Bayesian Inverse Problem

Collector: Nguyen Quan Ba Hong

August 3, 2021

Contents

Bibliography   **50**

Chapter 1

Introduction

1.1 Inverse problem

References.

- Aster, Borchers, and Thurber, 2018 [pdf]

- Ito and Jin, 2015 [pdf]

- Kirsch, 2021 [pdf]

- [**W/IP**] Wikipedia/Inverse Problem.

Sketches.

1. **Definition.** [**W/IP**] "An *inverse problem* in science is the process of calculating from a set of observations the causal factors that produced them."

2. **Intuition.** "It is called an inverse problem because it starts with the effects and then calculates the causes. It is the inverse of a *forward problem*, which starts with the causes and then calculates the effects."

3. **Motivation.** "Inverse problems are some of the most important mathematical problems in science and mathematics because they tell us about parameters that we cannot directly observe."

4. **Application.** Inverse problems "have wide application in system identification, optics, radar, acoustics, communication theory, signal processing, medical imaging, computer vision, geophysics, oceanography, astronomy, remote sensing, natural language processing, machine learning, nondestructive testing, slope stability analysis and many other fields.

Communities.

- WIAS/Research Group 4: *Nonlinear Optimization & Inverse Problem*.

1.1.1 Forward problem vs. inverse problem

Physical Model. $\mathcal{G}(u) \to y$.

- $u \in X$ parameter vector/parameter function

- $\mathcal{G} : X \to Y$ forward response operator

- y results/observations

- Evaluation of $\mathcal{G}$ expensive

Forward Problem. Find the output y for given parameters u.

Find the *unknown data $u \in X$* from *noisy observations*

$$y = \mathcal{G}(u) + \eta \text{ with } \eta \sim \mathcal{N}(0, \Gamma).$$

- $u \in X$ parameter vector/parameter function

- $\mathcal{G} : X \to Y$ forward response operator

- y result/observations

- Evaluation of $\mathcal{G}$ expensive.

Forward Problem. $\boxed{\text{Find the output } y \text{ for given parameters } u}$ $\to$ **well-posed**.

Inverse Problem. $\boxed{\text{Find the parameters } u \text{ from (noisy) observations } y}$ $\to$ **ill-posed**.

Example 1.1.1 (2D Inhomogeneous sideways heat equations)**.** *Consider the following 2D inhomogeneous sideways heat equation (SHE):*

$$\begin{cases} \partial_t u - \Delta u = f(t, x, y), & in \ [0, \infty) \times \mathbb{R}^2, \\ u(t, 0, y) = \phi(t, y), & in \ [0, \infty) \times \mathbb{R}, \\ \partial_x u(t, 0, y) = \psi(t, y) & in \ [0, \infty) \times \mathbb{R}, \end{cases} \quad \text{(2D-iSHE)}$$

where ϕ, ψ: temperature, flux data given in the accessible region $x = 0$, f: heat source density.

Approaches to represent solutions: separation of variables, transform methods, etc., see, e.g., Evans, 2010.

(i) ***Determination of temperature.*** *Denote the 2D Fourier transform of a function $f(t, y)$ by*

$$\hat{f}(\tau, \xi) = \mathcal{F}(f)(\tau, \xi) = \frac{1}{2\pi} \int_{\mathbb{R}^2} f(t, y) e^{-i(\tau t + \xi y)} \mathrm{d}y \mathrm{d}t, \ \forall (\tau, \xi) \in \mathbb{R}^2. \quad \text{(2D-Ft)}$$

extend the domain of time t from $[0, \infty)$ to $\mathbb{R}$, then applying the 2D Fourier transform (2D-Ft) *w.r.t. y, t to* (2D-iSHE)*, obtain in the frequency space the following 2nd-order ODE:*

$$\begin{cases} \partial_x^2 \hat{u} - (i\tau + \xi^2) \hat{u}(\tau, x, \xi) = \hat{f}(\tau, x, \xi), & in \ \mathbb{R}^3, \\ \hat{u}(\tau, 0, \xi) = \hat{\phi}(\tau, \xi), & in \ \mathbb{R}^2, \\ \partial_x \hat{u}(\tau, 0, \xi) = \hat{\psi}(\tau, \xi), & in \ \mathbb{R}^2. \end{cases}$$

Solving this ODE yields

$$\hat{u}(\tau, x, \xi) = \hat{\phi}(\tau, \xi) \cosh(x\theta) + \hat{\psi}(\tau, \xi) \frac{\sinh(x\theta)}{\theta} - \int_0^x \frac{\sinh((x - \eta)\theta)}{\theta} \hat{f}(\tau, \eta, \xi) \mathrm{d}\eta, \ \forall (\tau, x, \xi) \in \mathbb{R}^3, \quad (\hat{u})$$

where $\theta = \theta(\tau, \xi)$ is the principle value of $\sqrt{i\tau + \xi^2}$ and is given by

$$\theta(\tau, \xi) = \sqrt{\frac{\sqrt{\tau^2 + \xi^4} + \xi^2}{2}} + i \, \mathrm{sgn}(\tau) \sqrt{\frac{\sqrt{\tau^2 + \xi^4} - \xi^2}{2}}, \ \forall (\tau, \xi) \in \mathbb{R}^2.$$

This representation of $\hat{u}$ also indicates the ill-posedness of (2D-iSHE)*. Indeed, for $i = 1, 2$, consider u_i solves* (2D-iSHE) *with the data (f_i, ϕ_i, ψ_i), and then*

$$\hat{u}_i(\tau, x, \xi) = \hat{\phi}_i(\tau, \xi) \cosh(x\theta) + \hat{\psi}_i(\tau, \xi) \frac{\sinh(x\theta)}{\theta} - \int_0^x \frac{\sinh((x - \eta)\theta)}{\theta} \hat{f}_i(\tau, \eta, \xi) \mathrm{d}\eta, \ \forall (\tau, x, \xi) \in \mathbb{R}^3, \ for \ i = 1, 2.$$

Subtract these, term by term, to obtain:

$$(\hat{u}_1 - \hat{u}_2)(\tau, x, \xi) = (\hat{\phi}_1 - \hat{\phi}_2)(\tau, \xi) \cosh(x\theta) + (\hat{\psi}_1 - \hat{\psi}_2)(\tau, \xi) \frac{\sinh(x\theta)}{\theta} - \int_0^x \frac{\sinh((x - \eta)\theta)}{\theta} (\hat{f}_1 - \hat{f}_2)(\tau, \eta, \xi) \mathrm{d}\eta,$$

for all $(\tau, x, \xi) \in \mathbb{R}^3$.

The perturbations in the data, which are indicated by $(\hat{f}_1 - \hat{f}_2, \hat{\phi}_1 - \hat{\phi}_2, \hat{\psi}_1 - \hat{\psi}_2)$, are magnified by the hyperbolic function $\cosh$ and even a singular kernel $\frac{\sinh(x\theta)}{\theta}$ whose a singularity is the origin of the frequency space, i.e., $(\tau, \xi) = (0, 0)$.

Remark 1.1.1 (Generalization: $x = 0$ to $x = x_0 \in \mathbb{R}$). *Note that by using the 2D Fourier transform* (2D-Ft), *the solution of* (2D-iSHE) *where the Cauchy data* (ϕ, ψ) *is posed at* $x = x_0 \in \mathbb{R}$ *instead of* $x = 0$, *can be represented as*

$$\hat{u}(\tau, x, \xi) = \hat{\phi}(\tau, \xi) \cosh((x - x_0)\theta) + \hat{\psi}(\tau, \xi) \frac{\sinh((x - x_0)\theta)}{\theta} - \int_{x_0}^{x} \frac{\sinh((x - \eta)\theta)}{\theta} \hat{f}(\tau, \eta, \xi) \mathrm{d}\eta, \ \forall (\tau, x, \xi) \in \mathbb{R}^3.$$

- **Ill-posedness.** *Both kernels* $\cosh(x\theta)$, $\frac{\sinh(x\theta)}{\theta}$ *are unbounded for arbitrarily* $x \neq 0$ *when* $\|(\tau, \xi)\|_2^2 = \tau^2 + \xi^2 \to \infty$, *and thus small errors in the data can blow up and ultimately destroy the solution.*

- **Measured data.** *The measured data, denoted by* $(f_\delta, \phi_\delta, \psi_\delta)$, *where* $\delta > 0$ *is called an* error level *(see, e.g., Ito and Jin,* 2015; *Kirsch,* 2021), *is typically assumed to satisfy*

$$\max\{\|f_\delta - f\|_{L^2(\mathbb{R}^3)}, \|\phi_\delta - \phi\|_{L^2(\mathbb{R}^2)}, \|\psi_\delta - \psi\|_{L^2(\mathbb{R}^2)}\} \leq \delta. \tag{noise}$$

(ii) **Determination of the thermal flux.** *Differentiate the RHS of* ($\hat{u}$) *w.r.t.* x, *with the help of Leibnitz integral rule for the integral term, to obtain the following representation of the thermal flux w.r.t.* x

$$\partial_x \hat{u}(\tau, x, \xi) = \theta \hat{\phi}(\tau, \xi) \sinh(x\theta) + \hat{\psi}(\tau, \xi) \cosh(x\theta) - \int_0^x \cosh((x - \eta)\theta) \hat{f}(\tau, \eta, \xi) \mathrm{d}\eta, \ \forall (\tau, x, \xi) \in \mathbb{R}^3.$$

Blow-up kernels: $\sinh(x\theta)$, $\cosh((x - \eta)\theta) \to$ *ill-posedness.*

(iii) **Numerical implementation.** *The Cauchy data* $(u, \partial_x u)|_{x=0} = (\phi, \psi)$ *and the source term* $f(t, x, y)$ *are given and sampled at an equidistant grid in the domain* $[L, R]^2$ *with* $n \times n$ *points (n: sampling frequency), and the discretizations of* (f, ϕ, ψ) *are denoted by* $(\mathbf{F}, \Phi, \Psi)$, *to distinguish them and their continuous counterparts.*

Main tools: MATLAB *builtin-in function* `fft2`, *which is used to carry out the 2D fast Fourier transforms (abbr., 2D FFT) of the given data, and* `ifft2`, *which is used to obtain 2D inverse fast Fourier transforms (abbr., 2D IIFT).*

Denote $\bar{x} \in [0, 1]$ *the point at which we choose to observe the solutions and the thermal fluxes. Use Gauss-Legendre quadrature of order* m *for numerical integrations, where* $m \in \mathbb{Z}_{>0}$ *is chosen by users, i.e.,*

$$\int_0^{\bar{x}} \widehat{f_\delta}(\tau, \eta, \xi) \frac{\sinh((\bar{x} - \eta))}{\theta} \mathrm{d}\eta \approx \sum_{k=1}^{m} w_{m,k} \widehat{f_\delta}(\tau, x_{m,k}, \xi) \frac{\sinh((\bar{x} - x_{m,k})\theta)}{\theta},$$

$$\int_0^{\bar{x}} \widehat{f_\delta}(\tau, \eta, \xi) \cosh((\bar{x} - \eta)\theta) \mathrm{d}\eta \approx \sum_{k=1}^{m} w_{m,k} \widehat{f_\delta}(\tau, x_{m,k}, \xi) \cosh((\bar{x} - x_{m,k})\theta),$$

where $x_{m,k}$ *and* $w_{m,k}$ *are the nodes and weights of the Gauss-Legendre quadrature of degree* m *for the interval* $[0, \bar{x}]$, *resp. To apply these numerical integration formulas, it is necessary to carry out the 2D FFT of* f_δ *at each Gauss-Legendre quadrature's node.*

Add random noises to the (discrete) exact input data $(\mathbf{F}, \Phi, \Psi)$, *i.e.:*

$$\begin{cases} \mathbf{F}_\delta(\cdot, x_{m,k}, \cdot) = \mathbf{F}(\cdot, x_{m,k}, \cdot) + \delta\, \texttt{randn}(\texttt{size}(\mathbf{F}(\cdot, x_{m,k}, \cdot))), \ \forall k = 1, \ldots, m, \\ \Phi_\delta = \Phi + \delta\, \texttt{randn}(\texttt{size}(\Phi)), \ \Psi_\delta = \Psi + \delta\, \texttt{randn}(\texttt{size}(\Psi)), \end{cases} \tag{noise data}$$

where `randn()` *is a* MATLAB *built-in function.*

Denote $\|\cdot\|_{\overline{\mathrm{F}}}$ *the "average" Frobenius norm, i.e.:*

$$\|A\|_{\overline{\mathrm{F}}} := \frac{1}{\sqrt{mn}} \|A\|_{\mathrm{F}} = \frac{1}{\sqrt{mn}} \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \right)^{\frac{1}{2}}, \ \forall A \in \mathrm{Mat}_{\mathbb{C}}(m \times n),$$

where $\|\cdot\|_{\mathrm{F}}$ *is the Frobenius norm, we can define the total noises of the 3D array* $\mathbf{F}_\delta$ *and the matrices* $(\Phi_\delta, \Psi_\delta)$ *as*

$$\mathrm{TN}(\mathbf{F}, \mathbf{F}_\delta) := \frac{1}{n\sqrt{m}} \left(\sum_{k=1}^{m} \|(\mathbf{F}_\delta - \mathbf{F})(\cdot, x_{m,k}, \cdot)\|_{\mathrm{F}}^2 \right)^{\frac{1}{2}}, \ \mathrm{TN}(\Phi, \Phi_\delta) := \|\Phi_\delta - \Phi\|_{\overline{\mathrm{F}}}, \ \mathrm{TN}(\Psi, \Psi_\delta) := \|\Psi_\delta - \Psi\|_{\overline{\mathrm{F}}}, \ \forall \delta > 0.$$

$$\tag{total noise}$$

After doing this, we obtain the noisy data $(\mathbf{F}_\delta, \Phi_\delta, \Psi_\delta)$ *satisfying the condition* $\max\{\mathrm{TN}(\mathbf{F}, \mathbf{F}_\delta), \mathrm{TN}(\Phi, \Phi_\delta), \mathrm{TN}(\Psi, \Psi_\delta)\} \le$ δ, *which can be interpreted as the discrete analogue of* (noise). *We also use the norm* $\|\cdot\|_{\overline{\mathrm{F}}}$ *to calculate root mean square and relative errors, e.g.,*

$$\mathrm{RMSE}(\mathbf{U}_\mathcal{R}^\delta) := \|\mathbf{U}_\mathcal{R}^\delta - \mathbf{U}_{\mathrm{exact}}\|_{\overline{\mathrm{F}}}, \ \ \mathrm{RE}(\mathbf{U}_\mathcal{R}^\delta) := \frac{\|\mathbf{U}_\mathcal{R}^\delta - \mathbf{U}_{\mathrm{exact}}\|_{\overline{\mathrm{F}}}}{\|\mathbf{U}_{\mathrm{exact}}\|_{\overline{\mathrm{F}}}},$$

where $\mathbf{U}_{\mathrm{exact}}$, $\mathbf{U}_\mathcal{R}^\delta$ *is the discrete version of the solution* u *of* (2D-iSHE) *and its regularized analogue* $u_\mathcal{R}^\delta$ *on the sampled grid.*

Algorithm 1 Numerical implementation of a regularization method proposed for (2D-iSHE).

1: **Inputs:** observed point $x_{\mathrm{obs}} \in \mathbb{R}$, Cauchy data $[\phi, \psi]$ at $x = 0$, source term f, boundaries L and R of the computational domain $\Omega_{L,R}$, sampling frequency n, order $m \in \mathbb{Z}_{>0}$ of Gauss-Legendre quadrature, (approximate) nodes $x_{k,m}$ and (approximate) weights $w_{k,m}$ of the Gauss-Legendre quadrature of degree m for the interval $[0, x_{\mathrm{obs}}]$, noisy level $\delta > 0$, exact solution u or the set of its values at the grid points;
2: Generate an equidistant grid in the computational domain $\Omega_{L,R}$; ▷ Treat the origin $(0,0)$ carefully
3: Compute the discrete exact data $(\mathbf{F}, \Phi, \Psi)$;
4: Generate the discrete noisy data $(\mathbf{F}_\delta, \Phi_\delta, \Psi_\delta)$ by (noise data);
5: (Optional) Compute the triple of the averaged noises $(\mathrm{TN}(\mathbf{F}, \mathbf{F}_\delta), \mathrm{TN}(\Phi, \Phi_\delta), \mathrm{TN}(\Psi, \Psi_\delta))$ of the discrete noisy data generated in Step 4 by (total noise); ▷ This triple will slightly change in magnitude in each execution due to randomness.
6: Compute 2D FFT $(\widehat{\mathbf{F}_\delta}, \widehat{\Phi_\delta}, \widehat{\Psi_\delta})$ of the discrete noisy data in Step 4 by `fft2`;
7: Compute 2D FFT of the discrete versions $\widehat{\mathbf{U}_\delta}$ and $\widehat{\mathbf{U}_\mathcal{R}^\delta}$ of $(\hat{u})$ and the regularized one(s);
8: Carry out both the discrete versions $\mathbf{U}_\delta$ and $\mathbf{U}_\mathcal{R}^\delta$ of the non-regularized and regularized solutions u_δ and $u_\mathcal{R}^\delta$ by `ifft2`;
9: Compute the corresponding Root Mean Square Error $\mathrm{RMSE}(\mathbf{U}_\mathcal{R}^\delta)$ and Relative Error $\mathrm{RE}(\mathbf{U}_\mathcal{R}^\delta)$;
10: (Optional) 3D-Plot the discrete non-regularized solution $\mathbf{U}_\delta$ and regularized solution $\mathbf{U}_\mathcal{R}^\delta$ on the equidistant grid generated in Step 2;
11: **Output:** Triple of the averaged noises $(\mathrm{TN}(\mathbf{F}, \mathbf{F}_\delta), \mathrm{TN}(\Phi, \Phi_\delta), \mathrm{TN}(\Psi, \Psi_\delta))$ (optional), non-regularized discrete solution $\mathbf{U}_\delta$, regularized discrete solution $\mathbf{U}_\mathcal{R}^\delta$, errors $\mathrm{RMSE}(\mathbf{U}_\mathcal{R}^\delta)$, $\mathrm{RE}(\mathbf{U}_\mathcal{R}^\delta)$, 3D plots of non-regularized and regularized solutions.

The MATLAB scripts and a gallery for this algorithm for (2D-iSHE) can be downloaded at the following URLs:

- NQBH/math/seminar/Bayesian inverse problem/MATLAB

- NQBH/math/seminar/Bayesian inverse problem/gallery.

1.1.2 Deterministic Optimization Problem

Find the unknown data $u \in X$ from noisy observations

$$y = \mathcal{G}(u) + \eta.$$

Deterministic optimization problem

$$\min_{u \in X} \frac{1}{2} \|y - \mathcal{G}(u)\|^2 + R(u).$$

- $\|y - \mathcal{G}(u)\|$ potential/data misfit

- R regularization term

- Large-scale, deterministic optimization problem

- No quantification of the uncertainty in the unknown u

- Proper choice of the regularization term R

1.2   Uncertainty quantification

References.

- [**W/U**] Wikipedia/Uncertainty.

- [**W/UQ**] Wikipedia/Uncertainty Quantification.

Sketches.

1. **Definitions.** [**W/U**] "*Uncertainty* refers to epistemic situations involving imperfect or unknown information."

   [**W/UQ**] "*Uncertainty quantification* (UQ) is the science of quantitative characterization and reduction of uncertainties in both computational and real world applications."

2. **Intuition.** Uncertainty quantification "tries to determine how likely certain outcomes are if some aspects of the system are not exactly known".

3. **Motivation.** "Many problems in the natural sciences and engineering are also rife with sources of uncertainty." E.g., (see, e.g., [**Schillings_Teckentrup2017**], Slides 3–4) *environmental systems* (weather, climate, seismic, subsurface geophysics), *engineering systems* (automobiles, aircraft, bridges, structures), *biological systems* (health & medicine, pharmaceuticals, gene expression, cancer research), *physical systems* (nano-optics, quantum physics, radioactive decay), etc.

4. **Sources of uncertainty.** [**W/UQ**] "Uncertainty can enter mathematical models and experimental measurements in various contexts."

   Categorization of sources of uncertainty. Consider: parameter uncertainty, parametric variability, structural uncertainty/model inadequacy/model bias/model discrepancy, algorithmic uncertainty/numerical uncertainty/discrete uncertainty, experimental uncertainty/observation error, interpolation uncertainty.

   Classification of uncertainty.

   - *Aleatoric uncertainty*/*statistical uncertainty* is representative of unknowns that differ each time we run the same experiment. E.g., MATLAB built-in function `rand()`.

     "The quantification for the aleatoric uncertainties can be relatively straightforward, where traditional (frequentist) probability is the most basic form."

   - *Epistemic uncertainty*/*systematic uncertainty* is due to things one could in principle know but do not in practice, which may be because a measurement is not accurate, because the model neglects certain effects, or because particular data has been deliberately hidden. E.g., $g = 9.8\text{m/s}^2$ ignores the effects of air resistance.

     "To evaluate epistemic uncertainties, the efforts are made to understand the (lack of) knowledge of the system, process or mechanism."

   "Uncertainty quantification intends to explicitly express both types of uncertainty separately."

   Mathematical perspective of aleatoric & epistemic uncertainty. "In mathematics, uncertainty is often characterized in terms of a probability distribution. From that perspective, epistemic uncertainty means not being certain what the relevant probability distribution is, and aleatoric uncertainty means not being certain what a random sample drawn from a probability distribution will be."

5. **Quantification & minimization of uncertainties.**

   - Uncertainty quantification in numerical simulations
   - Minimization of uncertainties in design & control problems
   - Identification of uncertain parameters from noisy observations

Related research fields. Numerical analysis, methods for optimal control/optimization problems, approximation theory, stochastic, statistics, algorithms & data structures.

Communities.

- WIAS/Research Group 5: *Interacting Random Systems*.

- WIAS/Research Group 6: *Stochastic Algorithms & Nonparametric Statistics*.

References.

- [**Schillings_Teckentrup2017**] Schillings, Claudia & Teckentrup, Aretha, 2017. *Mathematical Foundations of Bayesian Inverse Problems: Well-posedness & Statistical Estimates in Bayesian Inverse Problems.* [pdf]

In Dashti and Stuart, 2017, the authors highlight the mathematical and computational structure relating to the formulation of, and development of algorithms for, the Bayesian approach to inverse problems in differential equations, which is a fundamental approach in the quantification of uncertainty within applications involving the blending of mathematical models with data.

1.3   Bayesian Inversion in $\mathbb{R}^n$

Consider the problem of finding $\mathbf{u} \in \mathbb{R}^n$ from $\mathbf{y} \in \mathbb{R}^J$ where $\mathbf{u}$ and $\mathbf{y}$ are related by the equation $\mathbf{y} = G(\mathbf{u})$.

We refer to $\mathbf{y}$ as *observed data* and to $\mathbf{u}$ as the *unknown*.

This problem may be difficult for a number of reasons.

Dashti & Stuart highlight 2 of these, both particularly relevant to their future developments.

1. The 1st difficulty, which may be illustrated in the case where $n = J$, concerns the fact that often the equation is perturbed by noise and so we should really consider the equation

$$\mathbf{y} = G(\mathbf{u}) + \boldsymbol{\eta}, \tag{1.3.1}$$

where $\boldsymbol{\eta} \in \mathbb{R}^J$ represents the *observational noise* which enters the observed data.

Assume further that G maps $\mathbb{R}^J$ into a proper subset of itself, Im_G, and that G has a unique inverse as a map from Im_G into $\mathbb{R}^J$.

It may then be the case that, because of the noise, $\mathbf{y} \notin \mathrm{Im}_G$ so that simply inverting G on the data $\mathbf{y}$ will not be possible.

Furthermore, the specific instance of $\boldsymbol{\eta}$ which enters the data may not be known to us; typically, at best, only the statistical properties of a typical noise $\boldsymbol{\eta}$ are known.

Thus we cannot subtract $\boldsymbol{\eta}$ from the observed data $\mathbf{y}$ to obtain something in Im_G.

Even if $\mathbf{y} \in \mathrm{Im}_G$, the uncertainty caused by the presence of noise $\boldsymbol{\eta}$ causes problems for the inversion.

2. The 2nd difficulty is manifest in the case where $n > J$ so that the system is *underdetermined*: the number of equations is smaller than the number of unknowns.

How do we attach a sensible meaning to the concept of solution in this case where, generically, there will be many solutions?

Thinking probabilistically enables us to overcome both of these difficulties.

We will treat $\mathbf{u}, \mathbf{y}$, and $\boldsymbol{\eta}$ as random variables and determine the joint probability distribution of $(\mathbf{u}, \mathbf{y})$.

We then define the "solution" of the inverse problem to be the probability distribution of $\mathbf{u}$ given $\mathbf{y}$, denoted $\mathbf{u}|\mathbf{y}$.

This allows us to $\boxed{\text{model the noise via its statistical properties}}$, even if we do not know the exact instance of the noise entering the given data.

And it also allows us to specify a priori the form of solutions that we believe to be more likely, thereby enabling us to attach weights to multiple solutions which explain the data.

This is the *Bayesian approach* to inverse problems.

To this end, we define a random variable $(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^J$ as follows.

We let $\mathbf{u} \in \mathbb{R}^n$ be a random variable with (Lebesgue) density $\rho_0(\mathbf{u})$.

Assume that $\mathbf{y}|\mathbf{u}$ ($\mathbf{y}$ given $\mathbf{u}$) is defined via the formula (1.3.1) where $G : \mathbb{R}^n \to \mathbb{R}^J$ is measurable and $\boldsymbol{\eta}$ is independent of $\mathbf{u}$ (we sometimes write this as $\boldsymbol{\eta} \perp \mathbf{u}$) and distributed according to measure $\mathbb{Q}_0$ with Lebesgue density $\rho(\boldsymbol{\eta})$.

Then $\mathbf{y}|\mathbf{u}$ is simply found by shifting $\mathbb{Q}_0$ by $G(\mathbf{u})$ to measure $\mathbb{Q}_\mathbf{u}$ with Lebesgue density $\rho(\mathbf{y} - G(\mathbf{u}))$.

It follows that $(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^J$ is a random variable with Lebesgue density $\rho(\mathbf{y} - G(\mathbf{u}))\rho_0(\mathbf{u})$.

The following theorem allows us to calculate the distribution of the random variable $\mathbf{u}|\mathbf{y}$:

Theorem 1.3.1 (Bayes' Theorem)**.** *Assume that*

$$Z := \int_{\mathbb{R}^n} \rho(\mathbf{y} - G(\mathbf{u}))\rho_0(\mathbf{u}) \mathrm{d}\mathbf{u} > 0.$$

Then $\mathbf{u}|\mathbf{y}$ is a random variable with Lebesgue density $\rho^{\mathbf{y}}(\mathbf{u})$ given by

$$\rho^{\mathbf{y}}(\mathbf{u}) = \frac{1}{Z}\rho(\mathbf{y} - G(\mathbf{u}))\rho_0(\mathbf{u}).$$

Remark 1.3.1. *The following remarks establish the nomenclature[1] of Bayesian statistics and also frame the previous theorem in a manner which generalizes to the infinite-dimensional setting.*

- *$\rho_0(\mathbf{u})$ is the **prior density**.*

- *$\rho(\mathbf{y} - G(\mathbf{u}))$ is the **likelihood**.*

- *$\rho^{\mathbf{y}}(\mathbf{u})$ is the **posterior density**.*

- *It will be useful in what follows to define*

$$\Phi(\mathbf{u}; \mathbf{y}) = -\log \rho(\mathbf{y} - G(\mathbf{u})).$$

   *We call Φ the **potential**.*

   *This is the **negative log likelihood**.*

- *Note that Z is the probability of $\mathbf{y}$.*

   Bayes' formula expresses

$$\mathbb{P}(\mathbf{u}|\mathbf{y}) = \frac{1}{\mathbb{P}(\mathbf{y})}\mathbb{P}(\mathbf{y}|\mathbf{u})\mathbb{P}(\mathbf{u}).$$

- *Let $\mu^{\mathbf{y}}$ be a measure on $\mathbb{R}^n$ with density $\rho^{\mathbf{y}}$ and μ_0 a measure on $\mathbb{R}^n$ with density ρ_0.*

   Then the conclusion of Theorem 1.3.1 may be written as:[2]

$$\frac{\mathrm{d}\mu^{\mathbf{y}}}{\mathrm{d}\mu_0}(\mathbf{u}) = \frac{1}{Z}\exp(-\Phi(\mathbf{u}; \mathbf{y})), \ \ Z = \int_{\mathbb{R}^n} \exp(-\Phi(\mathbf{u}; \mathbf{y}))\mu_0(\mathrm{d}\mathbf{u}).$$

---

[1]**nomenclaure**: [n] a system of naming things, especially in a branch of science.

[2]Indeed, since $\mu^{\mathbf{y}}$, μ_0 are measures on $\mathbb{R}^n$ with densities $\rho^{\mathbf{y}}, \rho_0$, respectively, we have $\mathrm{d}\mu^{\mathbf{y}}(\mathbf{u}) = \rho^{\mathbf{y}}\mathrm{d}\mathbf{u}$, $\mathrm{d}\mu_0(\mathbf{u}) = \rho_0(\mathbf{u})\mathrm{d}\mathbf{u}$, and thus the LHS can be written as

$$\frac{\mathrm{d}\mu^{\mathbf{y}}}{\mathrm{d}\mu_0}(\mathbf{u}) = \frac{\rho^{\mathbf{y}}(\mathbf{u})\mathrm{d}\mathbf{u}}{\rho_0(\mathbf{u})\mathrm{d}\mathbf{u}} = \frac{\rho^{\mathbf{y}}(\mathbf{u})}{\rho_0(\mathbf{u})}.$$

By definition of the potential Φ, we have

$$Z = \int_{\mathbb{R}^n} \exp(-\Phi(\mathbf{u}; \mathbf{y}))\mu_0(\mathrm{d}\mathbf{u}) = \int_{\mathbb{R}^n} \exp(\log \rho(\mathbf{y} - G(\mathbf{u})))\mu_0(\mathrm{d}\mathbf{u}) = \int_{\mathbb{R}^n} \rho(\mathbf{y} - G(\mathbf{u}))\rho_0(\mathbf{u})\mathrm{d}\mathbf{u},$$

which is assumed to be positive to make sense of the denominator. The RHS can be written as

$$\frac{1}{Z}\exp(-\Phi(\mathbf{u}; \mathbf{y})) = \frac{1}{Z}\exp(\log \rho(\mathbf{y} - G(\mathbf{y}))) = \frac{1}{Z}\rho(\mathbf{y} - G(\mathbf{u})).$$

Multiplying both sides with $\rho_0(\mathbf{u})$ yields the desired equality in Theorem 1.3.1.

Thus $\boxed{\textit{the posterior is absolutely continuous w.r.t. the prior}}$ *, and* the Radon-Nikodym derivative is proportional to the likelihood.

This is rewriting Bayes' formula in the form

$$\frac{1}{\mathbb{P}(\mathbf{u})}\mathbb{P}(\mathbf{u}|\mathbf{y}) = \frac{1}{\mathbb{P}(\mathbf{y})}\mathbb{P}(\mathbf{y}|\mathbf{u}).$$

- *The expression for the Radon-Nikodym derivative is to be interpreted as the statement that, for all measurable* $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}^{\mu^{\mathbf{y}}} f(\mathbf{u}) = \mathbb{E}^{\mu_0}\left(\frac{\mathrm{d}\mu^{\mathbf{y}}}{\mathrm{d}\mu_0}(\mathbf{u})f(\mathbf{u})\right).$$

Alternatively we may write this in integral form as

$$\int_{\mathbb{R}^n} f(\mathbf{u})\mu^{\mathbf{y}}(\mathrm{d}\mathbf{u}) = \int_{\mathbb{R}^n}\left(\frac{1}{Z}\exp(-\Phi(\mathbf{u};\mathbf{y}))f(\mathbf{u})\right)\mu_0(\mathrm{d}\mathbf{u}) = \frac{\int_{\mathbb{R}^n}\exp(-\Phi(\mathbf{u};\mathbf{y}))f(\mathbf{u})\mu_0(\mathrm{d}\mathbf{u})}{\int_{\mathbb{R}^n}\exp(-\Phi(\mathbf{u};\mathbf{y}))\mu_0(\mathrm{d}\mathbf{u})}.$$

1.3.1   Inverse Heat Equation

This inverse problem illustrates the 1st difficulty, which motivates the Bayesian approach to inverse problems.

Let $D \subset \mathbb{R}^d$ be a bounded open set, with Lipschitz boundary ∂D.

Then define the Hilbert space H and operator A as follows:

$$H = (L^2(D), \langle \cdot, \cdot \rangle, \|\cdot\|); \;\; A = -\Delta, \;\; \mathcal{D}(A) = H^2(D) \cap H_0^1(D).$$

We make the following assumption about the spectrum of A which is easily verified for simple geometries, but in fact holds quite generally.

Assumption 1.3.1. *The eigenvalue problem* $A\varphi_j = \alpha_j\varphi_j$ *has a countably infinite set of solutions, indexed by* $j \in \mathbb{Z}^+$.

They may be normalized to satisfy the L^2*-orthonormality condition*

$$\langle \varphi_j, \varphi_k \rangle = \delta_{jk} := \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases}$$

and form a basis for H.

Furthermore, the eigenvalues are positive and, if ordered to be increasing, satisfy $\alpha_j \asymp j^{2/d}$.

The notation $\asymp$ (i.e., "asymptotic to") denotes the existence of constants $C^{\pm} > 0$ s.t.

$$C^- j^{2/d} \leq \alpha_j \leq C^+ j^{2/d}, \;\; \forall j \in \mathbb{N}.$$

Any $w \in H$ can be written as

$$w = \sum_{j=1}^{\infty}\langle w, \varphi_j \rangle \varphi_j,$$

and we can define the Hilbert scale of spaces $\mathcal{H}^t = \mathcal{D}(A^{t/2})$ as explained in Sect. A.1.3 for any $t > 0$ and with the norm

$$\|w\|_{\mathcal{H}^t}^2 = \sum_{j=1}^{\infty} j^{\frac{2t}{d}}|w_j|^2, \;\; \text{where } w_j = \langle w, \varphi_j \rangle.$$

Consider the *heat conduction equation* on D, with homogeneous Dirichlet boundary conditions (i.e., $v = 0$ on ∂D), writing it as an ODE in H:

$$\frac{\mathrm{d}v}{\mathrm{d}t} + Av = 0, \;\; v(0) = u. \tag{hc}$$

We have the following:

Lemma 1.3.1. *Let Assumption 1 hold. Then for every $u \in H$ and every $s > 0$, there is a unique solution v of equation* (hc) *in the space* $C([0,\infty); H) \cap C((0,\infty); \mathcal{H}^s)$. *We write* $v(t) = \exp(-At)u$.

To motivate this statement, and in particular the high degree of regularity seen at each fixed t, we argue as follows.

Note that, if the initial condition is expanded in the eigenbasis as

$$u = \sum_{j=1}^{\infty} u_j \varphi_j, \ u_j = \langle u, \varphi_j \rangle,$$

then the solution of (hc) has the form

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

Thus

$$\|v(t)\|_{\mathcal{H}^s}^2 = \sum_{j=1}^{\infty} j^{2s/d} e^{-2\alpha_j t} |u_j|^2 \asymp \sum_{j=1}^{\infty} \alpha_j^s e^{-2\alpha_j t} |u_j|^2 = t^{-s} \sum_{j=1}^{\infty} (\alpha_j t)^s e^{-2\alpha_j t} |u_j|^2 \le Ct^{-s} \sum_{j=1}^{\infty} |u_j|^2 = Ct^{-s} \|u\|_H^2.$$

It follows that $v(t) \in \mathcal{H}^s$ for any $s > 0$, provided $u \in H$.

We are interested in the inverse problem of finding u from y where

$$y = v(1) + \eta = G(u) + \eta = e^{-A}u + \eta.$$

Here $\eta \in H$ is noise and $G(u) := v(1) = e^{-A}u$.

Formally this looks like an infinite-dimensional linear version of the inverse problem (10.1), extended from finite dimensions to a Hilbert space setting.

However, the infinite-dimensional setting throws up significant new issues.

To see this, assume that there is $\beta_c > 0$ s.t. η has regularity $\mathcal{H}^\beta$ iff $\beta < \beta_c$.

Then y is not in the image space of G which is, of course, contained in $\bigcap_{s>0} \mathcal{H}^s$.

Applying the formal inverse of G to y results in an object which is not in H.

To overcome this problem, we will apply a Bayesian approach and hence will need to put probability measures on the Hilbert space H; in particular we will want to study $\mathbb{P}(u)$, $\mathbb{P}(y|u)$, and $\mathbb{P}(u|y)$, all probability measures on H.

1.4   Elliptic Inverse Problem

1 motivation for adopting the Bayesian approach to inverse problems is that prior modeling is a transparent approach to dealing with under-determined inverse problems; it forms a rational approach to dealing with the 2nd difficulty, labeled 2 in Sect. 1.1.

The elliptic inverse problem we now describe is a concrete example of an under-determined inverse problem.

As in Sect. 1.2, $D \subset \mathbb{R}^d$ denotes a bounded open set, with Lispchitz boundary ∂D.

We define the Gelfand triple of Hilbert spaces $V \subset H \subset V^*$ by **(10.5)**

$$H = (L^2(D), \langle \cdot, \cdot \rangle, \|\cdot\|), \ V = (H_0^1(D), \langle \nabla \cdot, \nabla \cdot \rangle, \|\cdot\|_V = \|\nabla \cdot \|),$$

and V^* the dual of V w.r.t. the pairing induced by H.

Note that $\|\cdot\| \le C_{\mathrm{p}} \|\cdot\|_V$ for some constant C_{p}: the Poincaré inequality.

Let $\kappa \in X := L^\infty(D)$ satisfy **(10.6)**

$$\operatorname{ess\,inf}_{\mathbf{x} \in D} \kappa(\mathbf{x}) = \kappa_{\min} > 0.$$

Now consider the equation

$$\begin{cases} -\nabla \cdot (\kappa \nabla p) = f, \ \mathbf{x} \in D, \\ \qquad\qquad p = 0, \ \mathbf{x} \in \partial D. \end{cases} \tag{elliptic}$$

Lax-Milgram theory yields the following:

Lemma 1.4.1. *Assume that $f \in V^*$ and that κ satisfies (10.6). Then (10.7) has a unique weak solution $p \in V$. This solution satisfies*

$$\|p\|_V \leq \frac{\|f\|_{V^*}}{\kappa_{\min}},$$

and, if $f \in H$,

$$\|p\|_V \leq C_p \frac{\|f\|}{\kappa_{\min}}.$$

We will be interested in the inverse problem of finding κ from y where **(10.8)**

$$y_j = l_j(p) + \eta_j, \ j = 1, \ldots, J.$$

Here $l_j \in V^*$ is a continuous linear functional on V and η_j is a noise.

Notice that the unknown, $\kappa \in X$, is a function (infinite dimensional), whereas the data from which we wish to determine κ is finite dimensional: $\mathbf{y} \in \mathbb{R}^J$.

The problem is severely under-determined, illustrating point 2 from Sect. 1.1.

1 way to treat such problems is by adopting the Bayesian framework, using prior modeling to fill in missing informat

We will take the unknown function to be u where either $u = \kappa$ or $u = \log \kappa$.

In either case, we will define $G_j(u) = l_j(p)$ and, noting that p is then a nonlinear function of u, (10.8) may be written as **(10.9)**

$$\mathbf{y} = G(u) + \boldsymbol{\eta},$$

where $\mathbf{y}, \boldsymbol{\eta} \in \mathbb{R}^J$ and $G : X^+ \subset X \to \mathbb{R}^J$.

The set X^+ is introduced because G may not be defined on the whole of X.

In particular, the positivity constraint (10.6) is only satisfied on **(10.10)**

$$X^+ := \{u \in X ; \operatorname{ess\,inf}_{\mathbf{x} \in D} u(\mathbf{x}) > 0\} \subset X$$

in the case where $\kappa = u$.

On the other hand, if $\kappa = \exp(u)$, then the positivity constraint (10.6) is satisfied for any $u \in X$ and we may take $X^+ = X$.

Notice that we again need probability measures on function space, here the Banach space $X = L^\infty(D)$.

Furthermore, in the case where $u = \kappa$, these probability measures should charge only positive functions, in view of the desired inequality (10.6).

Probability on Banach spaces of functions is most naturally developed in the setting of separable spaces, which $L^\infty(D)$ is not.

This difficulty can be circumvented in various different ways as we describe in what follows.

1.5   Prior Modeling

Goal. *Show how to construct probability measures on a function space, adopting a constructive approach based on random series.*

As explained in Dashti and Stuart, 2017, Sect. A.2.2, the natural setting for probability in a function space is that of a separable Banach space.

A countable infinite sequence in the Banach space X will be used for our random series; in the case where X is not separable, the resulting probability measure will be constructed on a separable subspace X' of X.

We denote the prior measures constructed in this section by μ_0.

1.5.1   General setting

We let $\{\phi_j\}_{j=1}^\infty$ denote an infinite sequence in the Banach space X, with norm $\|\cdot\|$, of $\mathbb{R}$-valued functions defined on a domain D.

   We will either take $D \subset \mathbb{R}^d$, a bounded, open set with Lipschitz boundary or $D = \mathbb{T}^d$ the d-dimensional torus.
   We normalize these functions so that $\|\phi_j\| = 1$ for $j \in \mathbb{Z}_{>0}$.
   We also introduce another element $m_0 \in X$, not necessarily normalized to 1.
   Define the function u by

$$u := m_0 + \sum_{j=1}^\infty u_j \phi_j. \tag{rand-func}$$

By randomizing u $:= \{u_j\}_{j=1}^\infty$, we create real-valued random functions on D.
   (The extension to $\mathbb{R}^n$-valued random functions is straightforward, but omitted for brevity.)
   We now define the deterministic sequence $\gamma = \{\gamma_j\}_{j=1}^\infty$ and the i.i.d. random sequence $\xi = \{\xi_j\}_{j=1}^\infty$, and set $u_j := \gamma_j \xi_j$.
   We assume that ξ is centered, i.e., that it has mean zero.
   Formally we see that the average value of u is then m_0 so that this element of X should be thought of as the *mean function*.
   We assume that $\gamma \in l_w^p$ for some $p \in [1, \infty)$ and some positive weight sequence $\{w_j\}$ (see Dashti and Stuart, 2017, Sect. A.1.1).
   We define $\Omega = \mathbb{R}^\infty$ and view ξ as a random element in the probability space $(\Omega, \mathrm{B}(\Omega), \mathbb{P})$ of i.i.d. sequences equipped with the product σ-algebra; we let $\mathbb{E}$ denote expectation.
   This sigma algebra can be generated by cylinder sets if an appropriate distance d is defined on sequences.
   However, the distance d captures nothing of the properties of the random function u itself.
   For this reason we will be interested in the pushforward of the measure $\mathbb{P}$ on the measure space $(\Omega, \mathrm{B}(\Omega))$ into a measure μ on $(X', \mathrm{B}(X'))$, where X' is a separable Banach space and $\mathrm{B}(X')$ denotes its Borel σ-algebra.
   Sometimes X' will be the same as X but not always: the space X may not be separable; and, although we have stated the normalization of the ϕ_j in X, they may of course live in smaller spaces X', and u may do so too.
   For either of these reasons, X' may be a proper subspace of X.
   In dealing with the random series construction, we will also find it useful to consider the truncated random functions

$$u^N := m_0 + \sum_{j=1}^N u_j \phi_j, \ \ u_j := \gamma_j \xi_j.$$

1.5.2   Uniform Priors

To construct the random functions (rand-func), we take $X = L^\infty(D)$, choose the deterministic sequence $\gamma = \{\gamma_j\}_{j=1}^\infty \in l^1$ and specify the i.i.d. sequence $\xi = \{\xi_j\}_{j=1}^\infty$ by $\xi_j \sim U[-1, 1]$, uniform random variables on $[-1, 1]$ (see, e.g., Pope, 2000, Sect. 3.3).
   Assume further that there are finite, strictly positive constants $m_{\min}, m_{\max}$, and δ s.t.

$$\operatorname{ess\,inf}_{\mathbf{x} \in D} m_0(\mathbf{x}) \geq m_{\min}, \ \operatorname{ess\,sup}_{\mathbf{x} \in D} m_0(\mathbf{x}) \leq m_{\max}, \ \|\gamma\|_{l^1} = \frac{\delta}{1+\delta} m_{\min}.$$

The space X is not separable and so, instead, we work with the space X' found as the closure of the linear span of the functions $(m_0, \{\phi_j\}_{j=1}^\infty)$ w.r.t. the norm $\|\cdot\|_\infty$ on X.
   The Banach space $(X', \|\cdot\|_\infty)$ is separable.

Theorem 1.5.1. *The following holds $\mathbb{P}$-almost surely: the sequence of functions $\{u^N\}_{N=1}^\infty$ given by (10.12) is Cauchy in X', and the limiting function u given by (rand-func) satisfies*

$$\frac{1}{1+\delta} m_{\min} \leq u(\mathbf{x}) \leq m_{\max} + \frac{\delta}{1+\delta} m_{\min} \ a.e. \ \mathbf{x} \in D.$$

Proof. See Dashti and Stuart, 2017, pp. 322–323. □

Example 1.5.1. *Consider the random function* (rand-func) *as specified in this section.*
By Theorem 2 we have that, $\mathbb{P}$*-a.s.,* **(10.13)**

$$u(\mathbf{x}) \geq \frac{1}{1+\delta} m_{\min} > 0, \ \ a.e. \ \mathbf{x} \in D.$$

Set $\kappa = u$ *in the elliptic equation* (elliptic)*, so that the coefficient* κ *in the equation and the solution* p *are random variables on* $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{P})$.
Since (10.13) *holds* $\mathbb{P}$*-a.s., Lemma 2 shows that, again* $\mathbb{P}$*-a.s.,*

$$\|p\|_V \leq (1+\delta) \frac{\|f\|_{V^*}}{m_{\min}}.$$

Since the RHS is nonrandom, we have that for all $r \in \mathbb{Z}^+$ *the random variable* $p \in L^r_{\mathbb{P}}(\Omega; V)$:

$$\mathbb{E}\|p\|_V^r < \infty.$$

In fact $\mathbb{E} \exp(\alpha\|p\|_V^r) < \infty$ *for all* $r \in \mathbb{Z}^+$ *and* $\alpha \in (0, \infty)$.

Example 1.5.2 (1D Gaussian, linear example)**.** *Forward response operator* $g \in L(\mathbb{R}, \mathbb{R})$
Prior:

$$\mu_0 = \mathcal{N}(0, \sigma_0^2), \ \sigma_0 \in \mathbb{R}, \ Leb. \ dens. \ \rho_0(u) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{\|u\|^2}{\sigma_0^2}\right).$$

Noise:

$$\eta \sim \mathcal{N}(0, \gamma^2), \ \gamma \in \mathbb{R}, \ Leb. \ dens. \ \rho_0(\eta) = \frac{1}{\gamma\sqrt{2\pi}} \exp\left(-\frac{\|\eta\|^2}{2\gamma^2}\right).$$

Bayes' formula

$$\rho(u|y) = \frac{\rho(y|u)\rho(u)}{\int_{\mathbb{R}^2} \rho(y|u)\rho(u)\mathrm{d}u}.$$

Completing the square

$$\mu^y = \mathcal{N}\left(\frac{\sigma_0^2 g}{\gamma^2 + g^2\sigma_0^2}y, \sigma_0^2 - \frac{\sigma_0^4 g^2}{\gamma^2 + g^2\sigma_0^2}\right).$$

By assumption, we have for the prior density $\rho_0(u)$

$$\rho_0(u) \propto \exp\left(-\frac{1}{2\sigma_0^2}u^2\right),$$

and the noise has density

$$\rho(\eta) \propto \exp\left(-\frac{1}{2\gamma^2}\eta^2\right).$$

Thus, by Bayes' formula, the posterior is given by

$$\rho^y(u) \propto \exp\left(-\frac{1}{2\sigma_0^2}u^2 - \frac{1}{2\gamma^2}(y - gu)^2\right) = \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{g^2}{\gamma^2}\right)u^2 - 2\frac{gy}{\gamma^2}u + \frac{y^2}{u^2}\right).$$

Defining

$$a = \frac{1}{\sigma_0^2} + \frac{g^2}{\gamma^2}, \ b = \frac{gy}{\gamma^2}, \ c = \frac{y^2}{\gamma^2},$$

we are interested in constants m, K, σ, s.t.

$$\rho^\gamma(u) \propto \exp\left(-\frac{1}{2\sigma^2}(u-m)^2 + K\right).$$

By completing the square, we obtain

$$au^2 - 2bu + c = a\left(u^2 - 2\frac{b}{a}u + \frac{c}{a}\right) = a\left(u - \frac{b}{a}\right)^2 + c - \frac{b^2}{a},$$

and thus

$$\sigma = \frac{1}{a} = \frac{\sigma_0^2\gamma^2}{\gamma^2 + g^2\sigma_0^2}, \quad m = \frac{b}{a} = \frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2}y,$$

and for the constant K, which does not depend on u, we obtain

$$K = c - \frac{b^2}{a} = \frac{y^2}{\gamma^2} - \frac{\sigma_0^2 g^2 y^2}{y^4 + y^2 g^2 \sigma^2}.$$

Hence,

$$\rho^y(u) \propto \exp\left(-\frac{1}{2\left(\frac{\sigma_0^2\gamma^2}{\gamma^2 + g^2\sigma_0^2}\right)^2}\left(u - \frac{\sigma_0^2 g}{\gamma^2 + g^2\sigma_0^2}y\right)^2\right).$$

1.6   Estimators

- Maximum a posteriori estimate (MAP)

$$u_{\text{MAP}} = \arg\max_{u \in \mathbb{R}^n} \rho^y(u).$$

- Conditional mean (CM)

$$u_{\text{CM}} = \mathbb{E}(u|y) = \int_{\mathbb{R}^n} u\rho^y(u)\mathrm{d}u.$$

- Maximum likelihood estimate (ML)

$$u_{\text{ML}} = \arg\max_{u \in \mathbb{R}^n} \rho(y - \mathcal{G}(u)).$$

- Conditional variance (speed estimator)

$$\mathbb{V}(u|y) = \int_{\mathbb{R}^n} (u - u_{\text{CM}})(u - u_{\text{CM}})^\top \rho^y(u)\mathrm{d}u.$$

Example 1.6.1 (Linear Gaussian Case). ***Connection of MAP & CM & optimization.***
Forward response operator $A \in L(\mathbb{R}^n, \mathbb{R}^J)$
Prior: $\mu_0 = \mathcal{N}(m_0, C_0)$
Noise: $\eta \sim \mathcal{N}(0, \Gamma)$
Posterior:

$$\mu^y = \mathcal{N}(m, C) \text{ with } m = m_0 + C_0 A^*(AC_0 A^* + \Gamma)^{-1}(y - Am_0) \text{ and } C = C_0 - C_0 A^*(AC_0 A^* + \Gamma)^{-1}AC_0.$$

$$u_{\text{MAP}} = \arg\max_{u \in \mathbb{R}^n} \rho^y(u) = \arg\min_{u \in \mathbb{R}^n} (u - m)^\top C^{-1}(u - m) = m,$$
$$u_{\text{CM}} = \mathbb{E}(u|y) = m,$$
$$u_{\text{opt}} = \arg\min_{u \in \mathbb{R}^n} \|Ax - y\|_\Gamma^2 + \|u - m_0\|_{C_0}^2 = m.$$

15

Example 1.6.2 (Comparison of CM & MAP (1D))**.** *Let u be a real-valued rv and assume that the posterior density is given by*

$$\rho^y(u) = \frac{\alpha}{\sigma_0}\varphi\left(\frac{u}{\sigma_0}\right) + \frac{1-\alpha}{\sigma_1}\varphi\left(\frac{u-1}{\sigma_1}\right) \ \text{ with } 0 < \alpha < 1, \ \sigma_0, \sigma_1 > 0 \ \text{ and } \varphi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right).$$

Estimators.

$$u_{\mathrm{CM}} = 1 - \alpha, \ \ u_{\mathrm{MAP}} = \begin{cases} 0, & \text{if } \dfrac{\alpha}{\sigma_0} > \dfrac{1-\alpha}{\sigma_1}, \\[2ex] 1, & \text{if } \dfrac{\alpha}{\sigma_0} < \dfrac{1-\alpha}{\sigma_1}, \\[2ex] \{0,1\}, & \text{if } \dfrac{\alpha}{\sigma_0} = \dfrac{1-\alpha}{\sigma_1}. \end{cases}$$

Chapter 2

Prior Modeling

In this section the authors show how to construct probability measures on a function space, adopting a constructive approach based on random series.

As explained in Sect. A.2.2, the natural setting for probability in a function space is that of a separable Banach space.

A countable infinite sequence in the Banach space X will be used for our random series; in the case where X is not separable, the resulting probability measure will be constructed on a separable subspace X' of X (see the discussion in Sect. 2.1).

Sect. 2.1 describes this general setting, and Sects. 2.2–2.4 consider, in turn, 3 classes of priors termed uniform, Besov and Gaussian.

In Sect. 2.5 the authors link the random series construction to the widely used random field perspective on spatial stochastic processes and the authors summarize in Sect. 2.6.

We denote the prior measures constructed in this section by μ_0.

2.1 General Setting

We let $\{\phi_j\}_{j=1}^\infty$ denote an infinite sequence in the Banach space X, with norm $\|\cdot\|$, of $\mathbb{R}$-valued functions defined on a domain D.

We will either take $D \subset \mathbb{R}^d$, a bounded, open set with Lipschitz boundary or $D = \mathbb{T}^d$ the d-dimensional torus.

We normalize these functions so that $\|\phi_j\| = 1$ for $j = 1, \ldots, \infty$.

We also introduce another element $m_0 \in X$, not necessarily normalized to 1.

Define the function u by **(10.11)**

$$u = m_0 + \sum_{j=1}^\infty u_j \phi_j.$$

By randomizing $\mathrm{u} := \{u_j\}_{j=1}^\infty$, we create real-valued random functions on D.

(The extension to $\mathbb{R}^n$-valued random functions is straightforward, but omitted for brevity.)

We now define the deterministic sequence $\gamma = \{\gamma_j\}_{j=1}^\infty$ and the i.i.d. random sequence $\xi = \{\xi_j\}_{j=1}^\infty$, and set $u_j = \gamma_j \xi_j$.

We assume that ξ is centered, i.e., that it has mean zero.

Formally we see that the average value of u is then m_0 so that this element of X should be thought of as the *mean function*.

We assume that $\gamma \in l_w^p$ for some $p \in [1, \infty)$ and some positive weight sequence $\{w_j\}$ (see Sect. A.1.1).

We define $\Omega = \mathbb{R}^\infty$ and view ξ as a random element in the probability space $(\Omega, \mathrm{B}(\Omega), \mathbb{P})$ of i.i.d. sequences equipped with the product σ-algebra; we let $\mathbb{E}$ denote expectation.

This sigma algebra can be generated by cylinder sets if an appropriate distance d is defined on sequences.

However, the distance d captures nothing of the properties of the random function u itself.

For this reason we will be interested in the pushforward of the measure $\mathbb{P}$ on the measure space $(\Omega, \mathrm{B}(\Omega))$ into a measure μ on $(X', \mathrm{B}(X'))$, where X' is a separable Banach space and $\mathrm{B}(X')$ denotes its Borel σ-algebra.

Sometimes X' will be the same as X but not always: the space X may not be separable; and, although we have stated the normalization of the ϕ_j in X, they may of course live in smaller spaces X', and u may do so too.

For either of these reasons, X' may be a proper subspace of X.

In the next 3 subsections, the authors demonstrate how this general setting may be adapted to create a variety of useful prior measures on function space; the 4th subsection, which follows these 3, relates the random series construction, in the Gaussian case, to the standard construction of Gaussian random fields.

Dashti & Stuart will express many of their results in terms of the probability measure $\mathbb{P}$ i.i.d. sequences, but all such results will, of course, have direct implications for the induced pushforward measures on the function spaces where the random functions u live.

We discuss this perspective in the summary Sect. 2.6.

In dealing with the random series construction, we will also find it useful to consider the truncated random functions **(10.12)**

$$u^N = m_0 + \sum_{j=1}^{N} u_j \phi_j, \ \ u_j = \gamma_j \xi_j.$$

2.2   Uniform Priors

To construct the random functions (10.11), we take $X = L^\infty(D)$, choose the deterministic sequence $\gamma = \{\gamma_j\}_{j=1}^\infty \in l^1$ and specify the i.i.d. sequence $\xi = \{\xi_j\}_{j=1}^\infty$ by $\xi_j \sim U[-1, 1]$, uniform random variables on $[-1, 1]$.

Assume further that there are finite, strictly positive constants $m_{\min}, m_{\max}$, and δ s.t.

$$\operatorname{ess\,inf}_{\mathbf{x} \in D} m_0(\mathbf{x}) \geq m_{\min}, \ \operatorname{ess\,sup}_{\mathbf{x} \in D} m_0(\mathbf{x}) \leq m_{\max}, \ \|\gamma\|_{l^1} = \frac{\delta}{1+\delta} m_{\min}.$$

The space X is not separable and so, instead, we work with the space X' found as the closure of the linear span of the functions $(m_0, \{\phi_j\}_{j=1}^\infty)$ w.r.t. the norm $\|\cdot\|_\infty$ on X.

The Banach space $(X', \|\cdot\|_\infty)$ is separable.

Theorem 2.2.1. *The following holds $\mathbb{P}$-almost surely: the sequence of functions $\{u^N\}_{N=1}^\infty$ given by (10.12) is Cauchy in X', and the limiting function u given by (10.11) satisfies*

$$\frac{1}{1+\delta} m_{\min} \leq u(\mathbf{x}) \leq m_{\max} + \frac{\delta}{1+\delta} m_{\min} \ a.e. \ \mathbf{x} \in D.$$

Proof. See Dashti and Stuart, 2017, pp. 322–323. □

Example 2.2.1. *Consider the random function (10.11) as specified in this section.*

By Theorem 2 we have that, $\mathbb{P}$-a.s., **(10.13)**

$$u(\mathbf{x}) \geq \frac{1}{1+\delta} m_{\min} > 0, \ \ a.e. \ \mathbf{x} \in D.$$

Set $\kappa = u$ in the elliptic equation (10.6), so that the coefficient κ in the equation and the solution p are random variables on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{P})$.

Since (10.13) holds $\mathbb{P}$-a.s., Lemma 2 shows that, again $\mathbb{P}$-a.s.,

$$\|p\|_V \leq (1+\delta) \frac{\|f\|_{V^*}}{m_{\min}}.$$

Since the RHS is nonrandom, we have that for all $r \in \mathbb{Z}^+$ the random variable $p \in L^r_{\mathbb{P}}(\Omega; V)$:

$$\mathbb{E}\|p\|_V^r < \infty.$$

In fact $\mathbb{E}\exp(\alpha\|p\|_V^r) < \infty$ for all $r \in \mathbb{Z}^+$ and $\alpha \in (0, \infty)$.

We now consider the situation where the family $\{\phi_j\}_{j=1}^\infty$ has a uniform Hölder exponent α and study the implications for Hölder continuity of the random function u.

Specially we assume that there $C, a > 0$ and $\alpha \in (0,1]$ s.t., for all $j \geq 1$, **(10.14)–(10.15)**

$$|\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})| \leq Cj^a|\mathbf{x} - \mathbf{y}|^a, \ \mathbf{x}, \mathbf{y} \in D,$$
$$|m_0(\mathbf{x}) - m_0(\mathbf{y})| \leq C|\mathbf{x} - \mathbf{y}|^\alpha, \ \mathbf{x}, \mathbf{y} \in D.$$

Theorem 2.2.2. *Assume that u is given by (10.11) where the collection of functions $(m_0, \{\phi_j\}_{j=1}^\infty)$ satisfy (10.14) and (10.15). Assume further that $\sum_{j=1}^\infty |\gamma_j|^2 j^{a\theta} < \infty$ for some $\theta \in (0,2)$. Then $\mathbb{P}$-a.s. we have $u \in C^{0,\beta}(D)$ for all $\beta < \frac{a\theta}{2}$.*

Proof. See Dashti and Stuart, 2017, p. 324. $\square$

Example 2.2.2. *Let $\{\phi_j\}$ denote the Fourier basis for $L^2(D)$ with $D = [0,1]^d$.*
Then we may take $a = \alpha = 1$.
If $\gamma_j = j^{-s}$, then $s > 1$ ensures $\gamma \in l^1$.
Furthermore,

$$\sum_{j=1}^\infty |\gamma_j|^2 j^{a\theta} = \sum_j j^{\theta-2s} < \infty, \ for \ \theta < 2s - 1.$$

We thus deduce that $u \in C^{0,\beta}([0,1]^d)$ for all $\beta < \min\{s - \frac{1}{2}, 1\}$.

2.3 Besov Priors

For this construction of random functions, we take X to be the Hilbert space

$$X := \dot{L}^2(\mathbb{T}^d) = \left\{u : \mathbb{T}^d \to \mathbb{R}; \int_{\mathbb{R}^d} |u(\mathbf{x})|^2 d\mathbf{x} < \infty, \int_{\mathbb{T}^d} u(\mathbf{x})d\mathbf{x} = 0\right\}$$

of real valued periodic functions in dimension $d \leq 3$ with inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, resp.

We then set $m_0 = 0$ and let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis for X.

Consequently, for any $u \in X$, we have for a.e. $\mathbf{x} \in \mathbb{T}^d$, **(10.16)**

$$u(\mathbf{x}) = \sum_{j=1}^\infty u_j\phi_j(\mathbf{x}), \ u_j = \langle u, \phi_j \rangle.$$

Given a function $u : \mathbb{T}^d \to \mathbb{R}$ and the $\{u_j\}$ as defined in (10.16), we define the Banach space $X^{t,q}$ by

$$X^{t,q} = \left\{u : \mathbb{T}^d \to \mathbb{R}; \|u\|_{X^{t,q}} < \infty, \int_{\mathbb{T}^d} u(\mathbf{x})d\mathbf{x} = 0\right\},$$

where

$$\|u\|_{X^{t,q}} = \left(\sum_{j=1}^\infty j^{\frac{tq}{d}+\frac{q}{2}-1}|u_j|^q\right)^{\frac{1}{q}}, \ \text{with } q \in [1,\infty), \ t > 0.$$

If $\{\phi_j\}$ form the Fourier basis and $q = 2$, then $X^{t,2}$ is the Sobolev space $\dot{H}^t(\mathbb{T}^d)$ of mean-zero periodic functions with t (possibly non-integer) square-integrable derivatives; in particular $X^{0,2} = \dot{L}^2(\mathbb{T}^d)$.

On the other hand, if the $\{\phi_j\}$ form certain wavelet bases, then $X^{t,q}$ is the *Besov space* B_{qq}^t.

As described above, we assume that $u_j = \gamma_j\xi_j$ where $\xi = \{\xi_j\}_{j=1}^\infty$ is an i.i.d. sequence and $\gamma = \{\gamma_j\}_{j=1}^\infty$ is deterministic.

Here we assume that $\xi_j 1$ is drawn from the centered measure on $\mathbb{R}$ with density proportional to $\exp\left(-\frac{1}{2}|x|^q\right)$ for some $1 \leq q < \infty$ - we refer to this as a *q-exponential distribution*, noting that $q = 2$ gives a Gaussian and $q = 1$ a Laplace-distributed random variable.

Then for $s > 0$ and $\delta > 0$, we define **(10.17)**

$$\gamma_j = j^{-\left(\frac{s}{d}+\frac{1}{2}-\frac{1}{q}\right)}\left(\frac{1}{\delta}\right)^{\frac{1}{q}}.$$

The parameter δ is a key scaling parameter which will appear in the statement of exponential moment bounds below.

We now prove convergence of the series (found from (10.12) with $m_0 = 0$) **(10.18)**

$$u^N = \sum_{j=1}^{N} u_j \phi_j, \ u_j = \gamma_j \xi_j$$

to the limit function **(10.19)**

$$u(\mathbf{x}) = \sum_{j=1}^{\infty} u_j \phi_j(\mathbf{x}), \ u_j = \gamma_j \xi_j,$$

in an appropriate space.

To understand the sequence of functions $\{u^N\}$, it is useful to introduce the following function space:

$$L_{\mathbb{P}}^q(\Omega; X^{t,q}) := \left\{v : D \times \Omega \to \mathbb{R}; \mathbb{E}(\|v\|_{X^{t,q}}^q) < \infty\right\}.$$

This is a Banach space, when equipped with the norm $(\mathbb{E}(\|v\|_{X^{t,q}})^q)^{1/q}$.

Thus every Cauchy sequence is convergent in this space.

Theorem 2.3.1. *For $t < s - \frac{d}{q}$, the sequence of functions $\{u^N\}_{N=1}^{\infty}$, given by (10.18) and (10.17) with ξ_1 drawn from a centered q-exponential distribution, is Cauchy in the Banach space $L_{\mathbb{P}}^q(\Omega; X^{t,q})$. Thus the infinite series (10.19) exists as an $L_{\mathbb{P}}^q$-limit and takes values in $X^{t,q}$ almost surely, for all $t < s - \frac{d}{q}$.*

Proof. See Dashti and Stuart, 2017, p. 326. □

The previous theorem gives a sufficient condition, on t, for existence of the limiting random function.

The following theorem refines this to an iff statement, in the context of almost sure convergence.

Theorem 2.3.2. *Assume that u is given by (10.19) and (10.17) with ξ_1 drawn from a centered q-exponential distribution. Then the following are equivalent:*

(i) $\|u\|_{X^{t,q}} < \infty$ $\mathbb{P}$-a.s.;

(ii) $\mathbb{E}(\exp(\alpha\|u\|_{X^{t,q}}^q)) < \infty$ for any $\alpha \in [0, \frac{\delta}{2})$;

(iii) $t < s - \frac{d}{q}$.

Proof. See Dashti and Stuart, 2017, pp. 326–328. □

Lemma 2.3.1. *Let $\{I_j\}_{j=1}^{\infty}$ be an independent sequence of $\mathbb{R}^+$-valued random variables. Then*

$$\sum_{j=1}^{\infty} I_j < \infty \ a.s. \Leftrightarrow \sum_{j=1}^{\infty} \mathbb{E}(I_j \wedge 1) < \infty.$$

As in the previous subsection, we now study the situation where the family $\{\phi_j\}$ has a uniform Hölder exponent α and study the implications for Hölder continuity of the random function u.

In this case, however, the basis functions are normalized in $\dot{L}^2$ and not L^∞; thus we must make additional assumptions on the possible growth of the L^∞ norms of $\{\phi_j\}$ with j.

We assume that there are $C, a, b > 0$ and $\alpha \in (0, 1]$ s.t., for all $j \geq 0$, **(10.20)**

$$|\phi_j(\mathbf{x})| = \beta_j \leq Cj^b, \ \mathbf{x} \in D,$$
$$|\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})| \leq Cj^a|\mathbf{x} - \mathbf{y}|^\alpha, \ \mathbf{x}, \mathbf{y} \in D.$$

We also assume that $a > b$ as, since $\|\phi_j\|_{L^2} = 1$, it is natural that the pre-multiplication constant in the Hölder estimate on the $\{\phi_j\}$ grows in j at least as fast as the bound on the functions themselves.

Theorem 2.3.3. *Assume that u is given by (10.19) and (10.17) with ξ_1 drawn from a centered q-exponential distribution. Suppose also that (10.20) hold and that $s > d(b + q^{-1} + \frac{1}{2}\theta(a - b))$ for some $\theta \in (0, 2)$. Then $\mathbb{P}$-a.s. we have $u \in C^{0,\beta}(\mathbb{T}^d)$ for all $\beta < \frac{\alpha\theta}{2}$.*

Proof. See Dashti and Stuart, 2017, p. 329. ∎

We note that the result of Theorem 6 holds true when the mean function is nonzero if it satisfies

$$|m_0(\mathbf{x})| \leq C, \ \mathbf{x} \in D,$$
$$|m_0(\mathbf{x}) - m_0(\mathbf{y})| \leq C|\mathbf{x} - \mathbf{y}|^\alpha, \ \mathbf{x}, \mathbf{y} \in D.$$

We have the following sharper result if the family $\{\phi_j\}$ is regular enough to be a basis for B_{qq}^t instead of satisfying (10.20):

Theorem 2.3.4. *Assume that u is given by (10.19) and (10.17) with ξ_1 drawn from a centered q-exponential distribution. Suppose also that $\{\phi_j\}_{j\in\mathbb{N}}$ form a basis for B_{qq}^t for some $t < s - \frac{d}{q}$. Then $u \in C^{0,t}(\mathbb{T}^d)$ $\mathbb{P}$-almost surely.*

Proof. See Dashti and Stuart, 2017, p. 329. ∎

If the mean function m_0 is t-Hölder continuous, the result of the above theorem holds for a random series with nonzero mean function as well.

2.4   Gaussian Priors

Let X be a Hilbert space $\mathcal{H}$ of real-valued functions on bounded open $D \subset \mathbb{R}^d$ with Lipschitz boundary and with inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, resp.; e.g., $\mathcal{H} = L^2(D; \mathbb{R})$.

Assume that $\{\phi_j\}_{j=1}^\infty$ is an orthonormal basis for $\mathcal{H}$.

We study the Gaussian case where $\xi_1 \sim N(0, 1)$, and then equation (10.11) with $u_j = \gamma_j \xi_j$ generates random draws from the Gaussian measure $N(m_0, \mathcal{C})$ on $\mathcal{H}$ where the covariance operator $\mathcal{C}$ depends on the sequence $\gamma = \{\gamma_j\}_{j=1}^\infty$.

See the Appendix for background on Gaussian measures in a Hilbert space.

As in Sect. 2.3, we consider the setting in which $m_0 = 0$ so that the function u is given by (10.16) and has mean zero.

We thus focus on identifying $\mathcal{C}$ from the random series (10.16) and studying the regularity of random draws from $N(0, \mathcal{C})$.

Define the Hilbert scale of spaces $\mathcal{H}^t$ as in Sect. A.1.3 with, recall, norm

$$\|u\|_{\mathcal{H}^t}^2 = \sum_{j=1}^\infty j^{\frac{2t}{d}} |u_j|^2.$$

We choose $\xi_1 \sim N(0, 1)$ and study convergence of the series (10.18) for u^N to a limit function u given by (10.19); the spaces in which this convergence occurs will depend upon the sequence γ.

To understand the sequence of functions $\{u^N\}$, it is useful to introduce the following function space:

$$L^2_{\mathbb{P}}(\Omega; \mathcal{H}^t) := \left\{ v : D \times \Omega \to \mathbb{R}; \mathbb{E}(\|v\|_{\mathcal{H}^t})^2 < \infty \right\}.$$

This is in fact a Hilbert space, although we will not use the Hilbert space structure.

We will only use the fact that $L^2_{\mathbb{P}}$ is a Banach space when equipped with the norm $(\mathbb{E}(\|v\|^2_{\mathcal{H}^t}))^{1/2}$ and that hence every Cauchy sequence is convergent.

Theorem 2.4.1. *Assume that $\gamma_j \asymp j^{-s/d}$. Then the sequence of functions $\{u^N\}_{N=1}^{\infty}$ given by (10.18) is Cauchy in the Hilbert space $L^2_{\mathbb{P}}(\Omega; \mathcal{H}^t)$, $t < s - \frac{d}{2}$. Thus, the infinite series (10.19) exists as an $L^2_{\mathbb{P}}$ limit and takes values in $\mathcal{H}^t$ almost surely, for $t < s - \frac{d}{2}$.*

Proof. See Dashti and Stuart, 2017, p. 331. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Remark 2.4.1. *We make the following remarks concerning the Gaussian random functions constructed in the preceding theorem.*

- *The preceding theorem shows that the sum (10.18) has an $L^2_{\mathbb{P}}$ limit in $\mathcal{H}^t$ when $t < s - \frac{d}{2}$, as one can also see from the following direct calculation*

$$\mathbb{E}\|u\|^2_{\mathcal{H}^t} = \sum_{j=1}^{\infty} j^{\frac{2t}{d}} \mathbb{E}(\gamma_j^2 \xi_j^2) = \sum_{j=1}^{\infty} j^{\frac{2t}{d}} \gamma_j^2 \asymp \sum_{j=1}^{\infty} j^{\frac{2(t-s)}{d}} < \infty.$$

Thus $u \in \mathcal{H}^t$ a.s., for $t < s - \frac{d}{2}$.

- *From the preceding theorem, we see that, provided $s > \frac{d}{2}$, the random function in (10.19) generates a mean-zero Gaussian measure on $\mathcal{H}$.*

  The expression (10.19) is known as the Karhunen-Loève expansion *and the eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ as the* Karhunen-Loève basis.

- *The covariance operator $\mathcal{C}$ of a measure μ on $\mathcal{H}$ may then be viewed as a bounded linear operator from $\mathcal{H}$ into itself defined to satisfy **(10.24)***

$$\mathcal{C}l = \int_{\mathcal{H}} \langle l, u \rangle u \mu(\mathrm{d}u), \ \forall l \in \mathcal{H}.$$

*Thus, **(10.25)***

$$\mathcal{C} = \int_{\mathcal{H}} u \otimes u \mu(\mathrm{d}u).$$

The following formal calculation, which can be made rigorous if $\mathcal{C}$ is trace class on $\mathcal{H}$, gives an expression for the covariance operator:

$$\mathcal{C} = \mathbb{E}u \otimes u = \mathbb{E}\left(\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \gamma_j \gamma_k \xi_j \xi_k \phi_j \otimes \phi_k \right) = \left(\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \gamma_j \gamma_k \mathbb{E}(\xi_j \xi_k) \phi_j \otimes \phi_k \right)$$

$$= \left(\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \gamma_j \gamma_k \delta_{jk} \phi_j \otimes \phi_k \right) = \sum_{j=1}^{\infty} \gamma_j^2 \phi_j \otimes \phi_j.$$

From this expression for the covariance, we may find eigenpairs explicitly:

$$\mathcal{C}\phi_k = \left(\sum_{j=1}^{\infty} \gamma_j^2 \phi_j \otimes \phi_j \right) \phi_k = \sum_{j=1}^{\infty} \gamma_j^2 \langle \phi_j, \phi_k \rangle \phi_j = \sum_{j=1}^{\infty} \gamma_j^2 \delta_{jk} \phi_k = \gamma_k^2 \phi_k.$$

- *The Gaussian measure is denoted by $\mu_0 := N(0, \mathcal{C})$, a Gaussian with mean function 0 and covariance operator $\mathcal{C}$.*

  The eigenfunctions of $\mathcal{C}$, $\{\phi_j\}_{j=1}^\infty$, are known as the Karhunen-Loève basis *for measure μ_0.*

  The γ_j^2 are the eigenvalues associated with this eigenbasis, and thus γ_j is the standard deviation of the Gaussian measure in the direction ϕ_j.

In the case where $\mathcal{H} = \dot{L}^2(\mathbb{T}^d)$, we are in the setting of Sect. 2.3 and we briefly consider this case.

We assume that the $\{\phi_j\}_{j=1}^\infty$ constitute the Fourier basis.

Let $A = -\Delta$ denote the negative Laplacian equipped with periodic boundary conditions on $[0,1)^d$ and restricted to functions which integrate to zero over $[0,1)^d$.

This operator is positive self-adjoint and has eigenvalues which grow like $j^{2/d}$, analogously to Assumption 1 made in the case of Dirichlet boundary conditions.

It then follows that $\mathcal{H}^t = \mathcal{D}(A^{t/2}) = \dot{H}^t(\mathbb{T}^d)$, the Sobolev space of periodic functions on $[0,1)^d$ with spatial mean equal to zero and t (possibly negative or fractional) square integrable derivatives.

Thus, by the preceding Remarks 2, u defined by (10.19) is in the space $\dot{H}^t$ a.s., $t < s - \frac{d}{2}$.

In fact we can say more about regularity, using the Kolmogorov continuity test and Corollary 4; this we now do.

Theorem 2.4.2. *Consider the Karhunen-Loève expansion (10.19) so that u is a sample from the measure $N(0, \mathcal{C})$ in the case where $\mathcal{C} = A^{-s}$ with $A = -\Delta$, $\mathcal{D}(A) = \dot{H}^2(\mathbb{T}^d)$ and $s > \frac{d}{2}$. Then, $\mathbb{P}$-a.s., $u \in \dot{H}^t$, $t < s - \frac{d}{2}$, and $u \in C^{0,t}(\mathbb{T}^d)$ a.s., $t < 1 \wedge (s - \frac{d}{2})$.*

Proof. See Dashti and Stuart, 2017, p. 333. □

The previous example illustrates the fact that, although we have constructed Gaussian measures in a Hilbert space setting, and that they are naturally defined on a range of Hilbert (Sobolev-like) spaces defined through fractional powers of the Laplacian, they may also be defined on Banach spaces, e.g. the space of Hölder continuous functions.

We now return to the setting of the general domain D, rather than the d-dimensional torus.

In this general context, it is important to highlight the Fernique theorem, here restated from the Appendix because of its importance:

Theorem 2.4.3 (Fernique Theorem). *Let μ_0 be a Gaussian measure on the separable Banach space X. Then there exists $\beta_c \in (0, \infty)$ s.t., for all $\beta \in (0, \beta_c)$,*

$$\mathbb{E}^{\mu_0} \exp(\beta \|u\|_X^2) < \infty.$$

Remark 2.4.2. *We make 2 remarks concerning the Fernique theorem.*

- *Theorem 10, when combined with Theorem 9, shows that, with β sufficiently small, $\mathbb{E}^{\mu_0} \exp(\beta \|u\|_X^2) < \infty$ for both $X = \dot{H}^t$ and $X = C^{0,t}(\mathbb{T}^d)$, if $t < s - \frac{d}{2}$.*

- *Let $\mu_0 = N(0, A^{-s})$ where A is as in Theorem 9.*

  Then Theorem 5 proves the Fernique theorem 10 for $X = X^{t,2} = \dot{H}^t$, if $t < s - \frac{d}{2}$; the proof in the case of the torus is very different from the general proof of the result in the abstract setting of Theorem 10.

- *Theorem 5(ii) gives, in the Gaussian case, the Fernique theorem in the case that X is the Hilbert space $X^{t,2}$.*

  Furthermore, the constant β_c is specified explicitly in that setting.

  More explicit versions of the general Fernique Theorem 10 are possible, but the characterization of β_c is more involved.

Example 2.4.1. *Consider the random function (10.11) in the case where $\mathcal{H} = \dot{L}^2(\mathbb{T}^d)$ and $\mu_0 = N(0, A^{-s})$, $s > \frac{d}{2}$ as in the preceding example.*

Then we know that, μ_0-a.s., $u \in C^{0,t}$, $t < 1 \wedge \left(s - \frac{d}{2}\right)$.

Set $\kappa = e^u$ in the elliptic PDE (10.7) so that the coefficient κ and hence the solution p are random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Then $\kappa_{\min}$ given in (10.6) satisfies

$$\kappa_{\min} \geq \exp(-\|u\|_\infty).$$

By Lemma 2 we obtain

$$\|p\|_V \leq \exp(\|u\|_\infty)\|f\|_{V^*}.$$

Since $C^{0,t} \subset L^\infty(\mathbb{T}^d)$, $t \in (0,1)$, we deduce that

$$\|u\|_{L^\infty} \leq K_1 \|u\|_{C^{0,t}}.$$

Furthermore, for any $\epsilon > 0$, there is constant $K_2 = K_2(\epsilon)$ s.t. $\exp(K_1 r x) \leq K_2 \exp(\epsilon x^2)$ for all $x \geq 0$.

Thus

$$\|p\|_V^r \leq \exp(K_1 r \|u\|_{C^{0,t}})\|f\|_{V^*}^r \leq K_2 \exp(\epsilon \|u\|_{C^{0,t}}^2)\|f\|_{V^*}^t.$$

Hence, by Theorem 10, we deduce that

$$\mathbb{E}\|p\|_V^r < \infty, \ \ i.e. \ \ p \in L_\mathbb{P}^r(\Omega; V), \ \ \forall r \in \mathbb{Z}^+.$$

This result holds for any $r \geq 0$.

Thus, when the coefficient of the elliptic PDE is $\log$ normal, i.e., κ is the exponential of a Gaussian function, moments of all orders exist for the random variable p.

However, unlike the case of the uniform prior, we cannot obtain exponential moments on $\mathbb{E}\exp(\alpha\|p\|_V^r)$ for any $(r, \alpha) \in \mathbb{Z}^+ \times (0, \infty)$.

This is because the coefficient κ, while positive a.s., does not satisfy a uniform positive lower bound across the probability space.

2.5   Random Field Perspective

In this subsection the authors link the preceding constructions of random functions, through randomized series, to the notion of *random fields*.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with expectation denoted by $\mathbb{E}$, and $D \subseteq \mathbb{R}^d$ an open set.

For the random series constructions developed in the preceding subsections, $\Omega = \mathbb{R}^\infty$ and $\mathcal{F} = \mathrm{B}(\Omega)$; however, the development of the general theory of random fields does not require this specific choice.

A *random field* on D is a measurable mapping $u : D \times \Omega \to \mathbb{R}^n$.

Thus, for any $\mathbf{x} \in D$, $u(\mathbf{x}; \cdot)$ is an $\mathbb{R}^n$-valued random variable; on the other hand, for any $\omega \in \Omega$, $u(\cdot; \omega) : D \to \mathbb{R}^n$ is a vector field.

In the construction of random fields, it is commonplace to 1st construct the *finite-dimensional distributions*.

These are found by choosing any integer $K \geq 1$, and any set of points $\{x_k\}_{k=1}^K$ in D, and then considering the random vector $(u(x_1, \cdot)^*, \ldots, u(x_K; \cdot)^*)^* \in \mathbb{R}^{nK}$.

From the finite-dimensional distributions of this collection of random vectors, we would like to be able to make sense of the probability measure μ on X, a separable Banach space equipped with the Borel σ-algebra $\mathrm{B}(X)$, via the formula **(10.26)**

$$\mu(A) = \mathbb{P}(u(\cdot; \omega) \in A), \ \ A \in \mathrm{B}(X),$$

where ω is taken from a common probability space on which the random element $u \in X$ is defined.

It is thus necessary to study the joint distribution of a set of K $\mathbb{R}^n$-valued random variables, all on a common probability space.

Such $\mathbb{R}^{nK}$-valued random variables are, of course, only defined up to a set of zero measure.

It is desirable that all such finite-dimensional distributions are defined on a common subset $\Omega_0 \subset \Omega$ with full measure, so that u may be viewed as a function $u : D \times \Omega_0 \to \mathbb{R}^n$; such a choice of random field is termed a *modification*.

When reinterpreting the previous subsections in terms of random fields, statements about almost sure (regularity) properties should be viewed as statements concerning the existence of a modification possessing of the stated almost sure regularity property.

We may define the space of functions

$$L^q_{\mathbb{P}}(\Omega; X) := \left\{ v : D \times \Omega \to \mathbb{R}^n; \mathbb{E}(\|v\|^q_X) < \infty \right\}.$$

This is a Banach space, when equipped with the norm $(\mathbb{E}(\|v\|^q_X))^{1/q}$.

We have used such spaces in the preceding subsections when demonstrating convergence of the randomized series.

Note that we often simply write $u(x)$, suppressing the explicit dependence on the probability space.

A *Gaussian random field* is one where, for any integer $K \geq 1$, and any set of points $\{x_k\}_{k=1}^K$ in D, the random vector $(u(x_1, \cdot)^*, \ldots, u(x_K; \cdot)^*)^* \in \mathbb{R}^{nK}$ is a Gaussian random vector.

The *mean function* of a Gaussian random field is $m(x) = \mathbb{E}u(x)$.

The *covariance function* is $c(x, y) = \mathbb{E}(u(x) - m(x))(u(y) - m(y))^*$.

For Gaussian random fields, the *mean function* $m : D \to \mathbb{R}^n$ and the *covariance function* $c : D \times D \to \mathbb{R}^{n \times n}$ together completely specify the joint probability distribution for $(u(x_1; \cdot)^*, \ldots, u(x_K)^*)^* \in \mathbb{R}^{nK}$.

Furthermore, if we view the Gaussian random field as a Gaussian measure on $L^2(D; \mathbb{R}^n)$, then the covariance operator can be constructed from the covariance function as follows.

W.l.o.g., we consider the mean-zero case; the more general case follows by shift of origin.

Since the field has mean zero, we have, from (10.24), that for all $h_1, h_2 \in L^2(D; \mathbb{R}^n)$,

$$\langle h_1, \mathcal{C}h_2 \rangle = \mathbb{E}\langle h_1, u \rangle \langle u, h_2 \rangle = \mathbb{E} \int_D \int_D h_1(\mathbf{x})^* (u(\mathbf{x})u(\mathbf{y})^*) h_2(\mathbf{y}) \mathrm{d}\mathbf{y} \mathrm{d}\mathbf{x}$$

$$= \mathbb{E} \int_D h_1(\mathbf{x})^* \left(\int_D (u(\mathbf{x})u(\mathbf{y})^*) h_2(\mathbf{y}) \mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x} = \int_D h_1(\mathbf{x})^* \left(\int_D c(\mathbf{x}, \mathbf{y}) h_2(\mathbf{y}) \mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x},$$

and we deduce that, for all $\psi \in L^2(D; \mathbb{R}^n)$, **(10.27)**

$$(\mathcal{C}\psi)(\mathbf{x}) = \int_D c(\mathbf{x}, \mathbf{y}) \psi(\mathbf{y}) \mathrm{d}\mathbf{y}.$$

Thus the covariance operator of a Gaussian random field is an integral operator with kernel given by the covariance function.

As such we may also view the covariance function as the Green's function of the inverse covariance, or *precision*.

A mean-zero Gaussian random field is termed *stationary* if $c(\mathbf{x}, \mathbf{y}) = s(\mathbf{x} - \mathbf{y})$ for some matrix-valued function s, so that shifting the field by a fixed random vector does not change the statistics.

It is *isotropic* if it is stationary and, in addition, $s(\cdot) = \imath(|\cdot|)$, for some matrix-valued function $\imath$.

In the previous subsection, the authors demonstrated how the regularity of random fields maybe established from the properties of the sequences γ (deterministic, with decay) and ξ (i.i.d. random).

Here we show similar results but express them in terms of properties of the covariance function and covariance operator.

Theorem 2.5.1. *Consider an $\mathbb{R}^n$-valued Gaussian random field u on $D \subset \mathbb{R}^d$ with mean zero and with isotropic correlation function $c : D \times D \to \mathbb{R}^{n \times n}$. Assume that D is bounded and that $\operatorname{Tr} c(\mathbf{x}, \mathbf{y}) = k(|\mathbf{x} - \mathbf{y}|)$ where $k : \mathbb{R}^+ \to \mathbb{R}$ is Hölder with any exponent $\alpha \leq 1$. Then u is almost surely Hölder continuous on D with any exponent smaller than $\frac{1}{2}\alpha$.*

Proof. See Dashti and Stuart, 2017, p. 337. □

It is often convenient both algorithmically and theoretically to define the covariance operator through fractional inverse powers of a differential operator.

Indeed in the previous subsection, we showed that our assumptions on the random series construction we used could be interpreted as having a covariance operator which was an inverse fractional power of the Laplacian on zero spatial average functions with periodic boundary conditions.

We now generalize this perspective and consider covariance operators which are a fractional power of an operator A satisfying the following.

Assumption 2.5.1. *The operator A, densely defined on the Hilbert space $\mathcal{H} = L^2(D; \mathbb{R}^n)$, satisfies the following properties:*

1. *A is positive definite, self-adjoint and invertible;*

2. *the eigenfunctions $\{\phi_j\}_{j \in \mathbb{N}}$ of A form an orthonormal basis for $\mathcal{H}$;*

3. *the eigenvalues of A satisfy $\alpha_j \asymp j^{2/d}$;*

4. *there is $C > 0$ s.t.*

$$\sup_{j \in \mathbb{N}} \left(\|\phi_j\|_{L^\infty} + \frac{1}{j^{\frac{1}{d}}} \mathrm{Lip}(\phi_j) \right) \leq C.$$

These properties are satisfied by the Laplacian on a torus, when applied to functions with spatial mean zero.

But they are in fact satisfied for a much wider range of differential operators which are *Laplacian-like*.

E.g., the Dirichlet Laplacian on a bounded open set D in $\mathbb{R}^d$, together with various Laplacian operators perturbed by lower order terms, e.g., Schrödinger operators.

Inspection of the proof of Theorem 9 reveals that it only uses the properties of Assumption 2.

Thus we have:

Theorem 2.5.2. *Let u be a sample from the measure $N(0, \mathcal{C})$ in the case where $\mathcal{C} = A^{-s}$ with A satisfying Assumptions 2 and $s > \frac{d}{2}$. Then, $\mathbb{P}$-a.s., $u \in \dot{H}^t$, for $t < s - \frac{d}{2}$, and $u \in C^{0,t}(D)$, for $t < 1 \wedge \left(s - \frac{d}{2}\right)$.*

Example 2.5.1. *Consider the case $d = 2$, $n = 1$ and $D = [0, 1]^2$.*

Define the Gaussian random field through the measure $\mu = N(0, (-\Delta)^{-\alpha})$ where Δ is the Laplacian with domain $H_0^1(D) \cap H^2(D)$.

Then Assumption 2 is satisfied by $-\Delta$.

By Theorem 12 it follows that choosing $\alpha > 1$ suffices to ensure that draws from μ are almost surely in $L^2(D)$. It also follows that, in fact, draws from μ are almost surely in $C(D)$.

2.6 Summary

In the preceding 4 subsections, the authors have shown how to create random functions by randomizing the coefficients of a series of functions.

Using these random series, the authors have also studied the regularity properties of the resulting functions.

Furthermore we have extended our perspective in the Gaussian case to determine regularity properties from the properties of the covariance function or the covariance operator.

For the uniform prior, the authors have shown that the random functions all live in a subset of $X = L^\infty$ characterized by the upper and lower bounds given in Theorem 2 and found as the closure of the linear span of the set of functions $(m_0, \{\phi_j\}_{j=1}^\infty)$; denote this subset, which is a separable Banach space, by X'.

For the Besov priors, the authors have shown in Theorem 5 that the random functions live in the separable Banach spaces $X^{t,q}$ for all $t < s - \frac{d}{q}$; denote any 1 of these Banach spaces by X'.

And finally for the Gaussian priors, the authors have shown in Theorem 8 that the random function exists as an L^2 limit in any of the Hilbert spaces $\mathcal{H}^t$ for $t < s - \frac{d}{2}$.

Furthermore, the authors have indicated that, by use of the Kolmogorov continuity theorem, we can also show that the Gaussian random functions lie in certain Hölder spaces; these Hölder spaces are not separable but, by

the discussion in Sect. A.1.2, we can embed the spaces $C^{0,\gamma'}$ in the separable uniform Hölder spaces $C_0^{0,\gamma}$ for any $\gamma < \gamma'$; since the upper bound on the range of Hölder exponents established by use of Kolmogorov continuity theorem is open, this means we can work in the same range of Hölder exponents, but restricted to uniform Hölder spaces, thereby regaiing separability.

In this Gaussian case, we denote any of the separable Hilbert or Banach spaces where the Gaussian random function lives almost surely by X'.

Thus, in all of these examples, we have created a probability measure μ_0 which is the pushforward of the measure $\mathbb{P}$ on the i.i.d. sequence ξ under the map which takes the sequence into the random function.

The resulting measure lives on the separable Banach space X', and we will often write $\mu_0(X') = 1$ to denote this fact.

This is shorthand for saying that functions drawn from μ_0 are in X' almost surely.

Separability of X' naturally leads to the use of the Borel σ-algebra to define a canonical measurable space and to the development of an integration theory - Bochner integration - which is natural on this space; see Sect. A.2.2.

Chapter 3

Posterior Distribution

In this section the authors prove a Bayes' theorem appropriate for combining a likelihood with prior measures on separable Banach spaces as constructed in the previous section.

In Sect. 3.1, the authors start with some general remarks about conditioned random variables.

Sect. 3.2 contains the authors' statement and proof of a Bayes' theorem and specifically its application to Bayesian inversion.

Dashti & Stuart note here that, in their setting, the posterior μ^y will always be absolutely continuous w.r.t. the prior and they use the standard notation $\mu^y \ll \mu_0$ to denote this.

It is possible to construct examples, e.g., in the purely Gaussian setting, where the posterior is not absolutely continuous w.r.t. the prior.

Thus, it is certainly not necessary to work in the setting where $\mu^y \ll \mu_0$.

However, it is quite natural, from a modeling point of view, to work in this setting: absolute continuity ensures that almost sure properties built into the prior will be inherited by the posterior.

For these almost sure properties to be changed by the data would require that the data contains an infinite amount of information, something which is unnatural in most applications.

In Sect. 3.3, the authors study the example of the heat equation, introduced in Sect. 1.2, from the perspective of Bayesian inversion, and in Sect. 3.4 they do the same for the elliptic inverse problem of Sect. 1.3.

3.1   Conditional Random Variables

Key to the development of Bayes' theorem, and the posterior distribution, is the notion of conditional random variables.

In this section the authors state an important theorem concerning conditioning.

Let (X, A) and (Y, B) denote a pair of measurable spaces, and let ν and π be probability measures on $X \times Y$.

We assume that $\nu \ll \tau$.

Thus there exists π-measurable $\phi : X \times Y \to \mathbb{R}$ with $\phi \in L^1_\pi$ (see Sect. A.1.4 for definition of L^1_π) and **(10.28)**

$$\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x, y) = \phi(x, y).$$

I.e., for $(x, y) \in X \times Y$,

$$\mathbb{E}^\nu f(x, y) = \mathbb{E}^\pi(\phi(x, y), f(x, y)),$$

or, equivalently,

$$\int_{X \times Y} f(x, y)\nu(\mathrm{d}x, \mathrm{d}y) = \int_{X \times Y} \phi(x, y)f(x, y)\pi(\mathrm{d}x, \mathrm{d}y).$$

Theorem 3.1.1. *Assume that the condition random variable $x|y$ exists under π with probability distribution denoted $\pi^y(\mathrm{d}x)$. Then the condition random variable $x|y$ under ν exists, with probability distribution denoted by*

$\nu^y(\mathrm{d}x)$. *Furthermore,* $\nu^y \ll \pi^y$ *and if* $c(y) := \int_X \phi(x,y)\mathrm{d}\pi^y(x) > 0$, *then*

$$\frac{\mathrm{d}\nu^y}{\mathrm{d}\pi^y}(x) = \frac{1}{c(y)}\phi(x,y).$$

Example 3.1.1. *Let* $X = C([0,1];\mathbb{R})$, $Y = \mathbb{R}$.

Let π *denote the measure on* $X \times Y$ *induced by the random variable* $(w(\cdot), w(1))$, *where* w *is a draw from standard unit Wiener measure on* $\mathbb{R}$, *starting from* $w(0) = z$.

Let π^y *denote measure on* X *found by conditioning Brownian motion to satisfy* $w(1) = y$, *thus* π^y *is a Brownian bridge measure with* $w(0) = z$, $w(1) = y$.

Assume further that

$$\sup_{x \in X} \Phi(x,y) = \Phi^+(y) < \infty, \ \forall y \in \mathbb{R}.$$

Then

$$c(y) = \int_{\mathbb{R}} \exp(-\phi(x,y))\mathrm{d}\pi^y(x) > \exp(-\Phi^+(y)) > 0.$$

Thus $\nu^y(\mathrm{d}x)$ *exists and*

$$\frac{\mathrm{d}\nu^y}{\mathrm{d}\pi^y}(x) = \frac{1}{c(y)}\exp(-\Phi(x,y)).$$

We will use the preceding theorem to go from a construction of the joint probability distribution on unknown and data to the conditional distribution of the unknown, given data.

In constructing the joint probability distribution, we will need to establish measurability of the likelihood, for which the following will be useful:

Lemma 3.1.1. *Let* (Z, B) *be a Borel measurable topological space and assume that* $G \in C(Z;\mathbb{R})$ *and that* $\pi(Z) = 1$ *for some probability measure* π *on* (Z, B). *Then* G *is a* π*-measurable function.*

3.2   Bayes' Theorem for Inverse Problems

Let X and Y be separable Banach spaces, equipped with the Borel σ-algebra, and $G : X \to Y$ a measurable mapping.

We wish to solve the inverse problem of finding u from y where **(10.29)**

$$y = G(u) + \eta,$$

and $\eta \in Y$ denotes *noise*.

We employ a Bayesian approach to this problem in which we let $(u, y) \in X \times Y$ be a random variable and compute $u|y$.

We specify the random variable (u, y) as follows:

- **Prior**: $u \sim \mu_0$ measure on X.

- **Noise**: $\eta \sim \mathbb{Q}_0$ measure on Y, and (recalling that $\perp$ denotes independence) $\eta \perp u$.

The random variable $y|u$ is then distributed according to the measure $\mathbb{Q}_u$, the translate of $\mathbb{Q}_0$ by $G(u)$.

We *assume* throughout the following that $\mathbb{Q}_u \ll \mathbb{Q}_0$ for u μ_0-a.s.

Thus, for some *potential* $\Phi : X \times Y \to \mathbb{R}$, **(10.30)**

$$\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) = \exp(-\Phi(u;y)).$$

Thus, for fixed u, $\Phi(u;\cdot) : Y \to \mathbb{R}$ is measurable and $\mathbb{E}^{\mathbb{Q}_0}\exp(-\Phi(u;y)) = 1$.

For given instance of the data y, $-\Phi(\cdot; y)$ is termed the *log likelihood*.

Define ν_0 to be the product measure **(10.31)**

$$\nu_0(\mathrm{d}u, \mathrm{d}y) = \mu_0(\mathrm{d}u)\mathbb{Q}_0(\mathrm{d}y).$$

We *assume* in what follows that $\Phi(\cdot, \cdot)$ is ν_0 measurable.

Then the random variable $(u, y) \in X \times Y$ is distributed according to measure $\nu(\mathrm{d}u, \mathrm{d}y) = \mu_0(\mathrm{d}u)\mathbb{Q}_u(\mathrm{d}y)$.

Furthermore, it then follows that $\nu \ll \nu_0$ with

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(u, y) = \exp(-\Phi(u; y)).$$

We have the following infinite-dimensional analogue of Theorem 1.

Theorem 3.2.1 (Bayes' Theorem). *Assume that $\Phi : X \times Y \to \mathbb{R}$ is ν_0 measurable and that, for y $\mathbb{Q}_0$-a.s.,* **(10.32)**

$$Z := \int_X \exp(-\Phi(u; y))\mu_0(\mathrm{d}u) > 0.$$

Then the conditional distribution of $u|y$ exists under ν and is denoted by μ^y. Furthermore $\mu^y \ll \mu_0$ and, for y ν-a.s., **(10.33)**

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z}\exp(-\Phi(u; y)).$$

Proof. See Dashti and Stuart, 2017, p. 343. $\square$

Remark 3.2.1. *In order to implement the derivation of Bayes' formula (10.33), 4 essential steps are required:*

- *Define a suitable prior measure μ_0 and noise measure $\mathbb{Q}_0$ whose independent product form the reference measure ν_0.*

- *Determine the potential Φ s.t. formula (10.30) holds.*

- *Show that Φ is ν_0 measurable.*

- *Show that the normalization constant Z given by (10.32) is positive almost surely w.r.t. $y \sim \mathbb{Q}_0$.*

We will show how to carry out this program for 2 examples in the following subsections.

The following remark will also be used in studying 1 of the examples.

Remark 3.2.2. *The following comments on the setup above may be useful.*

- *In formula (10.33) we can shift $\Phi(u, y)$ by any constant $c(y)$, independent of u, provided the constant is finite $\mathbb{Q}_0$-a.s. and hence ν-a.s.*

  Such a shift can be absorbed into a redefinition of the normalization constant Z.

- *Our Bayes' theorem only asserts that the posterior is absolutely continuous w.r.t. the prior μ_0.*

  In fact equivalence (mutual absolute continuity) will occur when $\Phi(\cdot; y)$ is finite everywhere in X.

3.3   Heat Equation

We apply Bayesian inversion to the heat equation from Sect. 1.2.

Recall that for $G(u) = e^{-A}u$, we have the relationship

$$y = G(u) + \eta,$$

which we wish to invert.

Let $X = H$ and define

$$\mathcal{H}^t = \mathcal{D}(A^{t/2}) = \left\{ w; w = A^{-t/2}w_0, \ w_0 \in H \right\}.$$

Under Assumption 1, we have $\alpha_j \asymp j^{2/d}$ so that this family of spaces is identical with the Hilbert scale of spaces $\mathcal{H}^t$ as defined in Sects. 1.2 and 2.4.

We choose the prior $\mu_0 = N(0, A^{-\alpha})$, $\alpha > \frac{d}{2}$.

Thus $\mu_0(X) = \mu_0(H) = 1$.

Indeed the analysis in Sect. 2.4 shows that $\mu_0(\mathcal{H}^t) = 1$, $t < \alpha - \frac{d}{2}$.

For the likelihood we assume that $\eta \perp u$ with $\eta \sim \mathbb{Q}_0 = N(0, A^{-\beta})$, and $\beta \in \mathbb{R}$.

This measure satisfies $\mathbb{Q}_0(\mathcal{H}^t) = 1$ for $t < \beta - \frac{d}{2}$ and we thus choose $Y = \mathcal{H}^{t'}$ for some $t' < \beta - \frac{d}{2}$.

Notice that the authors' analysis includes the case of white observational noise, for which $\beta = 0$.

The Cameron-Martin theorem 32, together with the fact that $e^{-\lambda A}$ commutes with arbitrary fractional powers of A, can be used to show that $y|u \sim \mathbb{Q}_u := N(G(u), A^{-\beta})$ where $\mathbb{Q}_u \ll \mathbb{Q}_0$ with

$$\frac{d\mathbb{Q}_u}{d\mathbb{Q}_0}(y) = \exp(-\Phi(u; y)),$$

and

$$\Phi(u; y) = \frac{1}{2}\|A^{\frac{\beta}{2}}e^{-A}u\|^2 - \langle A^{\frac{\beta}{2}}e^{-\frac{A}{2}}y, A^{\frac{\beta}{2}}e^{-\frac{A}{2}}u \rangle.$$

In the following we repeatedly use the fact that $A^{\gamma}e^{-\lambda A}$, $\lambda > 0$, is a bounded linear operator from $\mathcal{H}^a$ to $\mathcal{H}^b$, any $a, b, \gamma \in \mathbb{R}$.

Recall that $\nu_0(\mathrm{d}u, \mathrm{d}y) = \mu_0(\mathrm{d}u)\mathbb{Q}_0(\mathrm{d}y)$.

Note that $\nu_0(H \times \mathcal{H}^{t'}) = 1$.

Using the boundedness of $A^{\gamma}e^{-\lambda A}$, it may be shown that $\Phi : H \times \mathcal{H}^{t'} \to \mathbb{R}$ is continuous and hence ν_0-measurable by Lemma 4.

Theorem 14 shows that the posterior is given by μ^y where

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z}\exp(-\Phi(u; y)), \ \ Z = \int_H \exp(-\Phi(u; y))\mu_0(\mathrm{d}u),$$

provided that $Z > 0$ for y $\mathbb{Q}_0$-a.s.

We establish this positivity in the remainder of the proof.

Since $y \in \mathcal{H}^t$ for any $t < \beta - \frac{d}{2}$, $\mathbb{Q}_0$-a.s., we have that $y = A^{-t'/2}w_0$ for some $w_0 \in H$ and $t' < \beta - \frac{d}{2}$.

Thus we may write **(10.34)**

$$\Phi(u; y) = \frac{1}{2}\|A^{\frac{\beta}{2}}e^{-A}u\|^2 - \langle A^{\frac{\beta-t'}{2}}e^{-\frac{A}{2}}w_0, A^{\frac{\beta}{2}}e^{-\frac{A}{2}}u \rangle.$$

Then, using the boundedness of $A^{\gamma}e^{-\lambda A}$, $\lambda > 0$, together with (10.34), we have

$$\Phi(u; y) \leq C(\|u\|^2 + \|w_0\|^2),$$

where $\|w_0\|$ is finite $\mathbb{Q}_0$-a.s.

Thus,

$$Z \geq \int_{\|u\|^2 \leq 1} \exp(-C(1 + \|w_0\|^2))\mu_0(\mathrm{d}u),$$

and, since $\mu_0(\|u\|^2 \leq 1) > 0$ (by Theorem 33 all balls have positive measure for Gaussians on a separable Banach space), the required positivity follows.

3.4   Elliptic Inverse Problem

We consider the elliptic inverse problem from Sect. 1.3 from the Bayesian perspective.

We consider the use of both uniform and Gaussian priors.

Before studying the inverse problem, however, it is important to derive some continuity properties of the forward problem.

Throughout this section, we consider equation (10.7) under the assumption that $f \in V^*$.

3.4.1   Forward Problem

Recall that in Sect. 1.3, equation (10.10), we defined **(10.35)**

$$X^+ = \{v \in L^\infty(D); \operatorname{ess\,inf}_{\mathbf{x} \in D} v(\mathbf{x}) > 0\}.$$

Then the map $\mathcal{R} : X^+ \to V$ by $\mathcal{R}(\kappa) = p$.

This map is well-defined by Lemma 2 and we have the following result.

Lemma 3.4.1. *For $i = 1, 2$, let*

$$-\nabla \cdot (\kappa_i \nabla p_i) = f, \ \mathbf{x} \in D,$$
$$p_i = 0, \ \mathbf{x} \in \partial D.$$

Then

$$\|p_1 - p_2\|_V \leq \frac{1}{\kappa_{\min}^2} \|f\|_{V^*} \|\kappa_1 - \kappa_2\|_{L^\infty},$$

where we assume that

$$\kappa_{\min} := \operatorname{ess\,inf}_{\mathbf{x} \in D} \kappa_1(\mathbf{x}) \wedge \operatorname{ess\,inf}_{\mathbf{x} \in D} \kappa_2(\mathbf{x}) > 0.$$

Thus the function $\mathcal{R} : X^+ \to V$ is locally Lipschitz.

Proof. See Dashti and Stuart, 2017, p. 346. □

3.4.2   Uniform Priors

We now study the inverse problem of finding κ from a finite set of continuous linear functionals $\{l_j\}_{j=1}^J$ on V, representing measurements of p; thus $l_j \in V^*$.

To match the notation from Sect. 3.2, we take $\kappa = u$ and we define the separable Banach space X' as in Sect. 2.2.

It is straightforward to see that Lemma 5 extends to the case where X^+ given by (10.35) is replaced by **(10.36)**

$$X^+ = \{v \in X'; \operatorname{ess\,inf}_{\mathbf{x} \in D} v(\mathbf{x}) > 0\}$$

since $X' \subset L^\infty(D)$.

When considering uniform priors for the elliptic problem, we work with this definition of X^+.

We define $G : X^+ \to \mathbb{R}^J$ by

$$G_j(u) = l_j(\mathcal{R}(u)), \ j = 1, \ldots, J,$$

where, recall, the l_j are elements of V^*: bounded linear functionals on V.

Then $G(u) = (G_1(u), \ldots, G_J(u))$ and we are interested in the inverse problem of finding $u \in X^+$ from y where $y = G(u) + \eta$ and η is the noise.

We assume $\eta \sim N(0, \Gamma)$, for positive symmetric $\Gamma \in \mathbb{R}^{J \times J}$.

(Use of other statistical assumptions on η is a straightforward extension of what follows whenever η has a smooth density on $\mathbb{R}^J$.)

Let μ_0 denote the prior measure constructed in Sect. 2.2.

Then μ_0-almost surely we have, by Theorem 2, **(10.37)**

$$u \in X_0^+ := \left\{ v \in X'; \frac{1}{1+\delta} m_{\min} \leq v(\mathbf{x}) \leq m_{\max} + \frac{\delta}{1+\delta} m_{\min} \text{ a.e. } \mathbf{x} \in D \right\}.$$

Thus $\mu_0(X_0^+) = 1$.

The likelihood is defined as follows.

Since $\eta \sim N(0, \Gamma)$, it follows that $\mathbb{Q}_0 = N(0, \Gamma)$, $\mathbb{Q}_u = N(G(u), \Gamma)$ and

$$\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) = \exp(-\Phi(u; y)), \ \ \Phi(u; y) = \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2 - \frac{1}{2}|\Gamma^{-\frac{1}{2}}y|^2.$$

Recall that $\nu_0(\mathrm{d}y, \mathrm{d}u) = \mathbb{Q}_0(\mathrm{d}y)\mu_0(\mathrm{d}u)$.

Since $G : X^+ \to \mathbb{R}^J$ is locally Lipschitz by Lemma 5, Lemma 4 implies that $\Phi : X^+ \times Y \to \mathbb{R}$ is ν_0-measurable.

Thus Theorem 14 shows that $u|y \sim \mu^y$ where **(10.38)**

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z} \exp(-\Phi(u; y)), \ \ Z = \int_{X^+} \exp(-\Phi(u; y))\mu_0(\mathrm{d}u),$$

provided $Z > 0$ for y $\mathbb{Q}_0$-almost surely.

To see that $Z > 0$, note that

$$Z = \int_{X_0^+} \exp(-\Phi(u; y))\mu_0(\mathrm{d}u),$$

since $\mu_0(X_0^+) = 1$.

On X_0^+ we have that $\mathcal{R}(\cdot)$ is bounded in V, and hence G is bounded in $\mathbb{R}^J$.

Furthermore y is finite $\mathbb{Q}_0$-almost surely.

Thus $\mathbb{Q}_0$-almost surely w.r.t. to y, $\Phi(\cdot; y)$ is bounded on X_0^+; we denote the resulting bound by $M = M(y) < \infty$.

Hence

$$Z \geq \int_{X_0^+} \exp(-M)\mu_0(\mathrm{d}u) = \exp(-M) > 0,$$

and the result is proved.

We may use Remark 5 to shift Φ by $\frac{1}{2}|\Gamma^{-1/2}y|^2$, since this is almost surely finite under $\mathbb{Q}_0$ and hence under $\nu(\mathrm{d}u, \mathrm{d}y) = \mathbb{Q}_u(\mathrm{d}y)\mu_0(\mathrm{d}u)$.

We then obtain the equivalent form for the posterior distribution μ^y: **(10.39)**

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z} \exp\left(-\frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2\right), \ \ Z = \int_X \exp\left(-\frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2\right)\mu_0(\mathrm{d}u).$$

3.4.3 Gaussian Priors

Dashti & Stuart conclude this subsection by discussing the same inverse problem, but using Gaussian priors from Sect. 2.4.

We now set $X = C(\overline{D})$ and $Y = \mathbb{R}^J$ and we note that X embeds continuously into $L^\infty(D)$.

We assume that we can find an operator A which satisfies Assumption 2.

We now take $\kappa = \exp(u)$ and define $G : X \to \mathbb{R}^J$ by

$$G_j(u) = l_j(\mathcal{R}(\exp(u))), \ j = 1, \ldots, J.$$

We take as prior on u the measure $N(0, A^{-s})$ with $s > \frac{d}{2}$.

Then Theorem 12 shows that $\mu(X) = 1$.

The likelihood is unchanged by the prior, since it concerns y given u, and is hence identical to that in the case of the uniform prior, although the mean shift from $\mathbb{Q}_0$ to $\mathbb{Q}_u$ by $G(u)$ now has a different interpretation since $\kappa = \exp(u)$ rather than $\kappa = u$.

Thus we again obtain (10.38) for the posterior distribution (albeit with a different definition of $G(u)$) provided that we can establish that, $\mathbb{Q}_0$-a.s.,

$$Z = \int_X \exp\left(\frac{1}{2}|\Gamma^{-\frac{1}{2}}y|^2 - \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2\right)\mu_0(\mathrm{d}u) > 0.$$

To this end, we use the fact that the unit ball in X, denoted B, has positive measure by Theorem 33 and that on this ball $\mathcal{R}(\exp(u))$ is bounded in V by $e^{-a}\|f\|_{V^*}$, by Lemma 2, for some finite positive constant a.

This follows from the continuous embedding of X into L^∞ and since the infimum of $\kappa = \exp(u)$ is bounded below by $e^{-\|u\|_{L^\infty}}$.

Thus G is bounded on B and, noting that y is $\mathbb{Q}_0$-a.s. finite, we have for some $M = M(y) < \infty$,

$$\sup_{u \in B}\left(\frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2 - \frac{1}{2}|\Gamma^{-\frac{1}{2}}y|^2\right) < M.$$

Hence

$$Z \geq \int_B \exp(-M)\mu_0(\mathrm{d}u) = \exp(-M)\mu_0(B) > 0,$$

since all balls have positive measure for Gaussian measure on a separable Banach space.

Thus we again obtain (10.39) for the posterior measure, now with the new definition of G, and hence Φ.

Chapter 4

Common Structures

4.1 Wellposedness

Ito and Jin, 2015, Chap. 3: "Inverse problems suffer from instability, which poses significantly challenges to their stable and accurate numerical solution. Therefore, specialized techniques are required."

Example 4.1.1. *Consider again the heat conduction equation*

$$
\begin{cases}
v_t + Av = 0, & in \ (0,T) \times \Omega, \\
v = 0, & on \ [0,T] \times \partial\Omega, \\
v(0, \cdot) = u, & in \ \Omega,
\end{cases}
$$

where $A = -\Delta$, $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$, and both perfect data $y = e^{-A}u$, derived from the forward model with no noise, and noisy data $y' = e^{-A}u + \eta$.

Consider the case where $\eta = \epsilon\varphi_j$ with $\epsilon \in (0,1)$ small and φ_j a normalized eigenfunction of A.

Thus $\|\eta\|_{L^2(\Omega)} = \epsilon$.

Obviously application of the inverse of e^{-A} to y returns the point u which gave rise to the perfect data.

It is natural to apply the inverse of e^{-A} to both y and to y' to understand the effect of the noise.

Doing so yields the identity

$$
\|e^A y - e^A y'\|_{L^2(\Omega)} = \|e^A(y-y')\|_{L^2(\Omega)} = \|e^A \eta\|_{L^2(\Omega)} = \epsilon\|e^A \varphi_j\|_{L^2(\Omega)} = \epsilon e^{\alpha_j}.
$$

Recall Assumption 1 which gives $\alpha_j \asymp j^{\frac{2}{d}} \to \infty$ as $j \to \infty$.

Now fix any $a > 0$ and choose j large enough to ensure that $\alpha_j \geq (a+1)\log(\epsilon^{-1})$.

It then follows that $\|y - y'\|_{L^2(\Omega)} = \epsilon$ while $\|e^A y - e^A y'\|_{L^2(\Omega)} = \epsilon e^{(a+1)\log(\epsilon^{-1})} = \epsilon\epsilon^{-(a+1)} = \epsilon^{-a}$.

This is a manifestation of ill posedness.

Furthermore, since $a > 0$ is arbitrary, the ill posedness can be made arbitrarily bad by considering $a \to \infty$ (then $\epsilon^{-a} \to \infty$).

Aim. To show that *this ill-posedness affect does not occur in the Bayesian posterior distribution: small changes in the data y lead to small changes in the measure μ^y.*

Let X and Y be separable Banach spaces, equipped with the Borel σ-algebra, and μ_0 a measure on X.

We will work under assumptions which enable us to make sense of the following measure $\mu^y \ll \mu_0$ defined, for some $\Phi : X \times Y \to \mathbb{R}$, by

$$
\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z(y)}\exp(-\Phi(u;y)), \ Z(y) := \int_X \exp(-\Phi(u;y))\mu_0(du).
$$

Question 4.1.1. *Why does this formula make sense?* → Require: Integrability of the integrand of $Z(y)$ (and thus a lower bound of the likelihood $\Phi(u;y)$), nonzero-being (and thus positivity) of the denominator $Z(y)$.

We make the following assumptions concerning Φ:

Assumption 4.1.1. *Let $X' \subseteq X$ and assume that $\Phi \in C(X' \times Y; \mathbb{R})$. Assume further that there are functions $M_1 : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $M_2 : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{> 0}$,[1] monotonic non-decreasing separately in each argument, s.t. for all $u \in X'$, $y, y_1, y_2 \in B_Y(0, r)$,*

$$\Phi(u; y) \geq -M_1(r, \|u\|_X),$$
$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq M_2(r, \|u\|_X)\|y_1 - y_2\|_Y.$$

Question 4.1.2. *Why do we need the Lipschitzian assumption of $\Phi(u; y)$ w.r.t. the 2nd argument?*

In order to measure the effect of changes in y on the measure μ^y, we need a metric on measures.
Dashti & Stuart use the Hellinger metric defined in Dashti and Stuart, 2017, Sect. A.2.4.

A brief summary of Hellinger metric & its properties. Assume that we have 2 probability measures μ and μ' on a separable Banach space denoted by B (actually the considerations here apply on a Polish space but we do not need this level of generality).

Assume that μ and μ' are both absolutely continuous w.r.t. a common reference measure ν, also defined on the same measure space, e.g., $\nu = \frac{1}{2}(\mu + \mu')$.

The resulting metrics that we define are independent of the choice of this common reference measure.

Definition 4.1.1. *The* total variation distance *between μ and μ' is*

$$d_{\mathrm{TV}}(\mu, \mu') := \frac{1}{2} \int_B \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu} \right| \mathrm{d}\nu.$$

In particular, if μ' is absolutely continuous w.r.t. μ, then

$$d_{\mathrm{TV}}(\mu, \mu') := \frac{1}{2} \int_B \left| 1 - \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \right| \mathrm{d}\mu.$$

Definition 4.1.2. *The* Hellinger distance *between μ and μ' is*

$$d_{\mathrm{Hell}}(\mu, \mu') := \sqrt{\frac{1}{2} \int_B \left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}} \right)^2 \mathrm{d}\nu}.$$

In particular, if μ' is absolutely continuous w.r.t. μ, then

$$d_{\mathrm{Hell}}(\mu, \mu') = \sqrt{\frac{1}{2} \int_B \left(1 - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}} \right)^2 \mathrm{d}\mu}.$$

Note that the numerical constant $\frac{1}{2}$ appearing in both definitions is chosen in such a way as to ensure the bounds

$$0 \leq d_{\mathrm{TV}}(\mu, \mu') \leq 1, \ 0 \leq d_{\mathrm{Hell}}(\mu, \mu') \leq 1.$$

The Hellinger and total variation distances are related as follows, which shows in particular that they both generate the same topology:

Lemma 4.1.1. *The total variation and Hellinger metrics are related by the inequalities*

$$\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\mu, \mu') \leq d_{\mathrm{Hell}}(\mu, \mu') \leq d_{\mathrm{TV}}(\mu, \mu')^{\frac{1}{2}}.$$

Furthermore, the Hellinger distance is particularly useful for estimating the difference between expectation values of functions of random variables under different measures:

---

[1] M_2 is required to be strictly positive to prevent the case that $\Phi(u; y) = \mathrm{const}$ for all $y \in B_Y(0, r)$ which is deduced from the 2nd inequality.

Lemma 4.1.2. *Let μ and μ' be 2 probability measures on a separable Banach space X. Assume also that $f : X \to E$, where $(E, \|\cdot\|)$ is a separable Banach space, is measurable and has 2nd moments w.r.t. both μ and μ'. Then*

$$\|\mathbb{E}^{\mu} f - \mathbb{E}^{\mu'} f\| \le 2 \left(\mathbb{E}^{\mu} \|f\|^2 + \mathbb{E}^{\mu'} \|f\|^2 \right)^{\frac{1}{2}} d_{\text{Hell}}(\mu, \mu').$$

Furthermore, if E is a separable Hilbert space and $f : X \to E$ as before has 4th moments, then

$$\|\mathbb{E}^{\mu}(f \otimes f) - \mathbb{E}^{\mu'}(f \otimes f)\| \le 2 \left(\mathbb{E}^{\mu} \|f\|^4 + \mathbb{E}^{\mu'} \|f\|^4 \right)^{\frac{1}{2}} d_{\text{Hell}}(\mu, \mu').$$

Proof. See Dashti and Stuart, 2017, p. 410. $\square$

Remark 4.1.1. *Note, in particular, that choosing $X = E$, and with f chosen to be the identity mapping, we deduce that the differences between the mean (resp., covariance operator) of 2 measures are bounded above by their Hellinger distance, provided that one has some a priori control on the 2nd (resp., 4th) moments.*

Go back to our main discussion:

Theorem 4.1.1. *Let Assumptions 4.1.1 hold. Assume that $\mu_0(X') = 1$ and that $\mu_0(X' \cap B) > 0$ for some bounded set B in X. Assume additionally that, for every fixed $r > 0$,*

$$\exp(M_1(r, \|u\|_X)) \in L^1_{\mu_0}(X; \mathbb{R}).$$

Then, for every $y \in Y$, $Z(y)$ given by

$$Z(y) := \int_X \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u).$$

is positive and finite and the probability measure μ^y given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z(y)} \exp(-\Phi(u; y)),$$

is well defined.

Question 4.1.3. *Why need the assumption $\mu_0(X' \cap B) > 0$ for some bounded set B in X?*

Proof. See Dashti and Stuart, 2017, p. 351.

The boundedness of $Z(y)$ follows directly from the lower bound on Φ in Assumption 1, together with the assumed integrability condition in the theorem:

$$Z(y) = \int_X \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) \le \int_X \exp(M_1(r, \|u\|_X)) \mu_0(\mathrm{d}u) < \infty.$$

Since $u \sim \mu_0$ satisfies $u \in X'$ a.s., we have

$$Z(y) = \int_{X'} \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u).$$

Note that $B' := X' \cap B$ is bounded in X. Define

$$R_1 := \sup_{u \in B'} \|u\|_X < \infty.$$

Since $\Phi : X' \times Y \to \mathbb{R}$ is continuous, Φ is finite at every point in $B' \times \{y\}$.

Thus, by the continuity of $\Phi(\cdot; \cdot)$ implied by Assumptions 4.1.1,

$$\sup_{(u,y) \in B' \times B_Y(0,r)} \Phi(u; y) = R_2 < \infty.$$

Indeed,

$$\forall u \in B' \subset X, \ \ \forall y \in B_Y(0,r): \ \ \Phi(u;y) \leq \Phi(u;0) + M_2(r, \|u\|_X)\|y\|_Y \leq \Phi(u;0) + rM_2(r, \|u\|_X),$$

and thus

$$\sup_{(u,y) \in B' \times B_Y(0,r)} \Phi(u;y) \leq \sup_{u \in B'} \Phi(u;0) + rM_2(r, \|u\|_X) \leq \sup_{u \in B'} \Phi(u;0) + rM_2(r, \sup_{u \in B'} \|u\|_X)$$

$$= \sup_{u \in B'} \Phi(u;0) + rM_2(r, R_1) =: R_2.$$

Hence

$$Z(y) = \int_{X'} \exp(-\Phi(u;y))\mu_0(\mathrm{d}u) \geq \int_{B'} \exp(-\Phi(u;y))\mu_0(\mathrm{d}u)$$

$$\geq \int_{B'} \exp(-R_2)\mu_0(\mathrm{d}u) = \exp(-R_2)\mu_0(B') > 0. \tag{4.1.1}$$

Since $\mu_0(B')$ is assumed positive and R_2 is finite, we deduce that $Z(y) > 0$. $\qquad\square$

Remark 4.1.2. *The following remarks apply to the preceding and following theorem.*

- *In the preceding theorem, we are not explicitly working in a Bayesian setting: we are showing that, under the stated conditions on Φ, the measure is well defined and normalizable.*

  In Theorem 14, we did not need to check normalizability because μ^y was defined as a regular conditional probability, via Theorem 13, and therefore automatically normalizable.

- *The lower bound* (4.1.1) *is used repeatedly in what follows, without comment.*

- *Establishing the integrability conditions for both the preceding*

$$\exp(M_1(r, \|u\|_X)) \in L^1_{\mu_0}(X; \mathbb{R}),$$

  and following theorem is often achieved for Gaussian μ_0 by appealing to the Fernique theorem.

Theorem 4.1.2 (Fernique theorem)**.** *Let μ_0 be a Gaussian measure on the separable Banach space X. Then there exists $\beta_c \in (0, \infty)$ s.t.,*[2]

$$\mathbb{E}^{\mu_0} \exp(\beta \|u\|_X^2) < \infty, \ \ \forall \beta \in (0, \beta_c).$$

Theorem 4.1.3. *Let Assumption 1 hold. Assume that $\mu_0(X') = 1$ and that $\mu_0(X' \cap B) > 0$ for some bounded set B in X. Assume additionally that, for every fixed $r > 0$,*

$$\exp(M_1(r, \|u\|_X)) \left(1 + M_2(r, \|u\|_X)^2\right) \in L^1_{\mu_0}(X; \mathbb{R}).$$

Then there is $C = C(r) > 0$ s.t., for all $y, y' \in B_Y(0,r)$

$$d_{\mathrm{Hell}}(\mu^y, \mu^{y'}) \leq C\|y - y'\|_Y.$$

Cf. integrability conditions of these 2 theorems:

$$\exp(M_1(r, \|u\|_X)) \in L^1_{\mu_0}(X; \mathbb{R}),$$

$$\exp(M_1(r, \|u\|_X)) \left(1 + M_2(r, \|u\|_X)^2\right) \in L^1_{\mu_0}(X; \mathbb{R}).$$

---

[2]The subscript c stands for "critical value" which is commonly used in the analysis of PDEs.

Proof. See Dashti and Stuart, 2017, pp. 352–354.

Throughout this proof, we use C to denote a constant independent of u, but possibly depending on the fixed value of r; it may change from occurrence to occurrence.

Since $M_2(r, \cdot)$ is monotonic non-decreasing and strictly positive on $[0, \infty)$,

$$\exp(M_1(r, \|u\|_X))M_2(r, \|u\|_X) \leq \exp(M_1(r, \|u\|_X)) \left(1 + M_2(r, \|u\|_X)^2\right), \tag{4.1.2}$$

$$\exp(M_1(r, \|u\|_X)) \leq \exp(M_1(r, \|u\|_X)) \left(1 + M_2(r, \|u\|_X)^2\right). \tag{4.1.3}$$

Let $Z = Z(y)$ and $Z' = Z(y')$ denote the *normalization constants* for μ^y and $\mu^{y'}$ so that, by Theorem 15,

$$Z = \int_{X'} \exp(-\Phi(u; y))\mu_0(\mathrm{d}u) > 0,$$

$$Z' = \int_{X'} \exp(-\Phi(u; y'))\mu_0(\mathrm{d}u) > 0.$$

Then, using the local Lipschitz property of the exponential and the assumed Lipschitz continuity of $\Phi(u, \cdot)$, together with (4.1.2), we have

$$|Z - Z'| \leq \int_{X'} |\exp(-\Phi(u; y)) - \exp(-\Phi(u; y'))|\mu_0(\mathrm{d}u) \leq \int_{X'} \exp(M_1(r, \|u\|_X))|\Phi(u; y) - \Phi(u; y')|\mu_0(\mathrm{d}u)$$

$$\leq \left(\int_{X'} \exp(M_1(r, \|u\|_X))M_2(r, \|u\|_X)\mu_0(\mathrm{d}u)\right) \|y - y'\|_Y$$

$$\leq \left(\int_{X'} \exp(M_1(r, \|u\|_X)) \left(1 + M_2(r, \|u\|_X)^2\right) \mu_0(\mathrm{d}u)\right) \|y - y'\|_Y \leq C\|y - y'\|_Y.$$

The last line follows because the integrand is in $L^1_{\mu_0}$ by assumption.

From the definition of Hellinger distance, we have

$$\left(d_{\mathrm{Hell}}(\mu^y, \mu^{y'})\right)^2 \leq I_1 + I_2,$$

where

$$I_1 := \frac{1}{Z} \int_{X'} \left(\exp\left(-\frac{1}{2}\Phi(u; y)\right) - \exp\left(-\frac{1}{2}\Phi(u; y')\right)\right)^2 \mu_0(\mathrm{d}u),$$

$$I_2 := |Z^{-\frac{1}{2}} - (Z')^{-\frac{1}{2}}|^2 \int_{X'} \exp(-\Phi(u; y'))\mu_0(\mathrm{d}u).$$

Indeed, recall that

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) := \frac{1}{Z}\exp(-\Phi(u; y)) \text{ where } Z := Z(y) := \int_X \exp(-\Phi(u; y))\mu_0(\mathrm{d}u) = \int_{X'} \exp(-\Phi(u; y))\mu_0(\mathrm{d}u),$$

$$\frac{\mathrm{d}\mu^{y'}}{\mathrm{d}\mu_0}(u) := \frac{1}{Z'}\exp(-\Phi(u; y')) \text{ where } Z' := Z(y') := \int_X \exp(-\Phi(u; y'))\mu_0(\mathrm{d}u) = \int_{X'} \exp(-\Phi(u; y'))\mu_0(\mathrm{d}u),$$

then

$$\left(d_{\mathrm{Hell}}(\mu^y, \mu^{y'})\right)^2 = \frac{1}{2}\int_{X'} \left(\sqrt{\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}} - \sqrt{\frac{\mathrm{d}\mu^{y'}}{\mathrm{d}\mu_0}}\right)^2 \mathrm{d}\mu_0(\mathrm{d}u)$$

$$= \frac{1}{2}\int_{X'} \left[Z^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\Phi(u; y)\right) - (Z')^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\Phi(u; y')\right)\right]^2 \mu_0(\mathrm{d}u)$$

$$= \frac{1}{2}\int_{X'} \left[Z^{-\frac{1}{2}}\left(\exp\left(-\frac{1}{2}\Phi(u; y)\right) - \exp\left(-\frac{1}{2}\Phi(u; y')\right)\right) + \left(Z^{-\frac{1}{2}} - (Z')^{-\frac{1}{2}}\right)\exp\left(-\frac{1}{2}\Phi(u; y')\right)\right]^2 \mu_0\mathrm{d}u$$

$$\leq \frac{1}{Z} \int_{X'} \left(\exp\left(-\frac{1}{2}\Phi(u;y)\right) - \exp\left(-\frac{1}{2}\Phi(u;y')\right) \right)^2 \mu_0(\mathrm{d}u) + |Z^{-\frac{1}{2}} - (Z')^{-\frac{1}{2}}|^2 \int_{X'} \exp(-\Phi(u;y'))\mu_0(\mathrm{d}u).$$

Note that, again using similar Lipschitz calculations to those above, using the fact that $Z > 0$ and Assumption 4.1.1,

$$I_1 \leq \cdots \leq C\|y - y'\|_Y^2.$$

Also, using Assumptions 4.1.1, together with 4.1.3,

$$I_2 \leq \cdots \leq C\|y - y'\|_Y^2.$$

The result is complete.   □

Remark 4.1.3. *The Hellinger metric has the very desirable property that it translates directly into bounds on expectations.*

For functions f which are in $L^2_{\mu^y}(X;\mathbb{R})$ and $L^2_{\mu^{y'}}(X;\mathbb{R})$, the closeness of the Hellinger metric implies closeness of expectations of f.

To be precise, for $y, y' \in B_Y(0, r)$, we have

$$|\mathbb{E}^{\mu^y} f(u) - \mathbb{E}^{\mu^{y'}} f(u)| \leq C d_{\mathrm{Hell}}(\mu^y, \mu^{y'}),$$

where constant C depends on r and on the expectations of $|f|^2$ under μ^y and $\mu^{y'}$.

It follows that

$$|\mathbb{E}^{\mu^y} f(u) - \mathbb{E}^{\mu^{y'}} f(u)| \leq C\|y - y'\|_Y,$$

for a possibly different constant C which also depends on r and on the expectations of $|f|^2$ under μ^y and $\mu^{y'}$.

4.2   Approximation

In this section we concentrate on continuity properties of the posterior measure w.r.t. approximation of the potential Φ.

The methods used are very similar to those in the previous subsection, and we establish a continuity property of the posterior distribution, in the Hellinger metric, w.r.t. small changes in the potential Φ.

Because the data y plays no explicit role in this discussion, we drop explicit reference to it.

Let X be a Banach space and μ_0 a measure on X.

Assume that μ and μ^N are both absolutely continuous w.r.t. μ_0 and given by **(10.43)**

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = \frac{1}{Z} \exp(-\Phi(u)), \ Z = \int_X \exp(-\Phi(u))\mu_0(\mathrm{d}u),$$

and **(10.44)**

$$\frac{\mathrm{d}\mu^N}{\mathrm{d}\mu_0}(u) = \frac{1}{Z^N} \exp(-\Phi^N(u)), \ Z^N = \int_X \exp(-\Phi^N(u))\mu_0(\mathrm{d}u),$$

resp.

The measure μ^N might arise, e.g., through an approximation of the forward map G underlying an inverse problem of the form (10.29).

It is natural to ask whether closeness of the forward map and its approximation imply closeness of the posterior measures.

Dashti & Stuart now address this question.

Assumption 4.2.1. *Let $X' \subseteq X$ and assume that $\Phi \in C(X'; \mathbb{R})$.*

Assume further that there are functions $M_i : \mathbb{R}^+ \to \mathbb{R}^+$, $i = 1, 2$, independent of N and monotonic non-decreasing separately in each argument, and with M_2 strictly positive, s.t. for all $u \in X'$,

$$\Phi(u) \geq -M_1(\|u\|_X),$$
$$\Phi^N(u) \geq -M_1(\|u\|_X),$$
$$|\Phi(u) - \Phi^N(u)| \leq M_2(\|u\|_X)\psi(N),$$

where $\psi(N) \to 0$ as $N \to \infty$.

The following 2 theorems are very similar to Theorems 15–16, and the proofs are adapted to estimate changes in the posterior caused by changes in the potential Φ, rather than the data y.

Theorem 4.2.1. *Let Assumptions 2 hold. Assume that $\mu_0(X') = 1$ and that $\mu_0(X' \cap B) > 0$ for some bounded set B in X. Assume additionally that, for every fixed $r > 0$,*

$$\exp(M_1(r, \|u\|_X)) \in L^1_{\mu_0}(X; \mathbb{R}).$$

Then Z and Z^N given by (10.43b) and (10.44b) are positive and finite and the probability measures μ and μ^N given by (10.43) and (10.44) are well defined. Furthermore, for sufficiently large N, Z^N given by (10.44b) is bounded below by a positive constant independent of N.

Proof. See Dashti and Stuart, 2017, pp. 355–356. □

Theorem 4.2.2. *Let Assumptions 2 hold. Assume that $\mu_0(X') = 1$ and that $\mu_0(X' \cap B) > 0$ for some bounded set B in X. Assume additionally that*

$$\exp(M_1(\|u\|_X)) \left(1 + M_2(\|u\|_X)^2\right) \in L^1_{\mu_0}(X; \mathbb{R}).$$

Then there is $C > 0$ s.t., for all N sufficiently large,

$$d_{\mathrm{Hell}}(\mu, \mu^N) \leq C\psi(N).$$

Proof. See Dashti and Stuart, 2017, pp. 356–358. □

Remark 4.2.1. *The following 2 remarks are relevant to establishing the conditions of the preceding 2 theorems and to applying them.*

- *As mentioned in the previous subsection concerning well posedness, the Fernique theorem can frequently be used to establish integrability conditions, e.g. those in the 2 preceding theorems when μ_0 is Gaussian.*

- *Using the ideas underlying Remark 1, the preceding theorem enables us to translate errors arising from approximation of the forward problem into errors in the Bayesian solution of the inverse problem.*

  Furthermore, the errors in the forward and inverse problems scale the same way w.r.t. N.

  For functions f which are in L^2_μ and $L^2_{\mu^N}$, uniformly w.r.t. N, the closeness of the Hellinger metric implies closeness of expectations of f:

$$|\mathbb{E}^\mu f(u) - \mathbb{E}^{\mu^N} f(u)| \leq C\psi(N).$$

4.3   MAP Estimators & Tikhonov Regularization

The aim of this section is to connect the probabilistic approach to inverse problems with the classical method of Tikhonov regularization.

Dashti & Stuart consider the setting in which the prior measure μ_0 is a Gaussian measure.

Dashti & Stuart then show that MAP estimators, points of maximal probability, coincide with minimizers of a Tikhonov-Phillips regularized least squares function, with regularization being w.r.t. the Cameron-Martin norm of the Gaussian prior.

The data y plays no explicit role in the authors' developments here, and so they work in the setting of equation (10.43).

Recall, however, that in the context of inverse problems, a classical methodology is to simply try and minimize (subject to some regularization) $\Phi(u)$.

Indeed for finite data and Gaussian observational noise with Gaussian distribution $N(0, \Gamma)$, we have

$$\Phi(u) = \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y - G(u))|^2.$$

Thus Φ is simply a covariance weighted model-data misfit least squares function.

In this section we show that maximizing probability under μ (in a sense that we will make precise in what follows) is equivalent to minimizing **(10.46)**

$$I(u) = \begin{cases} \Phi(u) + \frac{1}{2}\|u\|_E^2 & \text{if } u \in E, \\ +\infty & \text{else.} \end{cases}$$

Here $(E, \|\cdot\|_E)$ denotes the Cameron-Martin space associated to the Gaussian prior μ_0.

We view μ_0 as a Gaussian probability measure on a separable Banach space $(X, \|\cdot\|_X)$ so that $\mu_0(X) = 1$.

We make the following assumptions about the function Φ:

Assumption 4.3.1. *The function $\Phi : X \to \mathbb{R}$ satisfies the following conditions:*

(i) For every $\epsilon > 0$, there is an $M = M(\epsilon) \in \mathbb{R}$, s.t. for all $u \in X$,

$$\Phi(u) \geq M - \epsilon\|u\|_X^2.$$

(ii) Φ is locally bounded from above, i.e., for every $r > 0$ there exists $K = K(r) > 0$ s.t., for all $u \in X$ with $\|u\|_X < r$, we have

$$\Phi(u) \leq K.$$

(iii) Φ is locally Lipschitz continuous, i.e., for every $r > 0$ there exists $L = L(r) > 0$ s.t. for all $u_1, u_2 \in X$ with $\|u_1\|_X, \|u_2\|_X < r$, we have

$$|\Phi(u_1) - \Phi(u_2)| \leq L\|u_1 - u_2\|_X.$$

In finite dimensions, for measures which have a continuous density w.r.t. Lebesgue measure, there is an obvious notion of most likely point(s): simply the point(s) at which the Lebesgue density is maximized.

This way of thinking does not translate into the infinite-dimensional context, but there is a way of restating it which does.

Fix a small radius $\delta > 0$ and identify centers of balls of radius δ which have maximal probability.

Letting $\delta \to 0$ then recovers the preceding definition, when there is a continuous Lebesgue density.

Dashti & Stuart adopt this small ball approach in the infinite-dimensional setting.

For $z \in E$, let $B^\delta(z) \subset X$ be the open ball centered at $z \in X$ with radius δ in X.

Let

$$J^\delta(z) = \mu(B^\delta(z))$$

be the mass of the ball $B^\delta(z)$ under the measure μ.

Similarly we define

$$J_0^\delta(z) = \mu_0(B^\delta(z))$$

the mass of the ball $B^\delta(z)$ under the Gaussian prior.

Recall that at balls in a separable Banach space have positive Gaussian measure, by Theorem 33; it thus follows that $J_0^\delta(z)$ is finite and positive for any $z \in E$.

By Assumptions 3(i) and (ii) together with the Fernique Theorem 10, the same is true for $J^\delta(z)$.

Dashti & Stuart' 1st theorem encapsulates the idea that probability is maximized where I is minimized.

To see this, fix any point z_2 in the Cameron-Martin space E and notice that the probability of the small ball at z_1 is maximized, asymptotically as the radius of the ball tends to zero, at minimizers of I.

Theorem 4.3.1. *Let Assumptions 3 hold and assume that $\mu_0(X) = 1$. Then the function I defined by (10.46) satisfies, for any $z_1, z_2 \in E$,*

$$\lim_{\delta \to 0} \frac{J^\delta(z_1)}{J^\delta(z_2)} = \exp(I(z_2) - I(z_1)).$$

Proof. See Dashti and Stuart, 2017, pp. 360–361. $\square$

Dashti & Stuart have thus linked the Bayesian approach to inverse problems with a classical regularization technique.

Dashti & Stuart conclude the subsection by showing that, under the prevailing Assumption 3, the minimization problem for I is well defined.

1st recall a basic definition and lemma from the calculus of variations.

Definition 4.3.1. *The function $I : E \to \mathbb{R}$ is* weakly lower semicontinuous *if*

$$\liminf_{n \to \infty} I(u_n) \geq I(u)$$

whenever $u_n \rightharpoonup u$ in E.

The function $I : E \to \mathbb{R}$ is weakly continuous *if*

$$\lim_{n \to \infty} I(u_n) = I(u)$$

whenever $u_n \rightharpoonup u$ in E.

Clearly weak continuity implies weak lower semicontinuity.

Lemma 4.3.1. *If $(E, \langle \cdot, \cdot \rangle_E)$ is a Hilbert space with induced norm $\| \cdot \|_E$, then the quadratic form $J(u) \coloneqq \frac{1}{2}\|u\|_E^2$ is weakly lower semicontinuous.*

Proof. See Dashti and Stuart, 2017, p. 362. $\square$

Theorem 4.3.2. *Suppose that Assumption 3 hold and let E be a Hilbert space compactly embedded in X. Then there exists $\overline{u} \in E$ s.t.*

$$I(\overline{u}) = \overline{I} \coloneqq \inf\{I(u); u \in E\}.$$

Furthermore, if $\{u_n\}$ is a minimizing sequence satisfying $I(u_n) \to I(\overline{u})$, then there is a subsequence $\{u_{n'}\}$ that converges strongly to $\overline{u}$ in E.

Proof. See Dashti and Stuart, 2017, pp. 362–363. $\square$

Corollary 4.3.1. *Suppose that Assumptions 3 hold and the Gaussian measure μ_0 with Cameron-Martin space E satisfies $\mu_0(X) = 1$. Then there exists $\overline{u} \in E$ s.t.*

$$I(\overline{u}) = \overline{I} \coloneqq \inf\{I(u); u \in E\}.$$

Furthermore, if $\{u_n\}$ is a minimizing sequence satisfying $I(u_n) \to I(\overline{u})$, then there is a subsequence $\{u_{n'}\}$ that converges strongly to $\overline{u}$ in E.

Proof. See Dashti and Stuart, 2017, p. 363. $\square$

Appendix A

Tools

A.1 Probability Theory

A.1.1 Background in Probability Theory

Probability Space $(\Omega, \mathcal{A}, \mu)$

- Ω is a set

- $\mathcal{A}$ is a σ-algebra over Ω

  - $\emptyset \in \mathcal{A}$,
  - the complement $A^c := \{\omega \in \Omega; \omega \notin A\} \in \mathcal{A}$ for all $A \in \mathcal{A}$ and
  - the union $\bigcup_{j \in \mathbb{N}} A_j \in \mathcal{A}$ for every sequence $(A_j)_{j \in \mathbb{N}} \in \mathcal{A}$.

- A *measure* μ on a measurable space $(\Omega, \mathcal{A})$ is a mapping from $\mathcal{A}$ to $\mathbb{R}_+ \cup \{\infty\}$ s.t.

  - the empty set has measure zero, i.e., $\mu(\emptyset) = 0$ and
  - $\mu\left(\bigcup_{j \in \mathbb{N}} A_j\right) = \sum_{j \in \mathbb{N}} \mu(A_j)$ if $A_j \in \mathcal{A}$ are disjoint.

- μ is a *probability measure* iff

  - $\mu(\Omega) = 1$

- $\mathcal{B}(\mathbb{R})$ denotes the *Borel σ-algebra* and equals the smallest σ-algebra containing all open subsets of $\mathbb{R}$.

- The usual notion of volume of subsets of $\mathbb{R}^d$ gives rise to the *Lebesgue measure*.

A.1.2 Real-valued Random Variables

Let $(\Omega, \mathcal{A}, \mu)$ be a probability space.
A mapping $f : \Omega \to \mathbb{R}$ is called a *random variable*, if f is $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable, i.e.,

$$\forall B \in \mathcal{B}(\mathbb{R}), \ f^{-1}(B) = \{\omega \in \Omega; f(\omega) \in B\} \in \mathcal{A} \Rightarrow f^{-1}(\mathcal{B}(\mathbb{R})) \subset \mathcal{A}.$$

- Image measure/distribution

  - $\mu_f(B) := (\mu \circ f^{-1}(B)) = \mu(f^{-1}(B)) = \mu(\{f \in B\})$.
  - $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_f)$ is again a probability measure.

- Distribution function

  - $F_f(x) := \mu_f((-\infty, x]) = \mu(f \le x) = \mu(\{\omega \in \Omega; f(\omega) \le x\})$.

44

- Density function
  - If $\mu_f(B) := \int_B \rho(x)\mathrm{d}x$ for all $B \in \mathcal{B}(\mathbb{R})$ with a measurable function $\rho : \mathbb{R} \to [0, \infty)$.

- If h is a real-valued measurable function on $\mathbb{R} \Rightarrow h \circ f$ $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable

A.1.3 Expectations

- For $f \in \mathcal{L}^1(\mu)$ is the *expectation* defined as

$$\mathbb{E}(f) = \int_\Omega f \mathrm{d}\mu = \int_\Omega f(\omega)\mathrm{d}\mu(\omega).$$

- For a real-valued measurable function h
  - discretely distributed random variable: $\mu_f = \sum_{i=1}^n p_i \mathbf{1}_{x_i}$

$$\mathbb{E}(h \circ f) = \int_\Omega h \circ f \mathrm{d}\mu = \int_\mathbb{R} h \mathrm{d}\mu_f = \sum_{i=1}^n p_i f_{x_i}.???$$

  - continuously distributed random variable: $\mu_f([a, b]) = \int_a^b \rho(x)\mathrm{d}x$.

$$\mathbb{E}(h \circ f) = \int_\Omega h \circ f \mathrm{d}\mu = \int_\mathbb{R} h \mathrm{d}\mu_f = \int_\mathbb{R} h(x)\rho(x)\mathrm{d}x.$$

- For $f \in \mathcal{L}^2(\mu)$ is the *variance* defined as

$$\mathbb{V}(f) = \mathbb{E}((f - \mathbb{E}(f))^2) = \mathbb{E}(f^2) - (\mathbb{E}(f))^2.$$

A.1.4 Product Spaces & Joint Distributions

- Product space $(\bigtimes_{i=1}^N \Omega_i, \bigotimes_{i=1}^N \mathcal{A}_i, \bigotimes_{i=1}^N \mu_i)$
  - $\bigtimes_{i=1}^N \Omega_i = \{(\omega_1, \ldots, \omega_N); \omega_i \in \Omega_i, \ i = 1, \ldots, N\}$.
  - $\bigotimes_{i=1}^N \mathcal{A}_i$ smallest σ-algebra that contains all sets of the form $A = A_1 \times \cdots \times A_N$ with $A_i \in \mathcal{A}_i$, $i = 1, \ldots, N$.
  - $\bigotimes_{i=1}^N \mu_i$ is defined by

$$\bigotimes_{i=1}^N \mu_i = \mu_1(A_1) \times \cdots \times \mu_N(A_N), \ \forall A_1 \in \mathcal{A}_1, \ldots, A_N \in \mathcal{A}_N.$$

- For a real-valued random variables $(f_i, \ i = 1, \ldots, N)$, we define $f_1 \otimes \cdots \otimes f_N : \Omega \to \mathbb{R}^N$ via $\omega \mapsto (f_1(\omega), \ldots, f_N(\omega))$ a random variable in $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$.
  - Joint distribution is for $B \in \mathcal{B}(\mathbb{R}^N)$ given by

$$\mu_{f_1 \otimes \cdots \otimes f_N}(B) = \mu(f_1 \otimes \cdots \otimes f_N \in B).$$

  - Joint distribution function is given by

$$F_{f_1 \otimes \cdots \otimes f_N}(x_1, \ldots, x_N) = \mu_{f_1 \otimes \cdots \otimes f_N}((-\infty, x_1], \ldots, (-\infty, x_N]) = \mu(f_1 \leq x_1, \ldots, f_n \leq x_n).$$

A.1.5   Conditional Probabilities & Expectations

Let $(\Omega, \mathcal{A}, \mu)$ be a probability space.

- The *conditional probability* that an event $A \in \mathcal{A}$ occurs given that an event $B \in \mathcal{A}$ has occurred is defined by

$$\mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)} \text{ for } \mu(B) > 0.$$

- Given 2 random variables f and h with joint density function ρ and $x \in \mathbb{R}$ s.t. $\int_{-\infty}^{\infty} \rho(x, v) \mathrm{d}v > 0$,

  - the *conditional distribution function* of h given $f = x$ is given by

  $$F_{h|f}(y|x) = \int_{-\infty}^{y} \frac{\rho(x, v) \mathrm{d}v}{\int_{-\infty}^{\infty} \rho(x, v) \mathrm{d}v}.$$

  - the *conditional density function* of h given $f = x$ is given by

  $$\rho_{h|f}(y|x) = \frac{\rho(x, y)}{\int_{-\infty}^{\infty} \rho(x, v) \mathrm{d}v}.$$

  - the *conditional expectation* $\mathbb{E}(h|f)$ is defined by

  $$\mathbb{E}(h|f) = \int_{-\infty}^{\infty} y \rho_{h|f}(y|x) \mathrm{d}y.$$

A.1.6   Independent Random Variables

- For *independent random variables* $(f_i)_{i=1}^{N}$, it holds

$$\mu_{f_1 \otimes \cdots \otimes f_N}(B_1 \times \cdots \times B_N) = \mu(f_1 \in B_1, \ldots, f_N \in B_N) = \mu(f_1 \in B_1) \cdots \mu(f_N \in B_N)$$
$$= \mu_{f_1}(B_1) \cdots \mu_{f_N}(B_N) = \mu_{f_1} \otimes \cdots \otimes \mu_{f_N}(B_1 \times \cdots \times B_N).$$

- $(f_i)_{i=1}^{N}$ are independent $\Leftrightarrow$ the joint distribution is equal to the product measure.

A.1.7   Metrics on Probability Measures

When discussing well posedness and approximation theory for the posterior distribution, it is of interest to estimate the distance between 2 probability measures, and thus we will be interested in metrics between probability measures.

In this subsection the authors introduce 2 useful metrics on measures: the *total variation distance* and the *Hellinger distance*.

Dashti & Stuart discuss the relationships between the metrics and indicate how they may be used to estimate differences between expectations of random variables under 2 different measures.

Dashti & Stuart also discuss the *Kullback-Leibler divergence*, a useful distance measure which does not satisfy the axioms of a metric, but which may be used to bound both the Hellinger and total variation distances, and which is also useful in defining algorithms for finding the best approximation to a given measure from within some restricted class of measures, e.g. Gaussians.

Assume that we have 2 probability measures μ and μ' on a separable Banach space denoted by B (actually the considerations here apply on a Polish space but we do not need this level of generality).

Assume that μ and μ' are both absolutely continuous w.r.t. a common reference measure ν, also defined on the same measure space.

Such a measure always exists - take $\nu = \frac{1}{2}(\mu + \mu')$, e.g.

In the following, all integrals of real-valued functions over B are simply denoted by $\int$.

The following define 2 concepts of distance between μ and μ'.

The resulting metrics that we define are independent of the choice of this common reference measure.

Definition A.1.1. *The* total variation distance *between μ and μ' is*

$$d_{\mathrm{TV}}(\mu, \mu') := \frac{1}{2} \int \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu} \right| \mathrm{d}\nu.$$

In particular, if μ' is absolutely continuous w.r.t. μ, then **(10.118)**

$$d_{\mathrm{TV}}(\mu, \mu') := \frac{1}{2} \int \left| 1 - \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \right| \mathrm{d}\mu.$$

Definition A.1.2. *The* Hellinger distance *between μ and μ' is*

$$d_{\mathrm{Hell}}(\mu, \mu') := \sqrt{\frac{1}{2} \int \left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}} \right)^2 \mathrm{d}\nu}.$$

In particular, if μ' is absolutely continuous w.r.t. μ, then **(10.119)**

$$d_{\mathrm{Hell}}(\mu, \mu') = \sqrt{\frac{1}{2} \int \left(1 - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}} \right)^2 \mathrm{d}\mu}.$$

Note that the numerical constant $\frac{1}{2}$ appearing in both definitions is chosen in such a way as to ensure the bounds

$$0 \leq d_{\mathrm{TV}}(\mu, \mu') \leq 1, \ 0 \leq d_{\mathrm{Hell}}(\mu, \mu') \leq 1.$$

In the case of the total variation inequality, this is an immediate consequence of the triangle inequality, combined with the fact that both μ and μ' are probability measures, so that $\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathrm{d}\nu = 1$ and similarly for μ'.

In the case of the Hellinger distance, it follows by expanding the square and applying similar considerations.

The Hellinger and total variation distances are related as follows, which shows in particular that they both generate the same topology:

Lemma A.1.1. *The total variation and Hellinger metrics are related by the inequalities*

$$\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\mu, \mu') \leq d_{\mathrm{Hell}}(\mu, \mu') \leq d_{\mathrm{TV}}(\mu, \mu')^{\frac{1}{2}}.$$

Proof. See Dashti and Stuart, 2017, p. 408. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Example A.1.1. *Consider 2 Gaussian densities on $\mathbb{R}$: $N(m_1, \sigma_1^2)$ and $N(m_2, \sigma_2^2)$.*
The Hellinger distance between them is given by

$$d_{\mathrm{Hell}}(\mu, \mu')^2 = 1 - \sqrt{\exp\left(-\frac{(m_1 - m_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right) \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}.$$

To see this note that

$$d_{\mathrm{Hell}}(\mu, \mu')^2 = 1 - \frac{1}{(2\pi\sigma_1\sigma_2)^{\frac{1}{2}}} \int_{\mathbb{R}} \exp(-Q)\mathrm{d}x,$$

where

$$Q = \frac{1}{4\sigma_1^2}(x - m_1)^2 + \frac{1}{4\sigma_2^2}(x - m_2)^2.$$

Define σ^2 by

$$\frac{1}{\sigma^2} = \frac{1}{4\sigma_1^2} + \frac{1}{4\sigma_2^2}.$$

We change variable under the integral to y given by

$$y = x - \frac{m_1 + m_2}{2},$$

and note that then, by completing the square,

$$Q = \frac{1}{2\sigma^2}(y - m)^2 + \frac{1}{4(\sigma_1^2 + \sigma_2^2)}(m_2 - m_1)^2,$$

where m does not appear in what follows and so we do not detail it.

Note that the integral is then a multiple of a standard Gaussian $N(m, \sigma^2)$ gives the desired result.

In particular this calculation shows that the Hellinger distance between 2 Gaussians on $\mathbb{R}$ tends to 0 iff the means and variances of the 2 Gaussians approach one another.

Furthermore, by the previous lemma, the same is true for the total variation distance.

The preceding example generalizes to higher dimension and shows that, e.g., the total variation and Hellinger metrics cannot metrize weak convergence of probability measures (as one can also show that convergence in total variation metric implies strong convergence).

They are nonetheless useful distance measures, e.g., between families of measures which are mutually absolutely continuous.

Furthermore, the Hellinger distance is particularly useful for estimating the difference between expectation values of functions of random variables under different measures.

This is encapsulated in the following lemma:

Lemma A.1.2. *Let μ and μ' be 2 probability measures on a separable Banach space X. Assume also that $f : X \to E$, where $(E, \|\cdot\|)$ is a separable Banach space, is measurable and has 2nd moments w.r.t. both μ and μ'. Then*

$$\|\mathbb{E}^\mu f - \mathbb{E}^{\mu'} f\| \leq 2 \left(\mathbb{E}^\mu \|f\|^2 + \mathbb{E}^{\mu'} \|f\|^2 \right)^{\frac{1}{2}} d_{\mathrm{Hell}}(\mu, \mu').$$

Furthermore, if E is a separable Hilbert space and $f : X \to E$ as before has 4th moments, then

$$\|\mathbb{E}^\mu (f \otimes f) - \mathbb{E}^{\mu'}(f \otimes f)\| \leq 2 \left(\mathbb{E}^\mu \|f\|^4 + \mathbb{E}^{\mu'} \|f\|^4 \right)^{\frac{1}{2}} d_{\mathrm{Hell}}(\mu, \mu').$$

Proof. See Dashti and Stuart, 2017, p. 410. $\qquad\qquad\square$

Remark A.1.1. *Note, in particular, that choosing $X = E$, and with f chosen to be the identity mapping, we deduce that the differences between the mean (resp., covariance operator) of 2 measures are bounded above by their Hellinger distance, provided that one has some a priori control on the 2nd (resp., 4th) moments.*

We now define a 3rd widely used distance concept for comparing 2 probability measures.

Note, however, that it does not give rise to a metric in the strict sense, because it violates both symmetry and the triangle inequality.

Definition A.1.3. *The* Kullback-Leibler divergence *between 2 measures μ' and μ, with μ' absolutely continuous w.r.t. μ, is*

$$D_{\mathrm{KL}}(\mu' \| \mu) = \int \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \log \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \mathrm{d}\mu.$$

If μ is also absolutely continuous w.r.t. μ', so that the 2 measures are equivalent, then

$$D_{\mathrm{KL}}(\mu' \| \mu) = - \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\mu'} \mathrm{d}\mu',$$

and the 2 definitions coincide.

Example A.1.2. *Consider 2 Gaussian densities on $\mathbb{R}$: $N(m_1, \sigma_1^2)$ and $N(m_2, \sigma_2^2)$.*
  The Kullback-Leibler divergence between them is given by

$$D_{\mathrm{KL}}(\mu_1 \| \mu_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{(m_2 - m_1)^2}{2\sigma_2^2}.$$

To see this note that

$$\begin{aligned}
D_{\mathrm{KL}} &= \mathbb{E}^{\mu_1}\left(\log\sqrt{\frac{\sigma_2^2}{\sigma_1^2}} + \frac{1}{2\sigma_2^2}|x - m_2|^2 - \frac{1}{2\sigma_1^2}|x - m_1|^2\right) \\
&= \log\frac{\sigma_2}{\sigma_1} + \mathbb{E}^{\mu_1}\left(\left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)|x - m_1|^2\right) + \mathbb{E}^{\mu_1}\frac{1}{2\sigma_2^2}\left(|x - m_2|^2 - |x - m_1|^2\right) \\
&= \log\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{1}{2\sigma_2^2}\mathbb{E}^{\mu_1}\left(m_2^2 - m_1^2 + 2x(m_1 - m_2)\right) \\
&= \log\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{1}{2\sigma_2^2}(m_2 - m_1)^2
\end{aligned}$$

as required.

   As for Hellinger distance, this example shows that 2 Gaussians on $\mathbb{R}$ approach one another in the Kullback-Leibler divergence iff their means and variances approach one another.
   This generalizes to higher dimensions.
   The Kullback-Leibler divergence provides an upper bound for the square of the Hellinger distance and for the square of the total variation distance.

Lemma A.1.3. *Assume that 2 measures μ and μ' are equivalent. Then the bounds*

$$d_{\mathrm{Hell}}(\mu, \mu')^2 \leq \frac{1}{2}D_{\mathrm{KL}}(\mu \| \mu'), \ d_{\mathrm{TV}}(\mu, \mu')^2 \leq D_{\mathrm{KL}}(\mu \| \mu'),$$

hold.

Proof. See Dashti and Stuart, 2017, p. 412. $\square$

Bibliography

Aster, Richard C., Brian Borchers, and Clifford H. Thurber (2018). *Parameter estimation and inverse problems*. Third. Elsevier/Academic Press, Amsterdam, pp. xi+392. ISBN: 978-0-12-804651-7. DOI: 10.1016/C2015-0-02458-3. URL: https://doi.org/10.1016/C2015-0-02458-3.

Dashti, Masoumeh and Andrew M. Stuart (2017). "The Bayesian approach to inverse problems". In: *Handbook of uncertainty quantification. Vol. 1, 2, 3*. Springer, Cham, pp. 311–428.

Evans, Lawrence C. (2010). *Partial differential equations*. Second. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, pp. xxii+749. ISBN: 978-0-8218-4974-3. DOI: 10.1090/gsm/019. URL: https://doi.org/10.1090/gsm/019.

Ito, Kazufumi and Bangti Jin (2015). *Inverse problems*. Vol. 22. Series on Applied Mathematics. Tikhonov theory and algorithms. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, pp. x+318. ISBN: 978-981-4596-19-0.

Kirsch, Andreas (2021). *An introduction to the mathematical theory of inverse problems*. Third. Vol. 120. Applied Mathematical Sciences. Springer, Cham, p. 400. ISBN: 978-3-030-63343-1; 978-3-030-63342-4. DOI: 10.1007/978-3-030-63343-1. URL: https://doi.org/10.1007/978-3-030-63343-1.

Pope, Stephen B. (2000). *Turbulent flows*. Cambridge University Press, Cambridge, pp. xxxiv+771. ISBN: 0-521-59886-9. DOI: 10.1017/CBO9780511840531. URL: https://doi.org/10.1017/CBO9780511840531.