Chapter 6
Bayesian Inverse Problems

> It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.
>
> Mark Twain

This chapter provides a general introduction, at the high level, to the *backward* propagation of uncertainty/information in the solution of *inverse problems*, and specifically a Bayesian probabilistic perspective on such inverse problems. Under the umbrella of inverse problems, we consider parameter estimation and regression. One specific aim is to make clear the connection between regularization and the application of a Bayesian prior. The filtering methods of Chapter 7 fall under the general umbrella of Bayesian approaches to inverse problems, but have an additional emphasis on real-time computational expediency.

Many modern UQ applications involve inverse problems where the unknown to be inferred is an element of some infinite-dimensional function space, e.g. inference problems involving PDEs with uncertain coefficients. Naturally, such problems can be discretized, and the inference problem solved on the finite-dimensional space, but this is not always a well-behaved procedure: similar issues arise in Bayesian inversion on function spaces as arise in the numerical analysis of PDEs. For example, there are 'stable' and 'unstable' ways to discretize a PDE (e.g. the Courant–Friedrichs–Lewy condition), and analogously there are 'stable' and 'unstable' ways to discretize a Bayesian inverse problem. Sometimes, a discretized PDE problem has a solution, but the original continuum problem does not (e.g. the backward heat equation, or the control problem for the wave equation), and this phenomenon can be seen in the ill-conditioning of the discretized problem as the discretization dimension tends to infinity; similar problems can afflict a discretized Bayesian inverse

T.J. Sullivan, *Introduction to Uncertainty Quantification*, Texts in Applied Mathematics 63, DOI 10.1007/978-3-319-23395-6_6

problem. Therefore, one aim of this chapter is to present an elementary well-posedness theory for Bayesian inversion on the function space, so that this well-posedness will automatically be inherited by any finite-dimensional discretization. For a thorough treatment of all these questions, see the sources cited in the bibliography.

6.1 Inverse Problems and Regularization

Many mathematical models, and UQ problems, are *forward problems*, i.e. we are given some input u for a mathematical model H, and are required to determine the corresponding output y given by

$$y = H(u), \tag{6.1}$$

where $\mathcal{U}$, $\mathcal{Y}$ are, say, Banach spaces, $u \in \mathcal{U}$, $y \in \mathcal{Y}$, and $H \colon \mathcal{U} \to \mathcal{Y}$ is the *observation operator*. However, many applications require the solution of the *inverse problem*: we are given y and H and must determine u such that (6.1) holds. Inverse problems are typically ill-posed: there may be no solution, the solution may not be unique, or there may be a unique solution that depends sensitively on y. Indeed, very often we do not actually observe $H(u)$, but some noisily corrupted version of it, such as

$$y = H(u) + \eta. \tag{6.2}$$

The inverse problem framework encompasses that problem of *model calibration* (or *parameter estimation*), where a model H_θ relating inputs to outputs depends upon some parameters $\theta \in \Theta$, e.g., when $\mathcal{U} = \mathcal{Y} = \Theta$, $H_\theta(u) = \theta u$. The problem is, given some observations of inputs u_i and corresponding outputs y_i, to find the parameter value θ such that

$$y_i = H_\theta(u_i) \quad \text{for each } i.$$

Again, this problem is typically ill-posed.

One approach to the problem of ill-posedness is to seek a least-squares solution: find, for the norm $\| \cdot \|_{\mathcal{Y}}$ on $\mathcal{Y}$,

$$\operatorname*{arg\,min}_{u \in \mathcal{U}} \big\| y - H(u) \big\|_{\mathcal{Y}}^2 .$$

However, this problem, too, can be difficult to solve as it may possess minimizing sequences that do not have a limit in $\mathcal{U}$,[1] or may possess multiple minima, or may depend sensitively on the observed data y. Especially in this

[1] Take a moment to reconcile the statement "there may exist minimizing sequences that do not have a limit in $\mathcal{U}$" with $\mathcal{U}$ being a Banach space.

last case, it may be advantageous to not try to fit the observed data too closely, and instead *regularize* the problem by seeking

$$\arg\min \left\{ \left\| y - H(u) \right\|_{\mathcal{Y}}^2 + \left\| u - \bar{u} \right\|_{\mathcal{V}}^2 \,\middle|\, u \in \mathcal{V} \subseteq \mathcal{U} \right\}$$

for some Banach space $\mathcal{V}$ embedded in $\mathcal{U}$ and a chosen $\bar{u} \in \mathcal{V}$. The standard example of this regularization setup is *Tikhonov regularization*, as in Theorem 4.28: when $\mathcal{U}$ and $\mathcal{Y}$ are Hilbert spaces, given a compact, positive, self-adjoint operator R on $\mathcal{U}$, we seek

$$\arg\min \left\{ \left\| y - H(u) \right\|_{\mathcal{Y}}^2 + \left\| R^{-1/2}(u - \bar{u}) \right\|_{\mathcal{U}}^2 \,\middle|\, u \in \mathcal{U} \right\}.$$

The operator R describes the structure of the regularization, which in some sense is the practitioner's 'prior belief about what the solution should look like'. More generally, since it might be desired to weight the various components of y differently from the given Hilbert norm on $\mathcal{Y}$, we might seek

$$\arg\min \left\{ \left\| Q^{-1/2}(y - H(u)) \right\|_{\mathcal{Y}}^2 + \left\| R^{-1/2}(u - \bar{u}) \right\|_{\mathcal{U}}^2 \,\middle|\, u \in \mathcal{U} \right\}$$

for a given positive self-adjoint operator Q on $\mathcal{Y}$. However, this approach all appears to be somewhat ad hoc, especially where the choice of regularization is concerned.

Taking a probabilistic — specifically, Bayesian — viewpoint alleviates these difficulties. If we think of u and y as random variables, then (6.2) defines the conditioned random variable $y|u$, and we define the 'solution' of the inverse problem to be the conditioned random variable $u|y$. This allows us to model the noise, η, via its statistical properties, even if we do not know the exact instance of η that corrupted 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 essence of the Bayesian approach to inverse problems.

Remark 6.1. In practice the true observation operator is often approximated by some numerical model $H(\,\cdot\,; h)$, where h denotes a mesh parameter, or parameter controlling missing physics. In this case (6.2) becomes

$$y = H(u; h) + \varepsilon + \eta,$$

where $\varepsilon := H(u) - H(u; h)$. In principle, the observational noise η and the computational error ε could be combined into a single term, but keeping them separate is usually more appropriate: unlike η, ε is typically not of mean zero, and is dependent upon u.

To illustrate the central role that least squares minimization plays in elementary statistical estimation, and hence motivate the more general considerations of the rest of the chapter, consider the following finite-dimensional linear problem. Suppose that we are interested in learning some vector of parameters $u \in \mathbb{R}^n$, which gives rise to a vector $y \in \mathbb{R}^m$ of observations via

$$y = Au + \eta,$$

where $A \in \mathbb{R}^{m \times n}$ is a known linear operator (matrix) and η is a (not necessarily Gaussian) *noise vector* known to have mean zero and symmetric, positive-definite covariance matrix $Q := \mathbb{E}[\eta \otimes \eta] \equiv \mathbb{E}[\eta\eta^*] \in \mathbb{R}^{m \times m}$, with η independent of u. A common approach is to seek an estimate $\hat{u}$ of u that is a *linear* function Ky of the data y is *unbiased* in the sense that $\mathbb{E}[\hat{u}] = u$, and is the *best* estimate in that it minimizes an appropriate cost function. The following theorem, the Gauss–Markov theorem, states that there is precisely one such estimator, and it is the solution to the weighted least squares problem with weight Q^{-1}, i.e.

$$\hat{u} = \operatorname*{arg\,min}_{u \in \mathcal{H}} J(u), \quad J(u) := \frac{1}{2} \| Au - y \|_{Q^{-1}}^2.$$

In fact, this result holds true even in the setting of Hilbert spaces:

Theorem 6.2 (Gauss–Markov)**.** *Let $\mathcal{H}$ and $\mathcal{K}$ be separable Hilbert spaces, and let $A \colon \mathcal{H} \to \mathcal{K}$. Let $u \in \mathcal{H}$ and let $y = Au + \eta$, where η is a centred $\mathcal{K}$-valued random variable with self-adjoint and positive definite covariance operator Q. Suppose that $Q^{1/2}A$ has closed range and that $A^*Q^{-1}A$ is invertible. Then, among all unbiased linear estimators $K \colon \mathcal{K} \to \mathcal{H}$, producing an estimate $\hat{u} = Ky$ of u given y, the one that minimizes both the mean-squared error $\mathbb{E}[\|\hat{u} - u\|^2]$ and the covariance operator[2] $\mathbb{E}[(\hat{u} - u) \otimes (\hat{u} - u)]$ is*

$$K = (A^*Q^{-1}A)^{-1}A^*Q^{-1}, \tag{6.3}$$

and the resulting estimate $\hat{u}$ has $\mathbb{E}[\hat{u}] = u$ and covariance operator

$$\mathbb{E}[(\hat{u} - u) \otimes (\hat{u} - u)] = (A^*Q^{-1}A)^{-1}.$$

Remark 6.3. Indeed, by Theorem 4.28, $\hat{u} = (A^*Q^{-1}A)^{-1}A^*Q^{-1}y$ is also the solution to the weighted least squares problem with weight Q^{-1}, i.e.

$$\hat{u} = \operatorname*{arg\,min}_{u \in \mathcal{H}} J(u), \quad J(u) := \frac{1}{2} \| Au - y \|_{Q^{-1}}^2.$$

Note that the first and second derivatives (gradient and Hessian) of J are

$$\nabla J(u) = A^*Q^{-1}Au - A^*Q^{-1}y, \quad \text{and} \quad \nabla^2 J(u) = A^*Q^{-1}A,$$

so the covariance of $\hat{u}$ is the inverse of the Hessian of J. These observations will be useful in the construction of the Kálmán filter in Chapter 7.

[2] Here, the minimization is meant in the sense of positive semi-definite operators: for two operators A and B, we say that $A \leq B$ if $B - A$ is a positive semi-definite operator.

Proof of Theorem 6.2. It is easily verified that K as given by (6.3) is an unbiased estimator:

$$\hat{u} = (A^*Q^{-1}A)^{-1}A^*Q^{-1}(Au+\eta) = u + (A^*Q^{-1}A)^{-1}A^*Q^{-1}\eta$$

and so, taking expectations of both sides and using the assumption that η is centred, $\mathbb{E}[\hat{u}] = u$. Moreover, the covariance of this estimator satisfies

$$\mathbb{E}[(\hat{u}-u)\otimes(\hat{u}-u)] = K\mathbb{E}[\eta\otimes\eta]K = (A^*Q^{-1}A)^{-1},$$

as claimed.

Now suppose that $L = K + D$ is any linear unbiased estimator; note that $DA = 0$. Then the covariance of the estimate Ly satisfies

$$\begin{aligned}\mathbb{E}[(Ly-u)\otimes(Ly-u)] &= \mathbb{E}[(K+D)\eta\otimes\eta(K^*+D^*)]\\ &= (K+D)Q(K^*+D^*)\\ &= KQK^* + DQD^* + KQD^* + (KQD^*)^*.\end{aligned}$$

Since $DA = 0$,

$$KQD^* = (A^*Q^{-1}A)^{-1}A^*Q^{-1}QD^* = (A^*Q^{-1}A)^{-1}(DA)^* = 0,$$

and so

$$\mathbb{E}[(Ly-u)\otimes(Ly-u)] = KQK^* + DQD^* \geq KQK^*.$$

Since DQD^* is self-adjoint and positive semi-definite, this shows that

$$\mathbb{E}[(Ly-u)\otimes(Ly-u)] \geq KQK^*. \qquad \square$$

Remark 6.4. In the finite-dimensional case, if $A^*Q^{-1}A$ is not invertible, then it is common to use the estimator

$$K = (A^*Q^{-1}A)^\dagger A^*Q^{-1},$$

where $B^\dagger$ denotes the *Moore–Penrose pseudo-inverse* of B, defined equivalently by

$$\begin{aligned}B^\dagger &:= \lim_{\delta\to 0}(B^*B+\delta I)B^*,\\ B^\dagger &:= \lim_{\delta\to 0}B^*(BB^*+\delta I)B^*, \text{ or}\\ B^\dagger &:= V\Sigma^\dagger U^*,\end{aligned}$$

where $B = U\Sigma V^*$ is the singular value decomposition of B, and $\Sigma^\dagger$ is the transpose of the matrix obtained from Σ by replacing all the strictly positive singular values by their reciprocals. In infinite-dimensional settings, the use of regularization and pseudo-inverses is a more subtle topic, especially when the noise η has degenerate covariance operator Q.

Bayesian Interpretation of Regularization. The Gauss–Markov estimator is not ideal: for example, because of its characterization as the minimizer of a quadratic cost function, it is sensitive to large outliers in the data, i.e. components of y that differ greatly from the corresponding component of $A\hat{u}$. In such a situation, it may be desirable to not try to fit the observed data y too closely, and instead *regularize* the problem by seeking $\hat{u}$, the minimizer of

$$J(u) := \frac{1}{2}\|Au - y\|_{Q^{-1}}^2 + \frac{1}{2}\|u - \bar{u}\|_{R^{-1}}^2, \tag{6.4}$$

for some chosen $\bar{u} \in \mathbb{K}^n$ and positive-definite *Tikhonov matrix* $R \in \mathbb{K}^{n\times n}$. Depending upon the relative sizes of Q and R, $\hat{u}$ will be influenced more by the data y and hence lie close to the Gauss–Markov estimator, or be influenced more by the regularization term and hence lie close to $\bar{u}$. At first sight this procedure may seem somewhat ad hoc, but it has a natural Bayesian interpretation.

Let us make the additional assumption that, not only is η centred with covariance operator Q, but it is in fact Gaussian. Then, to a Bayesian practitioner, the observation equation

$$y = Au + \eta$$

defines the conditional distribution $y|u$ as $(y - Au)|u = \eta \sim \mathcal{N}(0, Q)$. Finding the minimizer of $u \mapsto \frac{1}{2}\|Au - y\|_{Q^{-1}}^2$, i.e. $\hat{u} = Ky$, amounts to finding the *maximum likelihood estimator* of u given y. The Bayesian interpretation of the regularization term is that $\mathcal{N}(\bar{u}, R)$ is a prior distribution for u. The resulting posterior distribution for $u|y$ has Lebesgue density $\rho(u|y)$ with

$$\begin{aligned}
\rho(u|y) &\propto \exp\left(-\frac{1}{2}\|Au - y\|_{Q^{-1}}^2\right) \exp\left(-\frac{1}{2}\|u - \bar{u}\|_{R^{-1}}^2\right) \\
&= \exp\left(-\frac{1}{2}\|Au - y\|_{Q^{-1}}^2 - \frac{1}{2}\|u - \bar{u}\|_{R^{-1}}^2\right) \\
&= \exp\left(-\frac{1}{2}\|u - Ky\|_{A^*Q^{-1}A}^2 - \frac{1}{2}\|u - \bar{u}\|_{R^{-1}}^2\right) \\
&= \exp\left(-\frac{1}{2}\|u - P^{-1}(A^*Q^{-1}AKy + R^{-1}\bar{u})\|_{P}^2\right)
\end{aligned}$$

where, by Exercise 6.1, P is the precision matrix

$$P = A^*Q^{-1}A + R^{-1}.$$

The solution of the regularized least squares problem of minimizing the functional J in (6.4) — i.e. minimizing the exponent in the above posterior distribution — is the *maximum a posteriori estimator* of u given y. However, the full posterior contains more information than the MAP estimator alone. In particular, the posterior covariance matrix $P^{-1} = (A^*Q^{-1}A + R^{-1})^{-1}$ reveals those components of u about which we are relatively more or less certain.

Non-Quadratic Regularization and Recovery of Sparse Signals. This chapter mostly deals with the case in which both the noise model (i.e. the likelihood) and the prior are Gaussian measures, which is the same as saying that the maximum a posteriori estimator is obtained by minimizing the sum of the squares of two Hilbert norms, just as in (6.4). However, there is no fundamental reason not to consider other regularizations — or, in Bayesian terms, other priors. Indeed, in many cases an appropriate choice of prior is a probability distribution with both a heavy centre and a heavy tail, such as

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}u}(u) \propto \exp\left(- \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} \right)$$

on $\mathbb{R}^n$, for $0 < p < 1$. Such regularizations correspond to a prior belief that the u to be recovered from noisy observations y is *sparse*, in the sense that it has a simple low-dimensional structure, e.g. that most of its components in some coordinate system are zero.

For definiteness, consider a finite-dimensional example in which it is desired to recover $u \in \mathbb{K}^n$ from noisy observations $y \in \mathbb{K}^m$ of Au, where $A \in \mathbb{K}^{m \times n}$ is known. Let

$$\|u\|_0 := \#\big\{ i \in \{1, \dots, n\} \big| u_i \neq 0 \big\}.$$

(Note well that, despite the suggestive notation, $\|\cdot\|_0$ is *not* a norm, since in general $\|\lambda u\|_0 \neq |\lambda| \|u\|_0$.) If the corruption of Au into y occurs through additive Gaussian noise distributed according to $\mathcal{N}(0, Q)$, then the ordinary least squares estimate of u is found by minimizing $\frac{1}{2}\|Au - y\|_{Q^{-1}}^2$. However, a prior belief that u is sparse, i.e. that $\|u\|_0$ is small, is reflected in the regularized least squares problem

$$\text{find } u \in \mathbb{K}^n \text{ to minimize } J_0(u) := \frac{1}{2}\|Au - y\|_{Q^{-1}}^2 + \lambda \|u\|_0, \tag{6.5}$$

where $\lambda > 0$ is a regularization parameter. Unfortunately, problem (6.5) is very difficult to solve numerically, since the objective function is not convex. Instead, we consider

$$\text{find } u \in \mathbb{K}^n \text{ to minimize } J_1(u) := \frac{1}{2}\|Au - y\|_{Q^{-1}}^2 + \lambda \|u\|_1. \tag{6.6}$$

Remarkably, the two optimization problems (6.5) and (6.6) are ‘often’ equivalent in the sense of having the same minimizers; this near-equivalence can be made precise by a detailed probabilistic analysis using the so-called *restricted isometry property*, which will not be covered here, and is foundational in the field of *compressed sensing*. Regularization using the 1-norm amounts to putting a Laplace distribution Bayesian prior on u, and is known in the statistical regression literature as the LASSO (least absolute shrinkage and selection operator); in the signal processing literature, it is known as *basis pursuit denoising*.

For a heuristic understanding of why regularizing using the norm $\|\cdot\|_1$ promotes sparsity, let us consider an even more general problem: let $R\colon \mathbb{K}^n \to \mathbb{R}$ be any convex function, and consider the problem

$$\text{find } u \in \mathbb{K}^n \text{ to minimize } J_R(u) := \|Au - Y\|_{Q^{-1}}^2 + R(u), \tag{6.7}$$

which clearly includes (6.4) and (6.6) as special cases. Observe that, by writing $r = R(x)$ for the value of the regularization term, we have

$$\inf_{u \in \mathbb{K}^n} J_R(u) = \inf_{r \geq 0} \left(r + \inf_{u : R(u) = r} \|Au - b\|_{Q^{-1}}^2 \right). \tag{6.8}$$

The equality constraint in (6.8) can in fact be relaxed to an inequality:

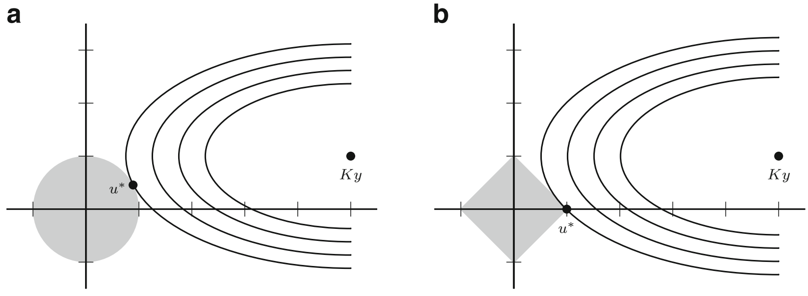
$$\inf_{u \in \mathbb{K}^n} J_R(u) = \inf_{r \geq 0} \left(r + \inf_{u : R(u) \leq r} \|Au - b\|_{Q^{-1}}^2 \right). \tag{6.9}$$

Note that convexity of R implies that $\{u \in \mathbb{K}^n \mid R(u) \leq r\}$ is a convex subset of $\mathbb{K}^n$. The reason for the equivalence of (6.8) and (6.9) is quite simple: if $(r, u) = (r^*, u^*)$ were minimal for the right-hand side and also $R(u^*) < r^*$, then the right-hand side could be reduced by considering instead $(r, u) = (R(u^*), u^*)$, which preserves the value of the quadratic term but decreases the regularization term. This contradicts the optimality of (r^*, u^*). Hence, in (6.9), we may assume that the optimizer has $R(u^*) = r^*$, which is exactly the earlier problem (6.8).

In the case that $R(u)$ is a multiple of the 1- or 2-norm of u, the region $R(u) \leq r$ is a norm ball centred on the origin, and the above arguments show that the minimizer u^* of J_1 or J_2 will be a boundary point of that ball. However, as indicated in Figure 6.1, in the 1-norm case, this u^* will 'typically' lie on one of the low-dimensional faces of the 1-norm ball, and so $\|u^*\|_0$ will be small and u^* will be sparse. There are, of course, y for which u^* is non-sparse, but this is the exception for 1-norm regularization, whereas it is the rule for ordinary 2-norm (Tikhonov) regularization.

6.2 Bayesian Inversion in Banach Spaces

This section concerns Bayesian inversion in Banach spaces, and, in particular, establishing the appropriate rigorous statement of Bayes' rule in settings where — by Theorem 2.38 — there is no Lebesgue measure with respect to which we can take densities. Therefore, in such settings, it is necessary to use as the prior a measure such as a Gaussian or Besov measure, often



Quadratic (ℓ^2) regularization.

Sparse (ℓ^1) regularization.

Fig. 6.1: Comparison of ℓ^2 versus ℓ^1 regularization of a least squares minimization problem. The shaded region indicates a norm ball centred on the origin for the appropriate regularizing norm. The black ellipses, centred on the unregularized least squares (Gauss–Markov) solution $Ky = (A^*Q^{-1}A)^{-1}A^*Q^{-1}y$, are contours of the original objective function, $u \mapsto \|Au - y\|^2_{Q^{-1}}$. By (6.9), the regularized solution u^* lies on the intersection of an objective function contour and the boundary of the regularization norm ball; for the 1-norm, u^* is sparse for 'most' y.

accessed through a sampling scheme such as a Karhunen–Loève expansion, as in Section 11.1. Note, however, then when the observation operator H is non-linear, although the prior may be a 'simple' Gaussian measure, the posterior will in general be a non-Gaussian measure with features such as multiple modes of different widths. Thus, the posterior is an object much richer in information than a simple maximum likelihood or maximum a posteriori estimator obtained from the optimization-theoretic point of view.

Example 6.5. There are many applications in which it is of interest to determine the permeability of subsurface rock, e.g. the prediction of transport of radioactive waste from an underground waste repository, or the optimization of oil recovery from underground fields. The flow velocity v of a fluid under pressure p in a medium or permeability κ is given by *Darcy's law*

$$v = -\kappa \nabla p.$$

The pressure field p within a bounded, open domain $\mathcal{X} \subset \mathbb{R}^d$ is governed by the elliptic PDE

$$-\nabla \cdot (\kappa \nabla p) = 0 \quad \text{in } \mathcal{X},$$

together with some boundary conditions, e.g. the Neumann (zero flux) boundary condition $\nabla p \cdot \hat{n}_{\partial \mathcal{X}} = 0$ on $\partial \mathcal{X}$; one can also consider a non-zero source term f on the right-hand side. For simplicity, take the permeability tensor field κ to be a scalar field k times the identity tensor; for mathematical and physical reasons, it is important that k be positive, so write $k = e^u$. The objective is to recover u from, say, observations of the pressure field at known points $x_1, \dots, x_m \in \mathcal{X}$:

$$y_i = p(x_i) + \eta_i.$$

Note that this fits the general '$y = H(u) + \eta$' setup, with H being defined implicitly by the solution operator to the elliptic boundary value problem.

In general, let u be a random variable with (prior) distribution μ_0 — which we do not at this stage assume to be Gaussian — on a separable Banach space $\mathcal{U}$. Suppose that we observe data $y \in \mathbb{R}^m$ according to (6.2), where η is an $\mathbb{R}^m$-valued random variable independent of u with probability density ρ with respect to Lebesgue measure. Let $\Phi(u; y)$ be any function that differs from $-\log \rho(y - H(u))$ by an additive function of y alone, so that

$$\frac{\rho(y - H(u))}{\rho(y)} \propto \exp(-\Phi(u; y))$$

with a constant of proportionality independent of u. An informal application of Bayes' rule suggests that the posterior probability distribution of u given y, $\mu^y \equiv \mu_0(\,\cdot\,|y)$, has Radon–Nikodým derivative with respect to the prior, μ_0, given by

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

The next theorem makes this argument rigorous:

Theorem 6.6 (Generalized Bayes' rule)**.** *Suppose that $H \colon \mathcal{U} \to \mathbb{R}^m$ is continuous, and that η is absolutely continuous with support $\mathbb{R}^m$. If $u \sim \mu_0$, then $u|y \sim \mu^y$, where $\mu^y \ll \mu_0$ and*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp(-\Phi(u; y)). \tag{6.10}$$

The proof of Theorem 6.6 uses the following technical lemma:

Lemma 6.7 (Dudley, 2002, Section 10.2)**.** *Let μ, ν be probability measures on $\mathcal{U} \times \mathcal{Y}$, where $(\mathcal{U}, \mathscr{A})$ and $(\mathcal{Y}, \mathscr{B})$ are measurable spaces. Assume that $\mu \ll \nu$ and that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \varphi$, and that the conditional distribution of $u|y$ under ν, denoted by $\nu^y(\mathrm{d}u)$, exists. Then the distribution of $u|y$ under μ, denoted $\mu^y(\mathrm{d}u)$, exists and $\mu^y \ll \nu^y$, with Radon–Nikodým derivative given by*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\nu^y}(u) = \begin{cases} \frac{\varphi(u,y)}{Z(y)}, & \text{if } Z(y) > 0, \\ 1, & \text{otherwise}, \end{cases}$$

where $Z(y) := \int_{\mathcal{U}} \varphi(u, y) \, \mathrm{d}\nu^y(u)$.

Proof of Theorem 6.6. Let $\mathbb{Q}_0(\mathrm{d}y) := \rho(y)\,\mathrm{d}y$ on $\mathbb{R}^m$ and $\mathbb{Q}(\mathrm{d}u|y) := \rho(y - H(u))\,\mathrm{d}y$, so that, by construction

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

Define measures ν_0 and ν on $\mathbb{R}^m \times \mathcal{U}$ by

$$\begin{aligned} \nu_0(\mathrm{d}y, \mathrm{d}u) &:= \mathbb{Q}_0(\mathrm{d}y) \otimes \mu_0(\mathrm{d}u), \\ \nu(\mathrm{d}y, \mathrm{d}u) &:= \mathbb{Q}_0(\mathrm{d}y|u)\mu_0(\mathrm{d}u). \end{aligned}$$

Note that ν_0 is a product measure under which u and y are independent, whereas ν is not. Since H is continuous, so is Φ; since $\mu_0(\mathcal{U}) = 1$, it follows that Φ is μ_0-measurable. Therefore, ν is well defined, $\nu \ll \nu_0$, and

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

Note that

$$\int_{\mathcal{U}} \exp(-\Phi(u;y))\,\mathrm{d}\mu_0(u) = C(y)\int_{\mathcal{U}} \rho(y - H(u))\,\mathrm{d}\mu_0(u) > 0,$$

since ρ is strictly positive on $\mathbb{R}^m$ and H is continuous. Since $\nu_0(\mathrm{d}u|y) = \mu_0(\mathrm{d}u)$, the result follows from Lemma 6.7. □

Exercise 6.2 shows that, if the prior μ_0 is a Gaussian measure and the potential Φ is quadratic in u, then, for all y, the posterior μ^y is Gaussian. In particular, if the observation operator is a continuous linear map and the observations are corrupted by additive Gaussian noise, then the posterior is Gaussian — see Exercise 2.8 for the relationships between the means and covariances of the prior, noise and posterior. On the other hand, if either the observation operator is non-linear or the observational noise is non-Gaussian, then a Gaussian prior is generally transformed into a non-Gaussian posterior.

6.3 Well-Posedness and Approximation

This section concerns the well-posedness of the Bayesian inference problem for Gaussian priors on Banach spaces. To save space later on, the following will be taken as our *standard assumptions* on the negative log-likelihood/potential Φ. In essence, we wish to restrict attention to potentials Φ that are Lipschitz in both arguments, bounded on bounded sets, and that do not decay to $-\infty$ at infinity 'too quickly'.

Assumptions on Φ. Assume that $\Phi \colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ satisfies:

(A1) For every $\varepsilon > 0$ and $r > 0$, there exists $M = M(\varepsilon, r) \in \mathbb{R}$ such that, for all $u \in \mathcal{U}$ and all $y \in \mathcal{Y}$ with $\|y\|_{\mathcal{Y}} < r$,

$$\Phi(u; y) \geq M - \varepsilon \|u\|_{\mathcal{U}}^2.$$

(A2) For every $r > 0$, there exists $K = K(r) > 0$ such that, for all $u \in \mathcal{U}$ and all $y \in \mathcal{Y}$ with $\|u\|_{\mathcal{U}}, \|y\|_{\mathcal{Y}} < r$,

$$\Phi(u; y) \leq K.$$

(A3) For every $r > 0$, there exists $L = L(r) > 0$ such that, for all $u_1, u_2 \in \mathcal{U}$ and all $y \in \mathcal{Y}$ with $\|u_1\|_{\mathcal{U}}, \|u_2\|_{\mathcal{U}}, \|y\|_{\mathcal{Y}} < r$,

$$\big|\Phi(u_1; y) - \Phi(u_2; y)\big| \leq L \big\|u_1 - u_2\big\|_{\mathcal{U}}.$$

(A4) For every $\varepsilon > 0$ and $r > 0$, there exists $C = C(\varepsilon, r) > 0$ such that, for all $u \in \mathcal{U}$ and all $y_1, y_2 \in \mathcal{Y}$ with $\|y_1\|_{\mathcal{Y}}, \|y_2\|_{\mathcal{Y}} < r$,

$$\big|\Phi(u; y_1) - \Phi(u; y_2)\big| \leq \exp\big(\varepsilon \|u\|_{\mathcal{U}}^2 + C\big) \big\|y_1 - y_2\big\|_{\mathcal{Y}}.$$

We first show that, for Gaussian priors, these assumptions yield a well-defined posterior measure for each possible instance of the observed data:

Theorem 6.8. *Let Φ satisfy standard assumptions* (A1), (A2), *and* (A3) *and assume that μ_0 is a Gaussian probability measure on $\mathcal{U}$. Then, for each $y \in \mathcal{Y}$, μ^y given by*

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

is a well-defined probability measure on $\mathcal{U}$.

Proof. Assumption (A2) implies that $Z(y)$ is bounded below:

$$Z(y) \geq \int_{\{u \mid \|u\|_{\mathcal{U}} \leq r\}} \exp(-K(r)) \, \mathrm{d}\mu_0(u) = \exp(-K(r)) \mu_0\big[\|u\|_{\mathcal{U}} \leq r\big] > 0$$

for $r > 0$, since μ_0 is a strictly positive measure on $\mathcal{U}$. By (A3), Φ is μ_0-measurable, and so μ^y is a well-defined measure. By (A1), for $\|y\|_{\mathcal{Y}} \leq r$ and ε sufficiently small,

$$\begin{aligned} Z(y) &= \int_{\mathcal{U}} \exp(-\Phi(u; y)) \, \mathrm{d}\mu_0(u) \\ &\leq \int_{\mathcal{U}} \exp(\varepsilon \|u\|_{\mathcal{U}}^2 - M(\varepsilon, r)) \, \mathrm{d}\mu_0(u) \\ &\leq C \exp(-M(\varepsilon, r)) < \infty, \end{aligned}$$

since μ_0 is Gaussian and we may choose ε small enough that the Fernique theorem (Theorem 2.47) applies. Thus, μ^y can indeed be normalized to be a probability measure on $\mathcal{U}$. □

Recall from Chapter 5 that the *Hellinger distance* between two probability measures μ and ν on $\mathcal{U}$ is defined in terms of any reference measure ρ with respect to which both μ and ν are absolutely continuous by

$$d_{\mathrm{H}}(\mu, \nu) := \sqrt{\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(u)} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\rho}(u)} \right|^2 \mathrm{d}\rho(u)}.$$

A particularly useful property of the Hellinger metric is that closeness in the Hellinger metric implies closeness of expected values of polynomially bounded functions: if $f \colon \mathcal{U} \to \mathcal{V}$, for some Banach space $\mathcal{V}$, then Proposition 5.12 gives that

$$\big\| \mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f] \big\| \leq 2 \sqrt{\mathbb{E}_{\mu}\big[\|f\|^2\big] + \mathbb{E}_{\nu}\big[\|f\|^2\big]}\, d_{\mathrm{H}}(\mu, \nu).$$

Therefore, Hellinger-close prior and posterior measures give similar expected values to quantities of interest; indeed, for fixed f, the perturbation in the expected value is Lipschitz with respect to the Hellinger size of the perturbation in the measure.

The following theorem shows that Bayesian inference with respect to a Gaussian prior measure is well-posed with respect to perturbations of the observed data y, in the sense that the Hellinger distance between the corresponding posteriors is Lipschitz in the size of the perturbation in the data:

Theorem 6.9. *Let Φ satisfy the standard assumptions* (A1), (A2), *and* (A4), *suppose that μ_0 is a Gaussian probability measure on $\mathcal{U}$, and that $\mu^y \ll \mu_0$ with density given by the generalized Bayes' rule for each $y \in \mathcal{Y}$. Then there exists a constant $C \geq 0$ such that, for all $y, y' \in \mathcal{Y}$,*

$$d_{\mathrm{H}}(\mu^y, \mu^{y'}) \leq C \|y - y'\|_{\mathcal{Y}}.$$

Proof. As in the proof of Theorem 6.8, (A2) gives a lower bound on $Z(y)$. We also have the following Lipschitz continuity estimate for the difference between the normalizing constants for y and y':

$$\begin{aligned} &|Z(y) - Z(y')| \\ &\quad \leq \int_{\mathcal{U}} \big| e^{-\Phi(u;y)} - e^{-\Phi(u;y')} \big| \,\mathrm{d}\mu_0(u) \\ &\quad \leq \int_{\mathcal{U}} \max\big\{ e^{-\Phi(u;y)}, e^{-\Phi(u;y')} \big\} \big| \Phi(u;y) - \Phi(u;y') \big| \,\mathrm{d}\mu_0(u) \end{aligned}$$

by the mean value theorem (MVT). Hence,

$$\begin{aligned}
&|Z(y) - Z(y')| \\
&\quad \leq \int_{\mathcal{U}} e^{\varepsilon\|u\|_{\mathcal{U}}^2 + M} \cdot e^{\varepsilon\|u\|_{\mathcal{U}}^2 + C} \|y - y'\|_{\mathcal{Y}} \, \mathrm{d}\mu_0(u) && \text{by (A1) and (A4)} \\
&\quad \leq C\|y - y'\|_{\mathcal{Y}} && \text{by Fernique.}
\end{aligned}$$

By the definition of the Hellinger distance, using the prior μ_0 as the reference measure,

$$\begin{aligned}
d_{\mathrm{H}}(\mu^y, \mu^{y'})^2 &= \int_{\mathcal{U}} \left| \frac{1}{\sqrt{Z(y)}} e^{-\Phi(u;y)/2} - \frac{1}{\sqrt{Z(y')}} e^{-\Phi(u;y')/2} \right|^2 \mathrm{d}\mu_0(u) \\
&= \frac{1}{Z(y)} \int_{\mathcal{U}} \left| e^{-\Phi(u;y)/2} - \sqrt{\frac{Z(y)}{Z(y')}} e^{-\Phi(u;y')/2} \right|^2 \mathrm{d}\mu_0(u) \\
&\leq I_1 + I_2,
\end{aligned}$$

where

$$\begin{aligned}
I_1 &:= \frac{1}{Z(y)} \int_{\mathcal{U}} \left| e^{-\Phi(u;y)/2} - e^{-\Phi(u;y')/2} \right|^2 \mathrm{d}\mu_0(u), \\
I_2 &:= \left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right|^2 \int_{\mathcal{U}} e^{-\Phi(u;y')/2} \, \mathrm{d}\mu_0(u).
\end{aligned}$$

For I_1, a similar application of the MVT, (A1) and (A4), and the Fernique theorem to the one above yields that

$$\begin{aligned}
I_1 &\leq \frac{1}{Z(y)} \int_{\mathcal{U}} \max\left\{ \tfrac{1}{2} e^{-\Phi(u;y)/2}, \tfrac{1}{2} e^{-\Phi(u;y')/2} \right\}^2 \cdot \left| \Phi(u;y) - \Phi(u;y') \right|^2 \mathrm{d}\mu_0(u) \\
&\leq \frac{1}{4Z(y)} \int_{\mathcal{U}} e^{\varepsilon\|u\|_{\mathcal{U}}^2 + M} \cdot e^{2\varepsilon\|u\|_{\mathcal{U}}^2 + 2C} \|y - y'\|_{\mathcal{Y}}^2 \, \mathrm{d}\mu_0(u) \\
&\leq C\|y - y'\|_{\mathcal{Y}}^2.
\end{aligned}$$

A similar application of (A1) and the Fernique theorem shows that the integral in I_2 is finite. Also, the lower bound on $Z(\,\cdot\,)$ implies that

$$\begin{aligned}
\left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right|^2 &\leq C \max\left\{ \frac{1}{Z(y)^3}, \frac{1}{Z(y')^3} \right\} |Z(y) - Z(y')|^2 \\
&\leq C\|y - y'\|_{\mathcal{Y}}^2.
\end{aligned}$$

Thus, $I_2 \leq C\|y - y'\|_{\mathcal{Y}}^2$, which completes the proof. □

Similarly, the next theorem shows that Bayesian inference with respect to a Gaussian prior measure is well-posed with respect to approximation of measures and log-likelihoods. The approximation of Φ by some Φ^N typically arises through the approximation of H by some discretized numerical model H^N. The importance of Theorem 6.10 is that it allows error estimates for the *forward* models H and H^N, which typically arise through non-probabilistic numerical analysis, to be translated into error estimates for the Bayesian *inverse* problem.

Theorem 6.10. *Suppose that the probability measures μ and μ^N are the posteriors arising from potentials Φ and Φ^N and are all absolutely continuous with respect to μ_0, and that Φ, Φ^N satisfy the standard assumptions* (A1) *and* (A2) *with constants uniform in N. Assume also that, for all $\varepsilon > 0$, there exists $K = K(\varepsilon) > 0$ such that*

$$\left|\Phi(u;y) - \Phi^N(u;y)\right| \leq K \exp(\varepsilon \|u\|_{\mathcal{U}}^2)\psi(N), \tag{6.11}$$

where $\lim_{N\to\infty} \psi(N) = 0$. *Then there is a constant C, independent of N, such that*

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

Proof. Exercise 6.4. □

Remark 6.11. Note well that, regardless of the value of the observed data y, the Bayesian posterior μ^y is absolutely continuous with respect to the prior μ_0 and, in particular, cannot associate positive posterior probability with any event of prior probability zero. However, the Feldman–Hájek theorem (Theorem 2.51) says that it is very difficult for probability measures on infinite-dimensional spaces to be absolutely continuous with respect to one another. Therefore, the choice of infinite-dimensional prior μ_0 is a very strong modelling assumption that, if it is 'wrong', cannot be 'corrected' even by large amounts of data y. In this sense, it is not reasonable to expect that Bayesian inference on function spaces should be well-posed with respect to apparently small perturbations of the prior μ_0, e.g. by a shift of mean that lies outside the Cameron–Martin space, or a change of covariance arising from a non-unit dilation of the space. Nevertheless, the infinite-dimensional perspective is not without genuine fruits: in particular, the well-posedness results (Theorems 6.9 and 6.10) are very important for the design of finite-dimensional (discretized) Bayesian problems that have good stability properties with respect to discretization dimension N.

6.4 Accessing the Bayesian Posterior Measure

For given data $y \in \mathcal{Y}$, the Bayesian posterior $\mu_0(\,\cdot\,|y)$ on $\mathcal{U}$ is determined as a measure that has a density with respect to the prior μ_0 given by Bayes' rule, e.g. in the form (6.10),

$$\frac{\mathrm{d}\mu_0(\,\cdot\,|y)}{\mathrm{d}\mu_0}(u) \propto \exp(-\Phi(u;y)).$$

The results outlined above have shown some of the analytical properties of this construction. However, in practice, this well-posedness theory is not the end of the story, principally because we need to be able to *access* this posterior measure: in particular, it is necessary to be able to (numerically) integrate with respect to the posterior, in order to form the posterior expected value of quantities of interest. (Note, for example, that (6.10) gives a non-normalized density for the posterior with respect to the prior, and this lack of normalization is sometimes an additional practical obstacle.)

The general problem of how to access the Bayesian posterior measure is a complicated and interesting one. Roughly speaking, there are three classes of methods for exploration of the posterior, some of which will be discussed in depth at appropriate points later in the book:

(a) Methods such as Markov chain Monte Carlo, to be discussed in Chapter 9, attempt to sample from the posterior directly, using the formula for its density with respect to the prior.
In principle, one could also integrate with respect to the posterior by drawing samples from some other measure (e.g. the prior, or some other reference measure) and then re-weighting according to the appropriate probability density. However, some realizations of the data may cause the density $\mathrm{d}\mu_0(\,\cdot\,|y)/\mathrm{d}\mu_0$ to be significantly different from 1 for most draws from the prior, leading to severe ill-conditioning. For this reason, 'direct' draws from the posterior are highly preferable.
An alternative to re-weighting of prior samples is to transform prior samples into posterior samples while preserving their probability weights. That is, one seeks a function $T^y \colon \mathcal{U} \to \mathcal{U}$ from the parameter space $\mathcal{U}$ to itself that pushes forward any prior to its corresponding posterior, i.e. $T^y_* \mu_0 = \mu_0(\,\cdot\,|y)$, and hence turns an ensemble $\{u^{(1)}, \dots, u^{(N)}\}$ of independent samples distributed according to the prior into an ensemble $\{T^y(u^{(1)}), \dots, T^y(u^{(N)})\}$ of independent samples distributed according to the posterior. Map-based approaches to Bayesian inference include the approach of El Moselhy and Marzouk (2012), grounded in optimal transportation theory, and will not be discussed further here.

(b) A second class of methods attempts to approximate the posterior, often through approximating the forward and observation models, and hence the likelihood. Many of the modelling methods discussed in Chapters 10–13 are examples of such approaches. For example, the Gauss–Markov theorem (Theorem 6.2) and Linear Kálmán Filter (see Section 7.2) provide optimal approximations of the posterior within the class of Gaussian measures, with linear forward and observation operators.

(c) Finally, as a catch-all term, there are the 'ad hoc' methods. In this category, we include the Ensemble Kálmán Filter of Evensen (2009), which will be discussed in Section 7.4.

6.5 Frequentist Consistency of Bayesian Methods

A surprisingly subtle question about Bayesian inference is whether it yields the ‘correct’ result, regardless of the prior used, when exposed to enough sample data. Clearly, when very few data points have been observed, the prior controls the posterior much more strongly than the observed data do, so it is necessary to answer such questions in an asymptotic limit. It is also necessary to clarify what is meant by ‘correctness’. One such notion is that of *frequentist consistency*:

> “While for a Bayesian statistician the analysis ends in a certain sense with the posterior, one can ask interesting questions about the properties of posterior-based inference from a frequentist point of view.” (Nickl, 2013)

To describe frequentist consistency, consider the standard setup of a Bayesian prior μ_0 on some space $\mathcal{U}$, together with a Bayesian likelihood model for observed data with values in another space $\mathcal{Y}$, i.e. a family of probability measures $\mu(\,\cdot\,|u) \in \mathcal{M}_1(\mathcal{Y})$ indexed by $u \in \mathcal{U}$. Now introduce a new ingredient, which is a probability measure $\mu^\dagger \in \mathcal{M}_1(\mathcal{Y})$ that is treated as the ‘truth’ in the sense that the observed data are in fact a sequence of independent and identically distributed draws from $\mu^\dagger$.

Definition 6.12. The likelihood model $\{\mu(\,\cdot\,|u) \mid u \in \mathcal{U}\}$ is said to be *well-specified* if there exists some $u^\dagger \in \mathcal{U}$ such that $\mu^\dagger = \mu(\,\cdot\,|u^\dagger)$, i.e. if there is some member of the model family that exactly coincides with the data-generating distribution. If the model is not well-specified, then it is said to be *misspecified*.

In the well-specified case, the model and the parameter space $\mathcal{U}$ admit some $u^\dagger$ that explains the frequentist ‘truth’ $\mu^\dagger$. The natural question to ask is whether exposure to enough independent draws $Y_1, \dots, Y_n$ from $\mu^\dagger$ will permit the model to identify $u^\dagger$ out of all the other possible $u \in \mathcal{U}$. If some sequence of estimators or other objects (such as Bayesian posteriors) converges as $n \to \infty$ to $u^\dagger$ with respect to some notion of convergence, then the estimator is said to be *consistent*. For example, Theorem 6.13 gives conditions for the maximum likelihood estimator (MLE) to be consistent, with the mode of convergence being convergence in probability; Theorem 6.17 (the Bernstein–von Mises theorem) gives conditions for the Bayesian posterior to be consistent, with the mode of convergence being convergence in probability, and with respect to the total variation distance on probability measures.

In order to state some concrete results on consistency, suppose now that $\mathcal{U} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^d$, and that the likelihood model $\{\mu(\,\cdot\,|u) \mid u \in \mathcal{U}\}$ can be written in the form of a parametric family of probability density functions with respect to Lebesgue measure on $\mathbb{R}^d$, which will be denoted by a function $f(\,\cdot\,|\,\cdot\,)\colon \mathcal{Y} \times \mathcal{U} \to [0, \infty)$, i.e.

$$\mu(E|u) = \int_E f(y|u)\,\mathrm{d}y \quad \text{for each measurable } E \subseteq \mathcal{Y} \text{ and each } u \in \mathcal{U}.$$

Before giving results about the convergence of the Bayesian posterior, we first state a result about the convergence of the *maximum likelihood estimator* (MLE) $\widehat{u}_n$ for $u^\dagger$ given the data $Y_1, \dots, Y_n \sim \mu^\dagger$, which, as the name suggests, is defined by

$$\widehat{u}_n \in \operatorname*{arg\,max}_{u \in \mathcal{U}} f(Y_1|u) \cdots f(Y_n|u).$$

Note that, being a function of the random variables $Y_1, \dots, Y_n$, $\widehat{u}_n$ is itself a random variable.

Theorem 6.13 (Consistency of the MLE)**.** *Suppose that $f(y|u) > 0$ for all $(u, y) \in \mathcal{U} \times \mathcal{Y}$, that $\mathcal{U}$ is compact, and that parameters $u \in \mathcal{U}$ are identifiable in the sense that*

$$f(\,\cdot\,|u_0) = f(\,\cdot\,|u_1) \text{ Lebesgue a.e.} \iff u_0 = u_1$$

and that

$$\int_{\mathcal{Y}} \sup_{u \in \mathcal{U}} |\log f(y|u)| f(y|u^\dagger)\,\mathrm{d}y < \infty.$$

Then the maximum likelihood estimator $\widehat{u}_n$ converges to $u^\dagger$ in probability, i.e. for all $\varepsilon > 0$,

$$\mathbb{P}_{Y_i \sim \mu^\dagger}\left[\left|\widehat{u}_n - u^\dagger\right| > \varepsilon\right] \xrightarrow[n \to \infty]{} 0. \tag{6.12}$$

The proof of Theorem 6.13 is omitted, and can be found in Nickl (2013). The next two results quantify the convergence of the MLE and Bayesian posterior in terms of the following matrix:

Definition 6.14. The *Fisher information matrix* $i_{\mathrm{F}}(u^\dagger) \in \mathbb{R}^{p \times p}$ of f at $u^\dagger \in \mathcal{U}$ is defined by

$$i_{\mathrm{F}}(u^\dagger)_{ij} := \mathbb{E}_{Y \sim f(\,\cdot\,|u^\dagger)}\left[\left.\frac{\partial \log f(Y|u)}{\partial u_i}\frac{\partial \log f(Y|u)}{\partial u_j}\right|_{u=u^\dagger}\right]. \tag{6.13}$$

Remark 6.15. Under the regularity conditions that will be used later, $i_{\mathrm{F}}(u^\dagger)$ is a symmetric and positive-definite matrix, and so can be viewed as a Riemannian metric tensor on $\mathcal{U}$, varying from one point $u^\dagger \in \mathcal{U}$ to another. In that context, it is known as the *Fisher–Rao metric tensor*, and plays an important role the field of information geometry in general, and geodesic Monte Carlo methods in particular.

The next two results, the lengthy proofs of which are also omitted, are both asymptotic normality results. The first shows that the error in the MLE is asymptotically a normal distribution with covariance operator given by the Fisher information; informally, for large n, $\widehat{u}_n$ is normally distributed with mean $u^\dagger$ and precision $n i_{\mathrm{F}}(u^\dagger)$. The second result — the celebrated

Bernstein–von Mises theorem or Bayesian CLT (central limit theorem) — shows that the entire Bayesian posterior distribution is asymptotically a normal distribution centred on the MLE, which, under the conditions of Theorem 6.13, converges to the frequentist 'truth'. These results hold under suitable regularity conditions on the likelihood model, which are summarized here for later reference:

Regularity Assumptions. The parametric family $f\colon \mathcal{Y} \times \mathcal{U} \to [0, \infty)$ will be said to *satisfy the regularity assumptions* with respect to a data-generating distribution $\mu^\dagger \in \mathcal{M}_1(\mathcal{Y})$ if

(a) for all $u \in \mathcal{U}$ and $y \in \mathcal{Y}$, $f(y|u) > 0$;

(b) the model is well-specified, with $\mu^\dagger = \mu(\,\cdot\,|u^\dagger)$, where $u^\dagger$ is an interior point of $\mathcal{U}$;

(c) there exists an open set U with $u^\dagger \in U \subseteq \mathcal{U}$ such that, for each $y \in \mathcal{Y}$, $f(y|\,\cdot\,) \in \mathcal{C}^2(U; \mathbb{R})$;

(d) $\mathbb{E}_{Y \sim \mu^\dagger}[\nabla_u^2 \log f(Y|u)|_{u=u^\dagger}] \in \mathbb{R}^{p \times p}$ is non-singular and

$$\mathbb{E}_{Y \sim \mu^\dagger} \left[\left\| \nabla_u \log f(Y|u) \big|_{u=u^\dagger} \right\|^2 \right] < \infty;$$

(e) there exists $r > 0$ such that $B = \mathbb{B}_r(u^\dagger) \subseteq U$ and

$$\mathbb{E}_{Y \sim \mu^\dagger} \left[\sup_{u \in B} \nabla_u^2 \log f(Y|u) \right] < \infty,$$

$$\int_{\mathcal{Y}} \sup_{u \in B} \left\| \nabla_u \log f(Y|u) \right\| \, \mathrm{d}y < \infty,$$

$$\int_{\mathcal{Y}} \sup_{u \in B} \left\| \nabla_u^2 \log f(Y|u) \right\| \, \mathrm{d}y < \infty.$$

Theorem 6.16 (Local asymptotic normality of the MLE)**.** *Suppose that f satisfies the regularity assumptions. Then the Fisher information matrix* (6.13) *satisfies*

$$i_{\mathrm{F}}(u^\dagger)_{ij} = -\mathbb{E}_{Y \sim f(\,\cdot\,|u^\dagger)} \left[\left. \frac{\partial^2 \log f(Y|u)}{\partial u_i \partial u_j} \right|_{u=u^\dagger} \right]$$

and the maximum likelihood estimator satisfies

$$\sqrt{n}\big(\widehat{u}_n - u^\dagger\big) \xrightarrow[n \to \infty]{d} X \sim \mathcal{N}(0, i_{\mathrm{F}}(u^\dagger)^{-1}), \tag{6.14}$$

where $\xrightarrow{d}$ denotes convergence in distribution (also known as weak convergence, q.v. Theorem 5.14), i.e. $X_n \xrightarrow{d} X$ if $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$ for all bounded continuous functions $\varphi\colon \mathbb{R}^p \to \mathbb{R}$.

Theorem 6.17 (Bernstein–von Mises)**.** *Suppose that f satisfies the regularity assumptions. Suppose that the prior $\mu_0 \in \mathcal{M}_1(\mathcal{U})$ is absolutely continuous* *with respect to Lebesgue measure and has* $u^\dagger \in \operatorname{supp}(\mu_0)$*. Suppose also that the model admits a uniformly consistent estimator, i.e. a* $T_n \colon \mathcal{Y}^n \to \mathbb{R}^p$ *such that, for all* $\varepsilon > 0$*,*

$$\sup_{u\in\mathcal{U}} \mathbb{P}_{Y_i\sim f(\,\cdot\,|u)}\left[\big\|T_n(Y_1,\dots,Y_n)-u\big\| > \varepsilon\right] \xrightarrow[n\to\infty]{} 0. \tag{6.15}$$

Let $\mu_n := \mu_0(\,\cdot\,|Y_1,\dots,Y_n)$ *denote the (random) posterior measure obtained by conditioning* μ_0 *on* n *independent* $\mu^\dagger$*-distributed samples* Y_i*. Then, for all* $\varepsilon > 0$*,*

$$\mathbb{P}_{Y_i\sim\mu^\dagger}\left[\left\|\mu_n - \mathcal{N}\left(\widehat{u}_n, \frac{i_{\mathrm{F}}(u^\dagger)^{-1}}{n}\right)\right\|_{\mathrm{TV}} > \varepsilon\right] \xrightarrow[n\to\infty]{} 0. \tag{6.16}$$

The Bernstein–von Mises theorem is often interpreted as saying that so long as the prior μ_0 is strictly positive — i.e. puts positive probability mass on every open set in $\mathcal{U}$ — the Bayesian posterior will asymptotically put all its mass on the frequentist 'truth' $u^\dagger$ (assuming, of course, that $u^\dagger \in \mathcal{U}$). Naturally, if $u^\dagger \notin \operatorname{supp}(\mu_0)$, then there is no hope of learning $u^\dagger$ in this way, since the posterior is always absolutely continuous with respect to the prior, and so cannot put mass where the prior does not. Therefore, it seems sensible to use 'open-minded' priors that are everywhere strictly positive; Lindley (1985) calls this requirement "Cromwell's Rule" in reference to Oliver Cromwell's famous injunction to the Synod of the Church of Scotland in 1650:

> "I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

Unfortunately, the Bernstein–von Mises theorem is no longer true when the space $\mathcal{U}$ is infinite-dimensional, and Cromwell's Rule is not a sufficient condition for consistency. In infinite-dimensional spaces, there are counterexamples in which the posterior either fails to converge or converges to something other than the 'true' parameter value — the latter being a particularly worrisome situation, since then a Bayesian practitioner will become more and more convinced of a *wrong answer* as more data come in. There are, however, some infinite-dimensional situations in which consistency properties do hold. In general, the presence or absence of consistency depends in subtle ways upon choices such as the topology of convergence of measures, and the types of sets for which one requires posterior consistency. See the bibliography at the end of the chapter for further details.

6.6 Bibliography

In finite-dimensional settings, the Gauss–Markov theorem is now classical, but the extension to Hilbert spaces appears is due to Beutler and Root (1976), with further extensions to degenerate observational noise due to Morley (1979). Quadratic regularization, used to restore well-posedness to ill-posed inverse problems, was introduced by Tikhonov (1943, 1963). An introduction to the general theory of regularization and its application to inverse problems is given by Engl et al. (1996). Tarantola (2005) and Kaipio and Somersalo (2005) provide a good introduction to the Bayesian approach to inverse problems, with the latter being especially strong in the context of differential equations.

The papers of Stuart (2010) and Cotter et al. (2009, 2010) set out the common structure of Bayesian inverse problems on Banach and Hilbert spaces, focussing on Gaussian priors. Theorems 6.8, 6.9, and 6.10 are Theorems 4.1, 4.2, and 4.6 respectively in Stuart (2010). Stuart (2010) stresses the importance of delaying discretization to the last possible moment, much as in PDE theory, lest one carelessly end up with a family of finite-dimensional problems that are individually well-posed but collectively ill-conditioned as the discretization dimension tends to infinity. Extensions to Besov priors, which are constructed using wavelet bases of L^2 and allow for non-smooth local features in the random fields, can be found in the articles of Dashti et al. (2012) and Lassas et al. (2009).

Probabilistic treatments of the deterministic problems of numerical analysis — including quadrature and the solution of ODEs and PDEs — date back to Poincaré (1896), but find their modern origins in the papers of Diaconis (1988), O'Hagan (1992), and Skilling (1991). More recent works examining the statistical character of discretization error for ODE and PDE solvers, its impact on Bayesian inferences, and the development of probabilistic solvers for deterministic differential equations, include those of Schober et al. (2014) and the works listed as part of the Probabilistic Numerics project http://www.probabilistic-numerics.org.

The use of 1-norm regularization was introduced in the statistics literature as the LASSO by Tibshirani (1996), and in the signal processing literature as basis pursuit denoising by Chen et al. (1998). The high-probability equivalence of 1-norm and 0-norm regularization, and the consequences for the recovery of sparse signals using compressed sensing, are due to Candès et al. (2006a,b). Chandrasekaran et al. (2012) give a general framework for the convex geometry of sparse linear inverse problems. An alternative paradigm for promoting sparsity in optimization and statistical inference problems is the reversible-jump Markov chain Monte Carlo method of Green (1995).

The classic introductory text on information geometry, in which the Fisher–Rao metric tensor plays a key role, is the book of Amari and Nagaoka (2000). Theorems 6.13, 6.16, and 6.17 on MLE and Bayesian posterior consistency are Theorems 2, 3, and 5 respectively in Nickl (2013), and their proofs can be found there. The study of the frequentist consistency of Bayesian procedures has a long history: the Bernstein–von Mises theorem, though attributed to Bernšteĭn (1964) and von Mises (1964) in the middle of the twentieth century, was in fact anticipated by Laplace (1810), and the first rigorous proof was provided by Le Cam (1953, 1986). Counterexamples to the Bernstein–von Mises phenomenon in 'large' spaces began to appear in the 1960s, beginning with the work of Freedman (1963, 1965) and continuing with that of Diaconis and Freedman (1998), Leahu (2011), Owhadi et al.

(2015) and others. There are also positive results for infinite-dimensional settings, such as those of Castillo and Nickl (2013, 2014) and Szabó et al. (2014, 2015). It is now becoming clear that the crossover from consistency to inconsistency depends subtly upon the topology of convergence and the geometry of the proposed credible/confidence sets.

6.7 Exercises

Exercise 6.1. Let $\mu_1 = \mathcal{N}(m_1, C_1)$ and $\mu_2 = \mathcal{N}(m_2, C_2)$ be non-degenerate Gaussian measures on $\mathbb{R}^n$ with Lebesgue densities ρ_1 and ρ_2 respectively. Show that the probability measure with Lebesgue density proportional to $\rho_1 \rho_2$ is a Gaussian measure $\mu_3 = \mathcal{N}(m_3, C_3)$, where

$$\begin{aligned} C_3^{-1} &= C_1^{-1} + C_2^{-1}, \\ m_3 &= C_3(C_1^{-1} m_1 + C_2^{-1} m_2). \end{aligned}$$

Note well the property that the precision matrices *sum*, whereas the covariance matrices undergo a kind of harmonic average. (This result is sometimes known as *completing the square*.)

Exercise 6.2. Let μ_0 be a Gaussian probability measure on $\mathbb{R}^n$ and suppose that the potential $\Phi(u; y)$ is quadratic in u. Show that the posterior $\mathrm{d}\mu^y \propto e^{-\Phi(u;y)} \, \mathrm{d}\mu_0$ is also a Gaussian measure on $\mathbb{R}^n$. Using whatever characterization of Gaussian measures you feel most comfortable with, extend this result to a Gaussian probability measure μ_0 on a separable Banach space $\mathcal{U}$.

Exercise 6.3. Let $\Gamma \in \mathbb{R}^{q \times q}$ be symmetric and positive definite. Suppose that $H \colon \mathcal{U} \to \mathbb{R}^q$ satisfies

(a) For every $\varepsilon > 0$, there exists $M \in \mathbb{R}$ such that, for all $u \in \mathcal{U}$,

$$\|H(u)\|_{\Gamma^{-1}} \leq \exp\bigl(\varepsilon \|u\|_{\mathcal{U}}^2 + M\bigr).$$

(b) For every $r > 0$, there exists $K > 0$ such that, for all $u_1, u_2 \in \mathcal{U}$ with $\|u_1\|_{\mathcal{U}}, \|u_2\|_{\mathcal{U}} < r$,

$$\|H(u_1) - H(u_2)\|_{\Gamma^{-1}} \leq K \bigl\|u_1 - u_2\bigr\|_{\mathcal{U}}.$$

Show that $\Phi \colon \mathcal{U} \times \mathbb{R}^q \to \mathbb{R}$ defined by

$$\Phi(u; y) := \frac{1}{2} \bigl\langle y - H(u), \Gamma^{-1}(y - H(u)) \bigr\rangle$$

satisfies the standard assumptions.

Exercise 6.4. Prove Theorem 6.10. Hint: follow the model of Theorem 6.9, with (μ, μ^N) in place of $(\mu^y, \mu^{y'})$, and using (6.11) instead of (A4).