)^2 d\pi(\mathbf{x}) + \sigma_\pi^2, \tag{17.7}$$

which is independent of the training set. It actually depends only on the distance between $\hat{h}$ and the regression function h^*. That is, it is completely determined by the approximation capability of the hypothesis space $\mathscr{H}$. Thus, $\mathscr{B}\left(\hat{h}\right)$ is often called the *bias*.

We are in the position to discuss *bias-variance tradeoff*. For a fixed $\mathscr{H}$, the sample error clearly decreases as the sample size N increases. Now if we fix the sample size N, enlarging the hypothesis space $\mathscr{H}$ clearly reduces the bias but increase the sample error in general (will be shown in the next chapters). A popular tradeoff is to increase the dimension of the hypothesis space $\mathscr{H}$ as the sample size N increases.

The question is how fast we should enlarge $\mathscr{H}$. To answer this question in a sensible and rigorous way, we need to first estimate both sample error and the bias, and then balance out these errors. This is the task of the following chapters.

17.4 Appendix

Theorem 17.1. *Let K be a compact set and $L : K \to \mathbb{R}$ be a continuous function. Then L is bounded. In particular, there exist $f, g \in K$ at which L attains its minimum and maximum.*

Chapter 18
Hypothesis space I

Recall from Chapter 17 we have assumed that the hypothesis space $\mathscr{H}$ is compact subset of $\mathbb{C}(X)$. In this chapter we show that this assumption is valid when the hypothesis space is taken as a bounded subset of the reproducing kernel Hilbert space of Mercer kernel.

18.1 Reproducing kernel Hilbert spaces (RKHS)

In this book we are interested in hypothesis space associated with RKHS. Assume X is a metric space and let $\mathsf{K} : X \times X \to \mathbb{R}$ be *symmetric positive semidefinite* in the sense $\mathsf{K}(\mathbf{x},\mathbf{x}') = \mathsf{K}(\mathbf{x}',\mathbf{x})$ and any $n \times n$ *Gramian* matrix whose ij-element is $\mathsf{K}(\mathbf{x}_i,\mathbf{x}_j)$ is semipositive definite for all n and $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. We say K is a *Mercer* kernel if it is continuous, symmetric, and positive semidefinite. Clearly the positive semidefiniteness implies $\mathsf{K}(\mathbf{x},\mathbf{x}) \geq 0$ for any $\mathbf{x} \in X$, and this allows us to define

$$C_{\mathsf{K}} := \sup_{\mathbf{x}\in X} \sqrt{\mathsf{K}(\mathbf{x},\mathbf{x})}.$$

Exercise 18.1. Show that

$$C_{\mathsf{K}} = \sup_{\mathbf{x},\mathbf{x}'\in X} \sqrt{|\mathsf{K}(\mathbf{x},\mathbf{x}')|}$$

•

Solution: From the positive semidefinite of the 2×2 Gramian matrix using two points $\mathbf{x},\mathbf{x}'$ we have

$$\left(\mathsf{K}(\mathbf{x},\mathbf{x}')\right)^2 \leq \mathsf{K}(\mathbf{x},\mathbf{x})\,\mathsf{K}(\mathbf{x}',\mathbf{x}'),$$

and hence

$$\sup_{\mathbf{x},\mathbf{x}'\in X}\sqrt{|\mathbf{K}(\mathbf{x},\mathbf{x}')|}\leq\sup_{\mathbf{x}\in X}\sqrt{\mathbf{K}(\mathbf{x},\mathbf{x})}\sup_{\mathbf{x}'\in X}\sqrt{\mathbf{K}(\mathbf{x}',\mathbf{x}')}=C_{\mathbf{K}}^{2}.$$

Since the equality is attainable, this ends the proof.

For any $\mathbf{x}\in X$, by $\mathbf{K}_{\mathbf{x}}$ we denote the function $\mathbf{K}_{\mathbf{x}}:X\ni\mathbf{x}'\mapsto\mathbf{K}_{\mathbf{x}}(\mathbf{x}'):=\mathbf{K}(\mathbf{x},\mathbf{x}')\in\mathbb{R}$. We form the space $\mathring{\mathscr{H}}_{\mathbf{K}}$ as

$$\mathring{\mathscr{H}}_{\mathbf{K}}:=\operatorname{span}\{\mathbf{K}_{\mathbf{x}}:\mathbf{x}\in X\},$$

and equip it with the following inner product: for $f=\sum_{i=1}^{I}f_i\mathbf{K}_{\mathbf{x}^i}\in\mathring{\mathscr{H}}_{\mathbf{K}}$ and $g=\sum_{j=1}^{J}g_j\mathbf{K}_{\mathbf{t}^j}\in\mathring{\mathscr{H}}_{\mathbf{K}}$,

$$(f,g)_{\mathscr{H}_{\mathbf{K}}}=\sum_{i,j}^{I,J}f_ig_j\mathbf{K}(\mathbf{x}^i,\mathbf{t}^j). \tag{18.1}$$

Theorem 18.1 (Reproducing Kernel Hilbert Space). *Let $\mathscr{H}_{\mathbf{K}}$ be a Hilbert space with with the following properties:*

1. $\mathbf{K}_{\mathbf{x}}\in\mathscr{H}_{\mathbf{K}}$ *for any* $\mathbf{x}\in X$.
2. $\mathring{\mathscr{H}}_{\mathbf{K}}$ *is dense in* $\mathscr{H}_{\mathbf{K}}$.
3. **Reproducing property:** *for any* $f\in\mathscr{H}_{\mathbf{K}}$ *and* $\mathbf{x}\in X$, *we have* $f(\mathbf{x})=(\mathbf{K}_{\mathbf{x}},f)_{\mathscr{H}_{\mathbf{K}}}$.

Then $\mathscr{H}_{\mathbf{K}}$ is unique. Moreover, $\mathscr{H}_{\mathbf{K}}\subset\mathbb{C}(X)$ and the inclusion $\mathrm{i}_{\mathbf{K}}:\mathscr{H}_{\mathbf{K}}\hookrightarrow\mathbb{C}(X)$ is bounded.

Proof. Note sure we need this result, so let's not type the proof for now. □

18.2 Hypothesis space associated with RKHS

Proposition 18.1. *Suppose $\mathbf{K}$ be a Mercer kernel on a compact metric space X, and $\mathscr{H}_{\mathbf{K}}$ is the associated RKHS. For any $R>0$, the ball $B(R):=\left\{f\in\mathscr{H}_{\mathbf{K}}:\|f\|_{\mathscr{H}_{\mathbf{K}}}\leq R\right\}$ is a closed subset of $\mathbb{C}(X)$.*

Proof. From Theorem 18.1 it is sufficient to show that $B(R)$ is closed in $\mathbb{C}(X)$. To that end, suppose $\{f_n\}_{n\in\mathbb{N}}\in B(R)$ converges to $f^*\in\mathbb{C}(X)$, i.e.,

$$\lim_n f_n(\mathbf{x})=f^*(\mathbf{x}),$$

and we need to show that $f^*\in B(R)$. Since, by Lemma 18.1, $B(R)$ is weakly compact, there exists a subsequence $\{f_{n_k}\}_{k\in\mathbb{N}}$ converging to $\hat{f}\in B(R)$, i.e.,

$$\lim_k\left(f_{n_k},g\right)_{\mathscr{H}_{\mathbf{K}}}=\left(\hat{f},g\right)_{\mathscr{H}_{\mathbf{K}}},\quad\forall g\in\mathscr{H}_{\mathbf{K}}.$$

Now taking $g = \mathbf{K}_{\mathbf{x}}$ and using the reproducing property we have

$$f^*(\mathbf{x}) = \lim_k f_{n_k}(\mathbf{x}) = \lim_k \left(f_{n_k}, \mathbf{K}_{\mathbf{x}}\right)_{\mathscr{H}_{\mathbf{K}}} = \left(\hat{f}, \mathbf{K}_{\mathbf{x}}\right)_{\mathscr{H}_{\mathbf{K}}} = \hat{f}(\mathbf{x}), \quad \forall \mathbf{x} \in X.$$

Since both f^* and $\hat{f}$ are continuous, they must be identical and this ends the proof. □

Theorem 18.2. *Suppose* $\mathbf{K}$ *be a Mercer kernel on a compact metric space* X*, and* $\mathscr{H}_{\mathbf{K}}$ *is the associated RKHS. The inclusion* $\mathrm{i}_{\mathbf{K}} : \mathscr{H}_{\mathbf{K}} \hookrightarrow \mathbb{C}(X)$ *is compact. In other words, the set* $\mathrm{i}_{\mathbf{K}}(B(R))$ *is compact for any* $R > 0$*.*

Proof. Theorem 18.1 and Proposition 18.1 show that $\mathrm{i}_{\mathbf{K}}(B(R))$ is closed and bounded. By the Arzelá-Ascoli theorem, what remains is to show the equicontinuity. We have

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{t})| &= \left|(f, \mathbf{K}_{\mathbf{x}} - \mathbf{K}_{\mathbf{t}})_{\mathscr{H}_{\mathbf{K}}}\right| \leq \|f\|_{\mathscr{H}_{\mathbf{K}}} \|\mathbf{K}_{\mathbf{x}} - \mathbf{K}_{\mathbf{t}}\|_{\mathscr{H}_{\mathbf{K}}} \\ &= R\sqrt{(\mathbf{K}_{\mathbf{x}} - \mathbf{K}_{\mathbf{t}}, \mathbf{K}_{\mathbf{x}} - \mathbf{K}_{\mathbf{t}})_{\mathscr{H}_{\mathbf{K}}}} = R\sqrt{(\mathbf{K}_{\mathbf{x}}(\mathbf{x}) - \mathbf{K}_{\mathbf{x}}(\mathbf{t}) + \mathbf{K}_{\mathbf{t}}(\mathbf{t}) - \mathbf{K}_{\mathbf{t}}(\mathbf{x}))_{\mathscr{H}_{\mathbf{K}}}} \end{aligned}$$

Now since $\mathbf{K}$ is continuous on the compact set $X \times X$, it is uniformly continuous on $X \times X$, i.e., for all $\mathbf{x}, \mathbf{t}, \mathbf{t}' \in X$ such that $\left\|\mathbf{t} - \mathbf{t}'\right\|_X \leq \delta$ (δ does not depend on $\mathbf{x}, \mathbf{t}, \mathbf{t}'$) implies

$$\left|\mathbf{K}_{\mathbf{x}}(\mathbf{t}) - \mathbf{K}_{\mathbf{x}}\left(\mathbf{t}'\right)\right| \leq \varepsilon.$$

We thus have

$$|f(\mathbf{x}) - f(\mathbf{t})| \leq R\sqrt{2\varepsilon}, \quad \forall \mathbf{t}, \mathbf{x} : \|\mathbf{x} - \mathbf{t}\| \leq \delta, \quad \forall f \in \mathrm{i}_{\mathbf{K}}(B(R)),$$

and this concludes the proof. □

In this book we generally take the hypothesis space $\mathscr{H}$ *as a compact subset of* $\mathbb{C}(X)$*. As shown in Theorem 18.2, this can be justified by choosing* $\mathscr{H}$ *as a closed ball in the RKHS associated with the kernel under consideration.*

18.3 Appendix

Lemma 18.1 (Weak compactness of closed balls in Hilbert space). *If* B *be a closed ball in a Hilbert space* $\mathscr{H}$*, it is weakly compact. In other words, every sequence* $\{f_n\}_{n \in \mathbb{N}} \subset B$ *has a weakly convergence subsequence* $\left\{f_{n_k}\right\}_{k \in \mathbb{N}}$*. That is, there exists some* $f^* \in B$ *such that*

$$\lim_{k \to \infty} \left(f_{n_k}, g\right)_{\mathscr{H}} = (f^*, g)_{\mathscr{H}}, \quad \forall g \in \mathscr{H}.$$

Definition 18.1 (Equicontinuity). A subset K of $\mathbb{C}(X)$ is equicontinuous at $\mathbf{x} \in X$ if for any $\varepsilon > 0$ there exists a neighborhood B of x such that $\forall \mathbf{t} \in X$ and $\forall f \in K$

we have $\|f(\mathbf{x}) - f(\mathbf{t})\|_\infty < \varepsilon$. K is equicontinuous if it is equicontinuous at every $\mathbf{x} \in X$.

Theorem 18.3 (Arzelá-Ascoli theorem). *Let X be compact. $K \subset \mathbb{C}(X)$ is compact if and only if K is closed, bounded, and equicontinuous.*

Chapter 19
Hypothesis space II

In this chapter we are going to construct the Reproducing kernel Hilbert space (RKHS) using the eigen-pairs of an integral operator defined by the Mercer kernel under consideration. We will show that this operator is a compact and self-adjoint and hence admits a spectral decomposition by the Hilbert-Schmidt theorem 1.1. This approach is more appealing as it follows the same direction of constructing Gaussian measure in Chapter **??**, and hence making the presentation of the book more coherent.

19.1 Kernel-based integral operators

Let us define the following integral operator

$$(L_{\mathbf{K}} f)(\mathbf{x}) := \int_X \mathbf{K}(\mathbf{x},\mathbf{t}) f(\mathbf{t}) \, d\pi(\mathbf{t}),$$

where we assume π is a probability measure on X though all the results, up to a constant, also hold for a general Borel measure. We observe

$$|(L_{\mathbf{K}} f)(\mathbf{x})| \leq \|\mathbf{K}_{\mathbf{x}}\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} \|f\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} \leq C_{\mathbf{K}} \|f\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))}, \tag{19.1}$$

which implies $L_{\mathbf{K}}$ as a map from $\mathbb{L}^2(X,\pi(\mathbf{x}))$ to $\mathbb{C}(X)$ is (Lipschitz) continuous. Since the inclusion $\mathbb{C}(X) \hookrightarrow \mathbb{L}^2(X,\pi(\mathbf{x}))$ is continuous, we see that $L_{\mathbf{K}}$ as a map from $\mathbb{L}^2(X,\pi(\mathbf{x}))$ into $\mathbb{L}^2(X,\pi(\mathbf{x}))$ is continuous and its operator norm is bounded as $\|L_{\mathbf{K}}\| \leq C_{\mathbf{K}}^2$.

Proposition 19.1. *$L_{\mathbf{K}} : \mathbb{L}^2(X,\pi(\mathbf{x})) \to \mathbb{C}(X)$ is a compact operator. If, in addition, $\mathbf{K}$ is a Mercer kernel, then $L_{\mathbf{K}}$ is self-adjoint positive semidefinite compact operator.*

Proof. Let B be a bounded set in $\mathbb{L}^2(X,\pi(\mathbf{x}))$ such that $\|f\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} \leq M, \forall f \in B$. From (19.1) we see that $L_{\mathbf{K}}(B)$ is uniformly bounded. A similar arguement as in (19.1) shows that

$$\left|\left(L_{\mathbf{K}}f\right)(\mathbf{x})-\left(L_{\mathbf{K}}f\right)\left(\mathbf{x}'\right)\right| \leq \sup_{\mathbf{t}\in X}\left|\mathbf{K}_{\mathbf{x}}(\mathbf{t})-\mathbf{K}_{\mathbf{x}'}(\mathbf{t})\right| \|f\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} \leq 2C_{\mathbf{K}}M,$$

which implies that $L_{\mathbf{K}}(B)$ is equicontinuous. By the Arzelá-Ascoli theorem 18.3, the closure of $L_{\mathbf{K}}(B)$ is compact in $\mathbb{C}(X)$. In other words, $L_{\mathbf{K}}(B)$ is relatively compact and by definition 19.2, $L_{\mathbf{K}}$ is a compact operator.

The self-adjointness is clear by the Fubini theorem. The positive semidefiniteness of $L_{\mathbf{K}}$ is a direct consquence of the positive semidefiniteness of $\mathbf{K}$. Indeed, since X is a compact subset of $\mathbb{R}^k$, without loss of generality, we can subdivide X into n subsets with equal volumes and with "centroids" $\mathbf{x}_1,\ldots,\mathbf{x}_n$. We then have

$$\begin{aligned}(f,L_{\mathbf{K}}f)_{\mathbb{L}^2(X,\pi(\mathbf{x}))} &= \int_{X\times X}\mathbf{K}(\mathbf{x},\mathbf{t})\,f(\mathbf{t})\,f(\mathbf{x})\,d\pi(\mathbf{t})\,d\pi(\mathbf{x})\\ &= \lim_{n\to\infty}\frac{\mathscr{V}(X)^2}{n^2}\sum_{i,j=1}^{n}\mathbf{K}(\mathbf{x}_i,\mathbf{t}_j)\,f(\mathbf{t}_j)\,f(\mathbf{x}_i)\geq 0.\end{aligned}$$

□

By the Hilbert-Schmidt Theorem 1.1, $L_{\mathbf{K}}$ admits a spectral decomposition with eigenpairs $(\lambda_i,\varphi_i)_{i=1}^{\infty}$, i.e.,

$$L_{\mathbf{K}}f=\sum_{i=1}^{\infty}a_i\lambda_i\varphi_i,$$

for any $f=\sum_{i=1}^{\infty}a_i\varphi_i\in\mathbb{L}^2(X,\pi(\mathbf{x}))$. Moreover the Mercer Theorem 19.3 holds.

Exercise 19.1. Show that

$$\sum_{i=1}^{\infty}\lambda_i=\int_X\mathbf{K}(\mathbf{x},\mathbf{x})\,d\pi(\mathbf{x})\leq\mathscr{V}(X)C_{\mathbf{K}}.$$

Deduce that $\lambda_k\leq\dfrac{\mathscr{V}(X)C_{\mathbf{K}}}{k}$. •

Solution: From Mercer theorem we have

$$\mathbf{K}(\mathbf{x},\mathbf{x})=\sum_{i=1}^{\infty}\lambda_i\varphi_i(\mathbf{x})^2,$$

which yields the desired result after integrating both sides and using the orthonormality of the eigenfunctions $\varphi_i(\mathbf{x})$. The second assertion is clear due to

$$i\lambda_i\leq\sum_{k=1}^{i}\lambda_k\leq\sum_{k=1}^{\infty}\lambda_k,$$

where we have ordered the eigenvalues such that λ_1 is the largest.

Without loss of generality, assume that $\lambda_i>0,\forall i$. Let us now define the following space

$$\widehat{\mathscr{H}_{\mathbf{K}}} := \left\{ f \in \mathbb{L}^2(X, \pi(\mathbf{x})) : f = \sum_{i=1}^{\infty} a_i \varphi_i \text{ with } \left(\frac{a_i}{\sqrt{\lambda_i}} \right) \in \ell^2 \right\},$$

and equip $\widehat{\mathscr{H}_{\mathbf{K}}}$ with the following inner product

$$(f,g)_{\mathbf{K}} := \sum_{i=1}^{\infty} \frac{a_i b_i}{\lambda_i} \tag{19.2}$$

for any $f = \sum_{i=1}^{\infty} a_i \varphi_i$ and $g = \sum_{i=1}^{\infty} b_i \varphi_i$ in $\widehat{\mathscr{H}_{\mathbf{K}}}$. Exercise 19.2 shows that $\widehat{\mathscr{H}_{\mathbf{K}}}$ is a Hilbert space.

Exercise 19.2. Show that $\widehat{\mathscr{H}_{\mathbf{K}}}$ with aforementioned inner product is a Hilbert space.

•

Since $L_{\mathbf{K}}^{\frac{1}{2}} : \mathbb{L}^2(X, \pi(\mathbf{x})) \ni f = \sum_{i=1}^{\infty} a_i \varphi_i \mapsto g = L_{\mathbf{K}}^{\frac{1}{2}} f := \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} \varphi_i \in \widehat{\mathscr{H}_{\mathbf{K}}}$ is an isometry, the space $\widehat{\mathscr{H}_{\mathbf{K}}}$ allows us to define a square root of $L_{\mathbf{K}}$.

Compare $L_{\mathbf{K}}^{\frac{1}{2}}$ and the Cameron-Martin map of a Gaussian measure.

Exercise 19.3. Show that the following hold true:

i) $L_{\mathbf{K}}^{\frac{1}{2}} : \mathbb{L}^2(X, \pi(\mathbf{x})) \ni f = \sum_{i=1}^{\infty} a_i \varphi_i \mapsto g = L_{\mathbf{K}}^{\frac{1}{2}} f := \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} \varphi_i \in \widehat{\mathscr{H}_{\mathbf{K}}}$ is an isometry.

ii) $g \in \mathbb{L}^2(X, \pi(\mathbf{x}))$.

iii) $L_{\mathbf{K}} = L_{\mathbf{K}}^{\frac{1}{2}} \cdot L_{\mathbf{K}}^{\frac{1}{2}}$.

•

Solution:

i) Clearly

$$\|f\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} = \sum_{i=1}^{\infty} a_i^2,$$

due to the orthonormality of the eigenfunctions φ_i (Hilbert-Schmidt Theorem 1.1).

On the other hand, using the norm induced from the inner product (19.2) we have

$$\|g\|_{\mathbf{K}} := \sum_{i=1}^{\infty} a_i^2.$$

ii) $\|g\|_{\mathbb{L}^2(X,\pi(\mathbf{x}))} = \sum_{i=1}^{\infty} \lambda_i a_i^2 \leq \lambda_1 \sum_{i=1}^{\infty} a_i^2 < \infty$.

iii) $L_{\mathbf{K}}^{\frac{1}{2}} g = \sum_{i=1}^{\infty} a_i \lambda_i \varphi_i = L_{\mathbf{K}} f$, where, in the second equality, we have used the fact that (λ_i, φ_i) are eigenpairs of $L_{\mathbf{K}}$.

Now comes the main result of this chapter.

Theorem 19.1. *$\widehat{\mathscr{H}_{\mathsf{K}}}$ and $\mathscr{H}_{\mathsf{K}}$ are identical.*

Proof. We just need to show that $\widehat{\mathscr{H}_{\mathsf{K}}}$ has three properties in Theorem 18.1.

- By definition of $\mathsf{K}_{\mathbf{x}}$, Mercer theorem, and the induced norm from (19.2) we have

$$\|\mathsf{K}_{\mathbf{x}}(\mathbf{t})\|_{\mathsf{K}} = \|\mathsf{K}(\mathbf{x},\mathbf{t})\|_{\mathsf{K}} = \left\|\sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x})\,\varphi_i(\mathbf{t})\right\|_{\mathsf{K}} = \sqrt{\sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x})^2}$$
$$= \sqrt{\mathsf{K}(\mathbf{x},\mathbf{x})} \leq C_{\mathsf{K}} < \infty,$$

  which shows that $\mathsf{K}_{\mathbf{x}} \in \widehat{\mathscr{H}_{\mathsf{K}}}$ for any $\mathbf{x} \in X$.
- To see that $\widehat{\mathscr{H}_{\mathsf{K}}}$ has the reprocducing property we take $f(\mathbf{t}) = \sum_{i=1}^{\infty} a_i \varphi_i(\mathbf{t}) \in \widehat{\mathscr{H}_{\mathsf{K}}}$ and show that $(f, \mathsf{K}\mathbf{x})_{\mathsf{K}} = f(\mathbf{x})$ but this is straightforward by Mercer theorem and the definition of the inner product in $\widehat{\mathscr{H}_{\mathsf{K}}}$. Indeed,

$$(f, \mathsf{K}_{\mathbf{x}})_{\mathsf{K}} = \sum_{i=1}^{\infty} \frac{a_i \lambda_i \varphi_i(\mathbf{x})}{\lambda_i} = f(\mathbf{x}).$$

- To show that $\mathring{\mathscr{H}}_{\mathsf{K}}$ is dense in $\mathscr{H}_{\mathsf{K}}$, it is sufficient to prove that, for any $f \in \widehat{\mathscr{H}_{\mathsf{K}}}$, $(f, \mathsf{K}_{\mathbf{x}}) = 0$ for any $\mathbf{x} \in X$ implies $f = 0$. But this is clear by the reproducing property.

□

We conclude this chapter with the definition and properties of a *feature* map.

Theorem 19.2. *Show the the feature map defined by*

$$\Phi : X \ni \mathbf{x} \mapsto \Phi(\mathbf{x}) := \left\{\sqrt{\lambda_k}\varphi_k(\mathbf{x})\right\}_{k\in\mathbb{N}} \in \ell^2$$

is well-defined, and continuous, and satisfies

$$\mathsf{K}(\mathbf{x},\mathbf{t}) = (\Phi(\mathbf{x}), \Phi(\mathbf{t}))_{\ell^2}.$$

Proof. It is obvious by Mercer theorem:

$$(\Phi(\mathbf{x}), \Phi(\mathbf{t}))_{\ell^2} = \sum_{k=1}^{\infty} \lambda_k \varphi_k(\mathbf{x})\,\varphi_k(\mathbf{t}) = \mathsf{K}(\mathbf{x},\mathbf{t}),$$

and

$$\|\Phi(\mathbf{x})\|_{\ell}^2 = \sqrt{\sum_{k=1}^{\infty} \lambda_k \varphi_k(\mathbf{x})^2} = \sqrt{\mathsf{K}(\mathbf{x},\mathbf{x})} \leq C_{\mathsf{K}} < \infty.$$

□

19.2 Appendix

Definition 19.1 (Relatively compact). A subset B of a metric space X is relatively compact if its closure is compact in X.

Definition 19.2 (Compact operator). $L : X \to y$ is a compact operator if for any bounded set $B \subset X$, $L(B)$ is relatively compact in y.

Theorem 19.3 (Mercer theorem). *Suppose X is a compact subset of $\mathbb{R}^k$ and K is a Mercer kernel. Let $(\lambda_i, \varphi_i)_{i=1}^{\infty}$ be the eigenpairs of L_{K}. For all $\mathbf{x}, \mathbf{t} \in X$, there holds*

$$\mathsf{K}(\mathbf{x}, \mathbf{t}) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{t}),$$

where the convergence is absolute and uniform on pointwise $X \times X$.

Chapter 20
Sample Error (Variance)

Recall that the sample error is bounded by

$$\mathscr{S}\left(\hat{h}_N\right) := \mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}\left(\hat{h}\right) \leq \left|\mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}_N\right)\right| + \left|\mathscr{R}_N\left(\hat{h}\right) - \mathscr{R}\left(\hat{h}\right)\right|. \quad (20.1)$$

The two terms on the right hand side of (20.1) are the sampling errors for $\mathscr{R}\left(\hat{h}_N\right)$ and $\mathscr{R}\left(\hat{h}\right)$ using Monte Carlo. Chapter **??** tells us that these errors for finite sample size[1] N can be bounded probalistically.

20.1 Sampling error for a function in $\mathscr{H}$

Lemma 20.1. *Assume that $\mathscr{H}$ is M-bounded in the sense of Assumption 17.1. For any $h \in \mathscr{H}$, there holds*

$$\mathbb{P}\left[\left|\mathscr{R}_N(h) - \mathscr{R}(h)\right| > t\right] \leq 2e^{-2N\frac{t^2}{M^4}}, \quad \forall t \geq 0.$$

Proof. The proof is obvious using the M-boundedness of $\mathscr{H}$, the i.i.d. assumption 16.1 on the training set S, and the result of Exercise 14.3. □

20.2 Sampling error for finite $\mathscr{H}$

We now extend Section 20.1 to hypothesis space $\mathscr{H}$ containing finite number of functions $h_1, \ldots, h_{\mathscr{N}}$. The task at hand is to probabilistly bound the worst error among $h_1, \ldots, h_{\mathscr{N}}$.

Lemma 20.2. *Let $\mathscr{H} = \{h_1, \ldots, h_{\mathscr{N}}\}$, H is M-bounded, and the training set S is i.i.d. in the sense of Assumption 16.1. There holds:*

[1] Again, this is the beauty of cencontration of measures for non-asymptotic theory.

$$\mathbb{P}\left[\left|\mathscr{R}_N(h)-\mathscr{R}(h)\right|>\varepsilon\right]\leq 2\mathscr{N}e^{-2N\frac{\varepsilon^2}{M^4}},\quad \forall\varepsilon\geq 0.$$

Proof. We start with the following observation

$$\sup_{h\in\mathscr{H}}\left|\mathscr{R}_N(h)-\mathscr{R}(h)\right|>\varepsilon \iff \exists i\leq m:\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon,$$

and hence the following identity of on the equivalent events

$$\begin{aligned}\left\{\sup_{h\in\mathscr{H}}\left|\mathscr{R}_N(h)-\mathscr{R}(h)\right|>\varepsilon\right\}&=\left\{\exists i\leq m:\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon\right\}\\&=\cup_{i=1}^{\mathscr{N}}\left\{\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon\right\}.\end{aligned}$$

This leads to

$$\begin{aligned}&\mathbb{P}\left[\sup_{h\in\mathscr{H}}\left|\mathscr{R}_N(h)-\mathscr{R}(h)\right|>\varepsilon\right]=\mathbb{P}\left[\exists i\leq m:\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon\right]\\&=\mathbb{P}\left[\cup_{i=1}^{\mathscr{N}}\left\{\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon\right\}\right]\leq\sum_{i=1}^{\mathscr{N}}\mathbb{P}\left[\left|\mathscr{R}_N(h_i)-\mathscr{R}(h_i)\right|>\varepsilon\right]\leq 2\mathscr{N}e^{-2N\frac{\varepsilon^2}{M^4}},\end{aligned}$$

where we have used the union bound in Lemma 12.3 in the second last inequality, and the i.i.d. nature of S together with Lemma 20.1 in the last inequality. □

20.3 Sampling error when $\mathscr{H}$ is a ball with radius R

Let $\mathscr{H}:=\{h\in\mathbb{C}(X):\|h\|_\infty\leq R\}$. Due to the continuity of the sampling error in Exercise 17.2, i.e.,

$$\left|\mathscr{E}(h_1)-\mathscr{E}(h_2)\right|\leq 4M\left\|h_1-h_2\right\|_\infty,\quad \forall h_1,h_2\in\mathscr{H},$$

we have

$$\mathscr{E}(h_c)\geq\mathscr{E}(h)-4MR,$$

where h_c is the center of the ball and h is an arbitrary function inside the ball. Thus

$$\mathscr{E}(h_c)\geq\sup_{h\in\mathscr{H}}\mathscr{E}(h)-4MR,$$

which, in turn, shows that if the following implication holds true:

$$\sup_{h\in\mathscr{H}}\left|\mathscr{E}(h)\right|\geq\varepsilon\Rightarrow\left|\mathscr{E}(h_c)\right|\geq\varepsilon-4MR.$$

We conclude that

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\right\}\right]\leq\mathbb{P}[\{|\mathscr{E}(h_c)|\geq\varepsilon-4MR\}]\leq 2e^{-2N\frac{(\varepsilon-4MR)^2}{M^4}},$$

where we have used Lemma 20.1 in the last inequality. We thus have proved the following result.

Lemma 20.3. *Assume $\mathscr{H}:=\{h\in\mathbb{C}(X):\|h\|_\infty\leq R\}$, and $\mathscr{H}$ is M-bounded in the sense of Assumption 17.1. Then the tail of the worst sampling error decays exponentially, i.e.,*

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\right\}\right]\leq 2e^{-2N\frac{(\varepsilon-4MR)^2}{M^4}}.$$

20.4 Sampling error when $\mathscr{H}$ is a union of $\mathscr{N}$ balls

Without loss of generality we assume that all the balls have the same radius of $\varepsilon/8M$, i.e., $\mathscr{H}=\cup_{i=1}^{\mathscr{N}}B_{h_i}\left(\frac{\varepsilon}{8M}\right)$.

Lemma 20.4. *Assume $\mathscr{H}$ is M-bounded in the sense of Assumption 17.1. Then the tail of the worst sampling error decays exponentially, i.e.,*

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\right\}\right]\leq 2\mathscr{N}e^{-N\frac{\varepsilon^2}{2M^4}}.$$

Proof. We proceed with the following observation

$$\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\iff\exists i\leq\mathscr{N}:\sup_{h\in B_{h_i}\left(\frac{\varepsilon}{8M}\right)}|\mathscr{E}(h)|\geq\varepsilon,$$

which, by the union bound in Lemma 12.3, implies

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\right\}\right]\leq\sum_{i=1}^{\mathscr{N}}\mathbb{P}\left[\left\{\sup_{h\in B_{h_i}\left(\frac{\varepsilon}{8M}\right)}|\mathscr{E}(h)|\geq\varepsilon\right\}\right],$$

which ends the proof by invoking Lemma 20.3. □

20.5 Sample error for finite dimensional $\mathscr{H}$

Let $W:=\operatorname{span}(\phi_1,\ldots,\phi_n)\subset\mathbb{C}(X)$ and the hypothesis space $\mathscr{H}$ be given as

$$\mathscr{H}:=\{h\in W:\|h\|_\infty\leq R\}.$$

Lemma 20.5. *Assume $\mathscr{H}$ is M-bounded in the sense of Assumption 17.1. Then the tail of the worst sampling error decays exponentially, i.e.,*

$$\mathbb{P}\left[\left\{\sup_{h\in\mathscr{H}}|\mathscr{E}(h)|\geq\varepsilon\right\}\right]\leq 2\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{8M}\right)e^{-N\frac{\varepsilon^2}{2M^4}},$$

where

$$\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{8M}\right)\leq\left(\frac{16RM}{\varepsilon}+1\right)^n.$$

Proof. The bound of the covering number of $\mathscr{H}$ using balls with radii $\varepsilon/8M$, $\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{8M}\right)$, is provided in Proposition 20.1. The assertion is readily available using Lemma 20.4. □

20.6 Sampling error when $\mathscr{H}$ is compact

In this section we assume that $\mathscr{H}$ is a compact subset of $\mathbb{C}(X)$. From Lemma 20.2 we know that the covering number $\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{8M}\right)$ for $\mathscr{H}$ is finite.[2] The sample error in Lemma 20.5 is still valid, but in this case we leave the covering number $\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{8M}\right)$ undefined/unestimated.

20.7 Sample error

We are now in the position to estimate the sample error $\mathscr{S}\left(\hat{h}_N\right)$ by bounding the two sampling errors on the right hand side of (20.1). To be concrete we consider the case when $\mathscr{H}$ is a compact subset of $\mathbb{C}(X)$, and this in fact covers the other cases except[3] the case in Section 20.3.

Since $\hat{h}$ is a deterministic function, we can use the sampling error estimation in Section 20.1 to conclude that

$$\left|\mathscr{R}_N\left(\hat{h}\right)-\mathscr{R}\left(\hat{h}\right)\right|\leq\frac{\varepsilon}{2}$$

with the probability at least

$$1-2e^{-N\frac{\varepsilon^2}{2M^4}}.$$

Nothing less will do!

Now, as remarked in Section 16.2, $\hat{h}_N$ is a random function and we have to employ the worst case error to bound $\left|\mathscr{R}\left(\hat{h}_N\right)-\mathscr{R}_N\left(\hat{h}_N\right)\right|$. Lemma 20.5 says that

$$\left|\mathscr{R}\left(\hat{h}_N\right)-\mathscr{R}_N\left(\hat{h}_N\right)\right|\leq\frac{\varepsilon}{2}$$

[2] The subject of estimating covering numbers is standard (but technical) in functional analysis, we refer the readers to [13].

[3] The reason is that balls in infinite dimensional spaces are not compact!

with the probability at least

$$1-2\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{16M}\right)e^{-N\frac{\varepsilon^2}{8M^4}}.$$

Combining these results we conclude that the sample error is bounded by any ε, i.e.,

$$\mathscr{S}\left(\hat{h}_N\right)\leq\varepsilon$$

with the probability at least

$$\left[1-2e^{-N\frac{\varepsilon^2}{2M^4}}\right]\times\left[1-2\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{16M}\right)e^{-N\frac{\varepsilon^2}{8M^4}}\right]$$
$$\geq 1-2\left[\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{16M}\right)+1\right]e^{-N\frac{\varepsilon^2}{8M^4}}.$$

In summary we have proved the following result.

Theorem 20.1 (Sample error estimation). *Suppose $\mathscr{H}$ is a compact subset of $\mathbb{C}(X)$ and $\mathscr{H}$ is M-bounded. The following estimation of the sample error*

$$\mathscr{S}\left(\hat{h}_N\right)\leq\varepsilon,\quad \varepsilon>0,$$

holds true with probability at least

$$1-2\left[\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{16M}\right)+1\right]e^{-N\frac{\varepsilon^2}{8M^4}}.$$

20.8 Appendix

Definition 20.1 (ε-net). Let $\mathscr{H}$ be a metric space with a metric $\|\cdot\|$. A subset W of $\mathscr{H}$ is called a ε-net of $\mathscr{H}$ if any point in $\mathscr{H}$ is within ε-distance from a point in W, i.e.,

$$\forall h\in\mathscr{H},\exists h_0\in W:\|h-h_0\|\leq\varepsilon.$$

Equivalently, a subset W of $\mathscr{H}$ is called a ε-net of $\mathscr{H}$ if and only if $\mathscr{H}$ can be covered by balls of radius ε with centers in W.

Definition 20.2 (Covering number). Let $\mathscr{H}$ be a metric space and $\eta>0$. The covering number $\mathscr{N}(\mathscr{H},\eta)$ is defined as the minimal number such that there exist $\mathscr{N}(\mathscr{H},\eta)$ balls in $\mathscr{H}$ with radius η covering $\mathscr{H}$.

Proposition 20.1 (Coverning number for a ball in finite dimensional space). *Let n be the dimension of a Banach space W and $\mathscr{H}:=\{h\in W:\|h\|_\infty\leq R\}=B_0(R)$. Then for $0<\eta<R$, we have*

$$\mathscr{N}(\mathscr{H},\eta)\leq\left(\frac{2R}{\eta}+1\right)^n.$$

Since $\mathscr{N}(B_0(R), \eta) = \mathscr{N}(B_0(1), \eta/R)$, it is sufficient to consider unit ball in W.

Proposition 20.2 (Finite cover of compact sets). *Let $\mathscr{H}$ be a compact subset of $\mathbb{C}(X)$, then there exists a finite cover for $\mathscr{H}$.*

Chapter 21
Approximation Error (Bias)

We remark in chapter 17, in particular the identity (17.7), that the approximation error (or the bias) depends only on the approximation capability of the hypothesis space $\mathscr{H}$. Estimating the bias is thus a problem in the classical approximation theory, which is vast and technical. In this chapter we shall consider a simple case in which the hypothesis space $\mathscr{H}$ is spanned by a finite number of eigenfunctions of a kernel K defined by the inverse of a fractional Laplacian. Note that the inverse of a fractional Laplacian (with correct power) can be used as the covariance operator of a Gaussian measure. Thus, the presentation of this chapter is valid for Gaussian prior or smooth hypothesis spaces.

FOLLOW CHAPTER II SECTION 1 OF SMALE'S PAPER FOR THIS CHAPTER

21.1 Sampling error for a function in $\mathscr{H}$

Chapter 22
Bias-Variance Tradeoff II

FOLLOW CHAPTER II SECTION 1 OF SMALE'S PAPER FOR THIS CHAPTER

22.1 Sampling error for a function in $\mathscr{H}$

Chapter 23
The universal approximation theorem for sigmoidal functions

As we have discussed in Chapters 18 and 19, we are limited to hypothesis spaces which are compact subset of $\mathbb{C}(X)$. Specifically we have considered hypothesis spaces defined as bounded subsets of (Mercer) kernel-based RKHS. Provided that the eigenfunctions φ_i of the corresponding integral operator are available, we can express the empirical target function $\hat{h}_N$ as a linear combination of φ_i:

$$\hat{h}_N = \sum_{i=1}^{\mathscr{N}} \alpha_i \varphi_i,$$

and the empirical risk minimization problem (23.1) is equivalent to

$$\boxed{\min_{\alpha:=[\alpha_1,\ldots,\alpha_{\mathscr{N}}]} \mathscr{R}_N(\alpha) = \frac{1}{N}\sum_{i}^{N}\left(\sum_{i=1}^{\mathscr{N}} \alpha_i \varphi_i(\mathbf{x}_i) - y_i\right)^2.} \tag{23.1}$$

In general, constructing (evaluating) φ_i is not a trivial task and we have to resort to universal bases that are available at no cost. One of the most popular universal bases in machine learning consists of sigmoidal functions. The purpose of this chapter is to show that, similar to the eigenfunctions φ, the sigmoidal functions is dense in $\mathbb{C}(X)$. We follow closely the original proof of Cybenko [14].

23.1 The universal approximation theorem

Definition 23.1 (Sigmoidal functions). Any function $\sigma : \mathbb{R} \ni t \mapsto \sigma(t) \in \mathbb{R}$ with the property

$$\sigma(t) \to \begin{cases} 1 & \text{when } t \to +\infty \\ 0 & \text{when } t \to -\infty \end{cases}$$

is called a *sigmoidal* function or simply *sigmoid*.

We denote by $\mathscr{M}(X)$ the space of *finite* signed regular Borel measures on X.

Definition 23.2 (Discriminatory). σ is called *discriminatory* if for any $\mu \in \mathscr{M}(X)$

$$\int_X \sigma\left(\mathbf{y}^T\mathbf{x}+b\right)\mu(d\mathbf{x}) = 0, \quad \forall \mathbf{y} \in \mathbb{R}^k \text{ and } b \in \mathbb{R}, \tag{23.2}$$

implies $\mu \equiv 0$.

Lemma 23.1 (Discriminatory of sigmoidal functions). *If a sigmoidal function is bounded and measurable, then it is discriminatory.*

Proof. We want to show that (23.2) implies $\mu \equiv 0$, and it is sufficient to consider the following class of sigmoidal functions

$$\sigma_\lambda\left(\lambda\left(\mathbf{y}^T\mathbf{x}+b\right)+\varphi\right) \overset{\lambda\to+\infty}{\to} \gamma(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{y}^T\mathbf{x}+b > 0 \\ 0 & \text{if } \mathbf{y}^T\mathbf{x}+b < 0 \\ \sigma(\varphi) & \text{if } \mathbf{y}^T\mathbf{x}+b = 0 \end{cases}.$$

Since $\sigma_\lambda\left(\lambda\left(\mathbf{y}^T\mathbf{x}+b\right)+\varphi\right)$ converges everywhere to $\gamma(\mathbf{x})$ and $\gamma(\mathbf{x})$ is integrable, i.e.,

$$\int_X \gamma(\mathbf{x})\,\mu(d\mathbf{x}) \le \int_X \max\{1,\sigma(\varphi)\}\,\mu(d\mathbf{x}) = \max\{1,\sigma(\varphi)\}\,\mu(X) < \infty,$$

we can apply the dominated convergence theorem 23.2 to obtain

$$\lim_{\lambda\to+\infty}\int_X \sigma\left(\lambda\mathbf{y}^T\mathbf{x}+b+\varphi\right)\mu(d\mathbf{x}) = \int_X \gamma(\mathbf{x})\,\mu(d\mathbf{x}) = \mu(H(\mathbf{y},b))+\sigma(\varphi)\,\mu(B(\mathbf{y},b)),$$

where $H(\mathbf{y},b) := \left\{\mathbf{x} : \mathbf{y}^T\mathbf{x}+b > 0\right\}$ and $B(\mathbf{y},b) := \left\{\mathbf{x} : \mathbf{y}^T\mathbf{x}+b = 0\right\}$. From (23.2), we have

$$\mu(H(\mathbf{y},b))+\sigma(\varphi)\,\mu(B(\mathbf{y},b)) = 0, \quad \forall \mathbf{y}, b, \text{ and } \varphi,$$

which, by passing φ to the limits $-\infty$ and then ∞, implies

$$\mu(H(\mathbf{y},b)) = 0, \text{ and } \mu(B(\mathbf{y},b)) = 0, \quad \forall \mathbf{y}, b,$$

that is, the measure of any half plane under μ vanishes. We now show that this implies $\mu \equiv 0$. To that end, we observe that

$$\int_X \mathbb{1}_{[\theta,\infty)}\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = \mu(H(\mathbf{y},-\theta))+\mu(B(\mathbf{y},-\theta)) = 0, \text{ and}$$

$$\int_X \mathbb{1}_{(\theta,\infty)}\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = \mu(H(\mathbf{y},-\theta)) = 0,$$

thus by linearity we conclude that

$$\int_X \chi\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = 0$$

for any *simple function* χ supported on any interval in $\mathbb{R}$. Since the set of all simple functions is dense in the space of all measurable functions (see Lemma 23.2), we

have

$$\hat{\mu}(\mathbf{y}) := \int_X e^{i\mathbf{y}^T\mathbf{x}}\mu(d\mathbf{x}) = \int_X \cos\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) + i\int_X \sin\left(\mathbf{y}^T\mathbf{x}\right)\mu(d\mathbf{x}) = 0, \forall \mathbf{y} \in \mathbb{R}^k,$$

that is, the characteristic function $\hat{\mu}(\mathbf{y})$ of μ is identically zero. Theorem 23.3 can now be applied to end the proof. □

Now comes the main result of the chapter.

Theorem 23.1 (Approximation capability of sigmoidal functions). *Let σ be a continuous discriminatory sigmoidal functions. The set*

$$M := \left\{\sigma\left(\mathbf{y}^T\mathbf{x} + b\right)\right\}, \quad \forall \mathbf{y} \in \mathbb{R}^k, b \in \mathbb{R},$$

is dense in $\mathbb{C}(X)$. In particular, for any $f \in \mathbb{C}(X)$ and any $\varepsilon > 0$, there exists an $n \in \mathbb{N}$ such that

$$\left\| \sum_{i=1}^{n} \alpha_i \sigma\left(\mathbf{y}_i^T\mathbf{x} + b_i\right) - f(\mathbf{x}) \right\|_\infty < \varepsilon.$$

Proof. We proceed by contradiction using the Hahn-Banach theorem 23.4. Suppose $M \subset \mathbb{C}(X)$ is not dense in $\mathbb{C}(X)$ and thus its closure $\overline{M}$ is a proper subset of $\mathbb{C}(X)$. By the Hahn-Banach extension theorem, there exists a bounded linear functional $\mathscr{L}$ on $\mathbb{C}(X)$ such that $\mathscr{L}|_{\overline{M}} = 0$, but $\mathscr{L} \neq 0$. By a Riesz representationt theorem 23.5, there exists a unique finite signed regular measure μ on $\mathbb{C}(X)$ such that

$$\mathscr{L}(f) = \int_X f\mu(d\mathbf{x}), \quad \forall f \in \mathbb{C}(X).$$

Now taking $f(\mathbf{x}) = \sigma\left(\mathbf{y}^t\mathbf{x} + b\right)$ we have

$$\mathscr{L}(\sigma) = \int_X \sigma\left(\mathbf{y}^T\mathbf{x} + b\right)\mu(d\mathbf{x}) = 0, \quad \forall \mathbf{y} \in \mathbb{R}^k, b \in \mathbb{R},$$

which, by discrimninatory of sigmoidal functions in Lemma 23.1, implies

$$\mu \equiv 0,$$

and hence $\mathscr{L} \equiv 0$, a contradiction, and this ends the proof. □

23.2 Appendix

Theorem 23.2 (Dominated convergence theorem). *Let $f_n(\mathbf{x})$ be sequence of measurable functions with respect to μ such that*

- $\lim_{n\to\infty} f_n(\mathbf{x}) = f(\mathbf{x})$ *for any* $\mathbf{x} \in X$,
- $|f_n(\mathbf{x})| \leq g(\mathbf{x})$, *for all* $\mathbf{x}$, n *and*

- $g(\mathbf{x}) \in L^1(X, \mu)$.

Then

i) $f(\mathbf{x}) \in L^1(X, \mu)$.
ii) $\lim_{n\to\infty} \int_X f_n(\mathbf{x})\, \mu(d\mathbf{x}) = \int_X f(\mathbf{x})\, \mu(d\mathbf{x})$.

That is, (a.e) pointwise convergence implies L^1-convergence.

Definition 23.3 (Simple functions). A simple function is a finite combination of characteristic (or indicator) functions. In particular,

$$f(\mathbf{x}) := \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\mathbf{x}),$$

where $A_i \subset X$ and $n \in \mathbb{N}$ is finite, is called a simple function.

Definition 23.4 (Denseness). A set $B \subset X$ is dense in X if for any $\mathbf{x} \in X$, there exists a sequence in B converges to $\mathbf{x}$ with respect to the topology in X.

Lemma 23.2 (Denseness of simple functions). *The space of simple functions is dense in the space of measurable functions. That is, for any measurable $F(\mathbf{x})$, there exists a sequence of simple functions $f^m(\mathbf{x})$ such that*

$$F(\mathbf{x}) = \lim_{m\to\infty} f^m(\mathbf{x}), \quad \forall \mathbf{x} \in X.$$

Definition 23.5 (Characteristic function of a measure). For a probability measure μ on X, the characteristic function is defined as

$$\hat{\mu}(\mathbf{y}) := \int_X e^{i\mathbf{y}^T \mathbf{x}} \mu(d\mathbf{x}),$$

for any $\mathbf{y} \in X$.

Theorem 23.3 (Equivalence of a measure and its characteristic functions). *For any two probability measures μ_1 and μ_2 on X, we have*

$$\hat{\mu}_2(\mathbf{y}) = \hat{\mu}_2(\mathbf{y}), \forall \mathbf{y} \in X \quad \textit{implies} \quad \mu_1 = \mu_2.$$

In particular, if $\hat{\mu}(\mathbf{y}) = 0$ for all $\mathbf{y} \in X$, then $\mu = 0$.

Theorem 23.4 (Hahn-Banach). *Let W be a real vector space, $V \subset W$, p a sublinear functional on W and L a linear functional on V such that $L(f) \leq p(f)$ for any $f \in V$. Then there exists a linear functional $\mathscr{L}$ on W such that $\mathscr{L}(f) \leq p(f)$ for all $f \in W$ and $\mathscr{L}|_V = L$.*

Theorem 23.5 (Riesz representation theorem for $\mathbb{C}(X)$). *Let X be a compact metric spaces and $\mathscr{L}$ be a linear functional on $\mathbb{C}(X)$. There exists a unique finite regular signed measure $\mu \in \mathscr{M}(X)$ such that*

$$\mathscr{L}(f) = \int_X f(\mathbf{x})\,\mu(d\mathbf{x}), \quad \forall f \in \mathbb{C}(X).$$

Chapter 24
Deep Neural Networks

24.1 Perceptrons as artificial neurons

A perceptron is a function f whose value is either 0 or 1 depending on weighted linear combination of the inputs. In particular, let $\mathbf{x} \in \mathbb{R}^k$ be the input vector, $\mathbf{w}$ the "weight" vector, and b the "bias", a perceptron $f : \mathbb{R}^k \to \{0,1\}$ is defined as

$$f(\mathbf{w}\cdot\mathbf{x}+b) = \begin{cases} 0 & \text{if } \mathbf{w}\cdot\mathbf{x}+b \leq 0 \\ 1 & \text{if } \mathbf{w}\cdot\mathbf{x}+b > 0 \end{cases}, \tag{24.1}$$

which can be understood as a device to make decision—Yes (1) or No (0)—based on weighting the evidence and at the same time taking the bias into account.

24.2 Sigmoid (Logistic) neurons

The main problem with perceptron neurons is that they are discontinuous functions. This makes the task of learning (optimizing) the weights $\mathbf{w}$ and the bias b nontrivial as small changes in either of them can make a large change in the perceptron output (an ill-posed construction). From Definition 23.1 perceptron neurons are sigmoidal functions. This observation suggests that we can use continuous sigmoids as continuous surrogates for perceptrons. The most popular one is the following *logistic* function (also known as sigmoid function)

$$\sigma(z) = \frac{1}{1+e^{-z}}.$$

24.3 One-layer neural networks

Chapter 23 shows that linear combinations of sigmoid neurons form a dense subset of $\mathbb{C}(X)$. In particular, we can express the empirical target function $\hat{h}_N$ as

$$\hat{h}_N(\mathbf{x}) = \sum_{i=1}^{\mathscr{N}} \alpha_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i),$$

which can approximate any function in $\mathbb{C}(X)$ (and hence $L^2(X)$ because $\mathbb{C}(X)$ is dense in $L^2(X)$) with any desired accuracy by appropriately choosing the number of sigmoids $\mathscr{N}$, the weights $\mathbf{w}_i$, and the biases b_i. For simplicity, let us assume that $\mathscr{N}$ is fixed and the task is to seek the best (the most appropriate) weights and biases. One way to accomplish this is to minimize the emprical risk

$$\boxed{\min_{\alpha,\mathbf{w}_j,b_j} \frac{1}{N}\sum_{i}^{N} \left(\hat{h}_N(\mathbf{x}_i) - y_i\right)^2 = \frac{1}{N}\sum_{j}^{N} \left(\sum_{i=1}^{\mathscr{N}} \alpha_j \sigma(\mathbf{w}_j \cdot \mathbf{x}_i + b_j) - y_i\right)^2.} \tag{24.2}$$

Intuitively, for binary classification, i.e. $y_i \in \{0,1\}$, each sigmoid neuron makes decision by weighting the input $\mathbf{x}$ plus biases, and the empirical target function then makes decision by combining decisions from the signmoids. Mathematically, (24.2) is a regression problem that best fits the data y_i. If y is a highly nonlinear function, we expect a large number of sigmoid neurons in order to best fit the data.

24.4 Deep neural networks

We know that complicated and/or nonlinear functions can be constructed/approximated by composing elementary functions. In this spirit, we can construct highly nonlinear/complex functions by composing several layers of sigmoid neurons. Let us denote by L the total number of layers and $\mathbf{a}^\ell$ the "activation" (decision) vector from the neurons at the $(\ell-1)$th layer with weights $\mathbf{w}^\ell$ (generally a matrix) and biases $\mathbf{b}^\ell$ (generally a vector), i.e.,

$$\mathbf{a}^\ell := \sigma\left(\mathbf{w}^\ell \cdot \mathbf{a}^{\ell-1} + \mathbf{b}^\ell\right), \quad \ell = 1,\ldots,L.$$

Here, the action of the sigmoid σ function on a vector is an elementwise operation, i.e.,

$$\mathbf{a}_i^\ell := \sigma\left(\mathbf{w}^\ell(i,:) \cdot \mathbf{a}^{\ell-1} + \mathbf{b}_i^\ell\right),$$

with $\mathbf{w}^\ell(i,:)$ denotes the ith row of $\mathbf{w}^\ell$. We end the definition of deep neural networks by assigning

$$\mathbf{a}^0 := \mathbf{x}, \quad \text{and} \quad \hat{h}_N = \sum_{i=1}^{\mathscr{N}} \alpha_i \mathbf{a}_i^L.$$

The emprical risk minimization problem now becomes

$$\boxed{\begin{array}{l}\min\limits_{\alpha,\mathbf{a}^\ell,\mathbf{b}^\ell,\ell=1,\ldots,L} \dfrac{1}{N}\displaystyle\sum_i^N \left(\hat{h}_N(\mathbf{x}_i)-y_i\right)^2 = \dfrac{1}{N}\displaystyle\sum_j^N \left(\sum_{i=1}^{\mathscr{N}} \alpha_j \mathbf{a}^L - y_i\right)^2. \\ \quad \text{subject} \quad \text{to} \\ \mathbf{a}^\ell = \sigma\left(\mathbf{w}^\ell \cdot \mathbf{a}^{\ell-1} + \mathbf{b}^\ell\right), \quad \ell = 1,\ldots,L. \\ \mathbf{a}^0 = \mathbf{x}_i.\end{array}} \tag{24.3}$$

We will solve the empirical risk mininimization problems (24.2) and (24.3) using the stochastic gradient approach developed in Chapter 25.

Chapter 25
Learning with Stochastic Gradient algorithm

The empirical risk minimization (aka learning) problems (24.2) and (24.3) are neither convex nor linear in general. Learning with massive amount of data, i.e. very large sample size N, can be time consuming and effective optimization method is desirable. One of the most popular approaches is the stochastic gradient method. This chapter studies and analyzes the convergence of this "gold-standard" method. For the sake of clarity, we follow the presentation of [5]. That is, we shall consider only strongly convex functions (risks).

25.1 Stochastic gradient algorithm

To make the exposition general, let us consider the cost function $F(\mathbf{w})$ where $\mathbf{w}$ is the optimization variable. For example, F could be either the risk or emprical risk and $\mathbf{w}$ is the combination of all the weights $\mathbf{w}^\ell$ and the biases $\mathbf{b}^\ell$. Let $g(\mathbf{w},\mathbf{x})$ be an approximation to the gradient of F where $\mathbf{x}$ is some random variable. For the empirical risk minimization problem (24.3), $g\left(\mathbf{w}^k,\mathbf{x}\right)$ is typically sub-sample gradient. For example, the approximation gradient at the kth iteration can be given as

$$g^k := g\left(\mathbf{w}^k,\mathbf{x}^k\right) := \nabla\left(\frac{1}{n^k}\sum_i^{n^k}\left(\hat{h}_N\left(\mathbf{x}_i^k\right)-y_i\right)^2\right),$$

where $1 \leq n^k \leq N$ is the sub-sample size at the kth iteration.

Assumption 25.1 (Smooth objective function). F is continuously differentiable with gradient ∇F and the gradient is Lipschitz continuous with Lipschitz constant C, i.e.,

$$\|\nabla F(\mathbf{w}) - \nabla F(\overline{\mathbf{w}})\| \leq C\|\mathbf{w}-\overline{\mathbf{w}}\|.$$

Lemma 25.1. *With assumption 25.1, there holds*

Algorithm 4 Stochastic gradient algorithm

1. Choose an intial guess $\mathbf{w}^1$.
2. **For** $k = 1, \ldots$ **do**:
   - Generate a realization of random variable $\mathbf{x}^k$.
   - Compute the stochastic gradient $g^k := g\left(\mathbf{w}^k, \mathbf{x}^k\right)$.
   - Choose stepsize α_k
   - Set the new iterate as
   $$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha_k g^k.$$
3. **End For**.

$$F(\mathbf{w}) \leq F(\mathbf{w}) + \nabla F(\overline{\mathbf{w}}) \cdot (\mathbf{w} - \overline{\mathbf{w}}) + \frac{1}{2} C \|\mathbf{w} - \overline{\mathbf{w}}\|^2.$$

Proof. From the fundamental theorem of calculus we have

$$F(\mathbf{w}) = F(\overline{\mathbf{w}}) + \int_0^1 \frac{\partial}{\partial t} F(\overline{\mathbf{w}} + t(\mathbf{w} - \overline{\mathbf{w}}))\, dt = F(\overline{\mathbf{w}}) + \int_0^1 \nabla F(\overline{\mathbf{w}} + t(\mathbf{w} - \overline{\mathbf{w}})) \cdot (\mathbf{w} - \overline{\mathbf{w}})\, dt$$

□

Chapter 26
Back-Propagation and Adjoint Method

Part VI
An Introduction to Infinite Dimensional Analysis

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Glossary

Use the template *glossary.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your glossary in the Springer layout.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

Solutions

Problems of Chapter ??

?? The solution is revealed here.

?? Problem Heading
(a) The solution of first part is revealed here.
(b) The solution of second part is revealed here.

References

1. Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687, 2003.
2. Nenad Antonić and Krešimir Burazin. Graph spaces of first-order linear partial differential operators. *Mathematical Communications*, 14(1):135–155, 2009.
3. Todd Arbogast and Jerry L. Bona. *Methods of Applied Mathematics*. University of Texas at Austin, 2008. Lecture notes in applied mathematics.
4. Z. Bai, G. Fahey, and G. Golub. Some large-scale matrix computation problems. *Journal of computational and applied mathematics*, 74(1-2):71–89, 1996.
5. L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
6. Tan Bui-Thanh, Leszek Demkowicz, and Omar Ghattas. A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. *SIAM J. Numer. Anal.*, 51(4):1933–1958, 2013. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/DPGunifiedRevised.pdf.
7. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):055001, 2012. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactI.pdf.
8. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves. *Inverse Problems*, 28(5):055002, 2012. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactII.pdf.
9. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves. Inverse Problems and Imaging, 2013. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/EM3Dmedium.pdf.
10. D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*. Springer, New York, 2007.
11. David Colton and Rainer Kress. *Integral equation methods in scattering theory*. John Wiley & Sons, 1983.
12. David Colton and Rainer Kress. *Inverse Acoustic and Electromagnetic Scattering*. Applied Mathematical Sciences, Vol. 93. Springer-Verlag, Berlin, Heidelberg, New-York, Tokyo, second edition, 1998.
13. F. Cucker and D.X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge monographs on applied and computational mathematics. Cambridge University Press, 2007.
14. G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.
15. Masoumeh Dashti and Andrew M. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM Journal on Numerical Analysis*, 49(6):2524–2542, 2011.
16. Rick Durret. *Probability: theory and examples*. Cambidge University Press, 2010.
17. Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.
18. Ping Li, Trevor J. Hastie, and Kenneth W. Church. Very sparse random projections. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 287, 2006.
19. D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, New York, 1969.
20. J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.
21. Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London, 1993.
22. J. Tinsley Oden and Leszek F. Demkowicz. *Applied functional analysis*. CRC Press, 2010.
23. Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):pp. 351–367, 2001.
24. Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

25. W. Rudin. *Functional Analysis*. McGraw-Hill, New York, St. Louis, San Francisco,..., 1973.
26. Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

Index