PART I

INVERSE PROBLEMS

1
Bayesian Inverse Problems and Well-Posedness

In this chapter we introduce the Bayesian approach to inverse problems in which the unknown parameter and the observed data are viewed as random variables. In this probabilistic formulation, the solution of the inverse problem is the posterior distribution on the parameter given the data. We will show that the Bayesian formulation leads to a form of well-posedness: small perturbations of the forward model or the observed data translate into small perturbations of the posterior distribution. Well-posedness requires a notion of distance between probability measures. We introduce the total variation and Hellinger distances, giving characterizations of them, and bounds relating them, that will be used throughout these notes. We prove well-posedness in the Hellinger distance.

The chapter is organized as follows. Section 1.1 introduces the formulation of Bayesian inverse problems. In Section 1.2 we derive a formula for the posterior pdf and explain how several estimators for the unknown parameter can be obtained using the posterior. Section 1.3 describes the well-posedness of the Bayesian formulation together with the necessary background on distances between probability measures. The chapter closes with bibliographical remarks in Section 1.4.

1.1 Formulation of Bayesian Inverse Problems

We consider the following setting. We let $G : \mathbb{R}^d \to \mathbb{R}^k$ define the forward model and aim to recover an unknown parameter $u \in \mathbb{R}^d$ from data $y \in \mathbb{R}^k$ given by

$$y = G(u) + \eta, \tag{1.1}$$

where $\eta \in \mathbb{R}^k$ represents observation noise. We view $(u, y) \in \mathbb{R}^d \times \mathbb{R}^k$ as a random variable, whose distribution is specified by means of the following

assumption on the distribution of $(u, \eta) \in \mathbb{R}^d \times \mathbb{R}^k$ and the relationship between u, y and η postulated in equation (1.1).

Assumption 1.1 *The distribution of the random variable $(u, \eta) \in \mathbb{R}^d \times \mathbb{R}^k$ is defined by:*

- $u \sim \rho(u), u \in \mathbb{R}^d$.
- $\eta \sim \nu(\eta), \eta \in \mathbb{R}^k$.
- *u and η are independent, written $u \perp \eta$.*

Here ρ and ν describe the pdfs of the random variables u and η, respectively. Then $\rho(u)$ is called the *prior* pdf and, for each fixed $u \in \mathbb{R}^d$, $y \mid u \sim \nu\big(y - G(u)\big)$ determines the *likelihood* function. In this probabilistic perspective, the solution to the inverse problem is the conditional distribution of u given y, which is called the *posterior* distribution, and will be denoted by $u \mid y \sim \pi^y(u)$. The posterior pdf determines, for any candidate parameter value in $\mathbb{R}^d$, how probable that parameter is, based on prior assumptions and the link between parameter and data, all expressed probabilistically. In particular, the posterior contains information about the level of uncertainty in the parameter recovery: for instance, large posterior covariance typically indicates that the data contains insufficient information to accurately recover the input parameter.

1.2 Formula for Posterior pdf: Bayes' Theorem

Bayes' theorem is a bridge connecting the prior, the likelihood, and the posterior.

Theorem 1.2 (Bayes' Theorem) *Let Assumption 1.1 hold, and assume that*

$$Z = Z(y) := \int_{\mathbb{R}^d} \nu\big(y - G(u)\big)\rho(u)du > 0.$$

Then $u \mid y \sim \pi^y(u)$, where

$$\pi^y(u) = \frac{1}{Z}\nu\big(y - G(u)\big)\rho(u). \tag{1.2}$$

Proof Denote by $\mathbb{P}(\cdot)$ the pdf of a random variable and by $\mathbb{P}(\cdot \mid \cdot)$ its conditional pdf. We have

$$\mathbb{P}(u, y) = \mathbb{P}(u \mid y)\,\mathbb{P}(y), \text{ if } \mathbb{P}(y) > 0,$$
$$\mathbb{P}(u, y) = \mathbb{P}(y \mid u)\,\mathbb{P}(u), \text{ if } \mathbb{P}(u) > 0.$$

Note that the marginal pdf on y is given by

$$\begin{aligned}\mathbb{P}(y) &= \int_{\mathbb{R}^d} \mathbb{P}(u, y)du \\ &= \int_{\mathbb{R}^d} \mathbb{P}(y \mid u)\, \mathbb{P}(u)du = Z > 0.\end{aligned}$$

Then

$$\mathbb{P}(u \mid y) = \frac{1}{\mathbb{P}(y)}\, \mathbb{P}(y \mid u)\, \mathbb{P}(u) = \frac{1}{\mathbb{P}(y)} \nu\big(y - G(u)\big)\rho(u) \tag{1.3}$$

for both $\mathbb{P}(u) = \rho(u) > 0$ and $\mathbb{P}(u) = \rho(u) = 0$. □

We will often denote the likelihood function by $\mathrm{l}(u) := \nu\big(y - G(u)\big)$. We then write

$$\pi^y(u) = \frac{1}{Z}\mathrm{l}(u)\rho(u),$$

omitting the data y in the likelihood function; when no confusion arises we will also simply write $\pi(u)$ for the posterior pdf, rather than $\pi^y(u)$.

***Remark* 1.3** The proof of Theorem 1.2 shows that in order to apply Bayes' formula (1.2) one needs to guarantee that the normalizing constant $\mathbb{P}(y) = Z$ is positive; in other words, the marginal density of the observed data y needs to be positive. This is simply the natural assumption that the observed data could indeed have been observed, given the probabilistic conditions in Assumption 1.1. From now on it will be assumed without further notice that $\mathbb{P}(y) = Z > 0$. Finally, we remark that throughout these notes we will denote normalizing constants generically by Z, and depending on the context the normalizing constant may sometimes be interpreted as the marginal density of an underlying data set. ⋄

The posterior distribution $\pi^y(u)$ contains all the knowledge on the parameter u available in the prior and the data. In applications it is often useful, however, to summarize the posterior distribution through a few numerical values. Summarizing the posterior is particularly important if the parameter is high-dimensional, since then visualizing the posterior or detecting regions of high posterior probability is nontrivial. Two natural numerical summaries are the posterior mean and the posterior mode.

Definition 1.4 The *posterior mean estimator* of u given data y is the mean of the posterior distribution:

$$u_{\mathrm{PM}} = \int_{\mathbb{R}^d} u\pi^y(u)\, du.$$

The *maximum a posteriori (MAP) estimator* of u given data y is the mode of the posterior distribution $\pi^y(u)$, defined as

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

This maximum may not be uniquely defined, in which case we talk about *a*, rather than *the*, MAP estimator. ◊

The importance of the MAP and the posterior mean already suggest the need to compute maxima (for the MAP estimator) and integrals (for the posterior mean) in order to extract actionable information from the Bayesian formulation of inverse problems and data assimilation. For this reason, optimization (to compute maxima) and sampling (to compute integrals) will play an important role in these notes. In practice it is often useful to quantify the uncertainty in the parameter reconstruction, and numerical summaries such as the posterior mean and the MAP estimators can be complemented by credible intervals; that is, parameter regions of prescribed posterior probability. In order to make tractable the computation of estimators and credible intervals, the posterior can be approximated by a simple distribution, such as a Gaussian or a Gaussian mixture; optimization can be used to determine such approximations. In a similar spirit, sampling may be viewed as approximating the posterior by a combination of Dirac masses to enable computation of integrals. An optimization perspective for inverse problems and data assimilation will be studied in Chapters 3 and 9, respectively, and Gaussian approximations will be discussed in Chapters 4 and 10, respectively; Dirac approximations constructed via sampling will be studied in Chapters 5 and 6 (inverse problems) and in Chapters 11 and 12 (data assimilation).

We next consider two simple examples of a direct application of Bayes' theorem.

Example 1.5 (MAP and Posterior Mean Estimators) Let $d = k = 1$, $\eta \sim \nu = \mathcal{N}(0, \gamma^2)$, and let

$$\rho(u) = \begin{cases} \frac{1}{2}, & u \in (-1,1), \\ 0, & u \in (-1,1)^c. \end{cases}$$

Suppose that the observation is generated by $y = u + \eta$. Using Bayes' Theorem 1.2, we derive the posterior pdf

$$\pi^y(u) = \begin{cases} \frac{1}{2Z}\exp(-\frac{1}{2\gamma^2}|y-u|^2), & u \in (-1,1), \\ 0, & u \in (-1,1)^c, \end{cases}$$

where Z is a normalizing constant ensuring that $\int_{\mathbb{R}} \pi^y(u)du = 1$. Now we find

the MAP estimator. From the explicit formula for π^y, we have

$$u_{\text{MAP}} = \arg\max_{u\in\mathbb{R}} \pi^y(u) = \begin{cases} y & \text{if } y \in (-1,1), \\ -1 & \text{if } y \le -1, \\ 1 & \text{if } y \ge 1. \end{cases}$$

In this example, the prior on u is supported on $(-1,1)$ and the posterior on $u \mid y$ is supported on $(-1,1)$. If the data lies in $(-1,1)$ then the MAP estimator is the data itself; otherwise it is the extremal point of the prior support which matches the sign of the data. The posterior mean is

$$u_{\text{PM}} = \frac{1}{2Z}\int_{-1}^{1} u \exp\Big(-\frac{1}{2\gamma^2}|y-u|^2\Big)\, du,$$

which may be approximated using, for instance, the sampling methods described in Chapters 5 and 6. ◇

The following example illustrates once again the application of Bayes' theorem, and shows that the posterior may concentrate near a low-dimensional manifold in the input parameter space $\mathbb{R}^d$. In such a case it is important to understand the geometry of the support of the posterior density, which cannot be captured by point estimation or Gaussian approximations.

Example 1.6 (Concentration of Posterior on a Manifold) Let $d = 2$, $k = 1$, $\rho \in C(\mathbb{R}^2, \mathbb{R})$, and suppose that there is $\rho_{\max} > 0$ such that, for all $u \in \mathbb{R}^2$, we have $0 < \rho(u) \le \rho_{\max} < \infty$. Suppose that the observation is generated by

$$\begin{aligned} y &= G(u) + \eta, \\ G(u) &= u_1^2 + u_2^2, \\ \eta &\sim \nu = \mathcal{N}(0,\gamma^2), \quad 0 < \gamma \ll 1, \end{aligned}$$

and assume that $y > 0$. Using Bayes' theorem we obtain the posterior pdf

$$\pi^y(u) = \frac{1}{Z}\exp\Big(-\frac{1}{2\gamma^2}|u_1^2 + u_2^2 - y|^2\Big)\rho(u).$$

We now show that the posterior concentrates near the manifold defined by the circumference $\{u \in \mathbb{R}^2 : u_1^2 + u_2^2 = y\}$. Denote $A^{\pm} := \{u \in \mathbb{R}^2 : |u_1^2 + u_2^2 - y|^2 \le \gamma^{2\pm\delta}\}$, for some fixed $\delta \in (0,2)$. The set A^- is defined so that it captures most of the posterior probability, and A^+ so that it captures little of the posterior probability. They are defined this way because the observational noise has variance γ^2; considering a neighborhood of the circumference which scales as γ raised to a power slightly smaller than 2 captures most of the posterior probability; considering a neighborhood of the circumference in which the exponent is slightly

larger than this captures little of the posterior probability. Define B to be the closed ball of radius $2\sqrt{y}$ centered at the origin. Let $u^+ \in A^+ \subset B, u^- \in (A^-)^c$ and let $\rho_{\min} = \inf_{u\in B} \rho(u)$. Since $\rho(u)$ is positive and continuous and B is compact, $\rho_{\min} > 0$. Taking the small noise limit yields

$$\frac{\pi^y(u^+)}{\pi^y(u^-)} \geq \exp\left(-\frac{1}{2}\gamma^\delta + \frac{1}{2}\gamma^{-\delta}\right)\frac{\rho_{\min}}{\rho_{\max}} \to \infty, \text{ as } \gamma \to 0^+.$$

Therefore, noting that $y > 0$, the posterior π^y concentrates, as $\gamma \to 0^+$, on the circumference with radius $\sqrt{y}$. ◊

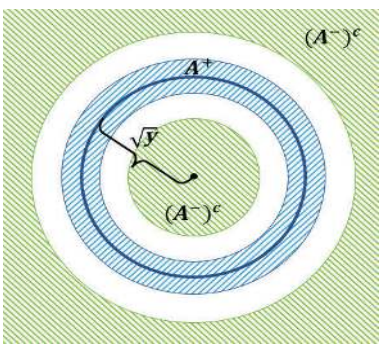

Figure 1.1 The posterior measure concentrates on a circumference with radius $\sqrt{y}$. Here, the blue shadow area is A^+ and the green shadow area is $(A^-)^c$.

1.3 Well-Posedness of Bayesian Inverse Problems

In this section we show that the Bayesian formulation of inverse problems leads to a form of well-posedness. More precisely, we study the sensitivity of the posterior pdf to perturbations of the forward model G. In many inverse problems the ideal forward model G is not accessible but can be approximated by some computable G_δ; consequently π^y is replaced by π^y_δ. An example that is often found in applications, to which the theory contained herein may be generalized, is when G is an operator acting on an infinite-dimensional space which is approximated, for the purposes of computation, by some finite-dimensional operator G_δ. We seek to prove that, under certain assumptions, the small difference between G and G_δ (forward error) leads to a similarly small difference between π^y and π^y_δ (inverse error):

Meta Theorem (Well-Posedness)

$$|G - G_\delta| = O(\delta) \implies d(\pi^y, \pi^y_\delta) = O(\delta)$$

for small enough $\delta > 0$ and some metric $d(\cdot,\cdot)$ on probability densities.

This result will be formalized in Theorem 1.15 below, which shows that the $O(\delta)$-convergence of π^y_δ with respect to some distance $d(\cdot,\cdot)$ can be guaranteed under certain assumptions on the likelihood. We will conclude the chapter by showing an example where these assumptions hold true. In order to discuss these issues we will need to introduce metrics on probability densities.

1.3.1 Metrics on Probability Densities

Here we introduce the total variation and the Hellinger distance, both of which have been used to show well-posedness results. In this chapter we will use the Hellinger distance to establish well-posedness of Bayesian inverse problems, and in Chapter 7 we employ the total variation distance to establish well-posedness of Bayesian formulations of filtering and smoothing in data assimilation.

Definition 1.7 The *total variation distance* between two pdfs π and π' is defined by

$$d_{\mathrm{TV}}(\pi,\pi') := \frac{1}{2}\int |\pi(u) - \pi'(u)|du = \frac{1}{2}\|\pi - \pi'\|_{L^1}.$$

The *Hellinger distance* between two pdfs π and π' is defined by

$$d_{\mathrm{H}}(\pi,\pi') := \left(\frac{1}{2}\int |\sqrt{\pi(u)} - \sqrt{\pi'(u)}|^2 du\right)^{1/2} = \frac{1}{\sqrt{2}}\|\sqrt{\pi} - \sqrt{\pi'}\|_{L^2}.$$

$\diamond$

In the rest of this subsection we will establish bounds between the Hellinger and total variation distance, and show how both distances can be used to bound the difference of expected values computed with two different densities; these results will be used in subsequent chapters. Before doing so, the next lemma motivates our choice of normalization constant $1/2$ for total variation distance and $1/\sqrt{2}$ for Hellinger distance: they are chosen so that the maximum possible distance between two densities is one. The proof also shows that π and π' have total variation and Hellinger distance equal to one if and only if they have disjoint supports; that is, if $\int \pi(u)\pi'(u)du = 0$.

Lemma 1.8 *For any pdfs π and π',*

$$0 \le d_{\mathrm{TV}}(\pi,\pi') \le 1, \quad 0 \le d_{\mathrm{H}}(\pi,\pi') \le 1.$$

Proof The lower bounds follow immediately from the definitions, so we only need to prove the upper bounds. For total variation distance,

$$d_{\mathrm{TV}}(\pi,\pi') = \frac{1}{2}\int |\pi(u)-\pi'(u)|du \le \frac{1}{2}\int \pi(u)du + \frac{1}{2}\int \pi'(u)du = 1,$$

and for Hellinger distance,

$$\begin{aligned} d_{\mathrm{H}}(\pi,\pi') &= \left(\frac{1}{2}\int \left|\sqrt{\pi(u)}-\sqrt{\pi'(u)}\right|^2 du\right)^{1/2} \\ &= \left(\frac{1}{2}\int \left(\pi(u)+\pi'(u)-2\sqrt{\pi(u)\pi'(u)}\ \right)du\right)^{1/2} \\ &\le \left(\frac{1}{2}\int \left(\pi(u)+\pi'(u)\right)du\right)^{1/2} \\ &= 1. \end{aligned}$$

□

The following result gives bounds between total variation and Hellinger distance.

Lemma 1.9 *For any pdfs π and π',*

$$\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\pi,\pi') \le d_{\mathrm{H}}(\pi,\pi') \le \sqrt{d_{\mathrm{TV}}(\pi,\pi')}.$$

Proof From the Cauchy–Schwarz inequality it follows that

$$\begin{aligned} d_{\mathrm{TV}}(\pi,\pi') &= \frac{1}{2}\int \left|\sqrt{\pi(u)}-\sqrt{\pi'(u)}\right|\left|\sqrt{\pi(u)}+\sqrt{\pi'(u)}\right|du \\ &\le \left(\frac{1}{2}\int \left|\sqrt{\pi(u)}-\sqrt{\pi'(u)}\right|^2 du\right)^{1/2}\left(\frac{1}{2}\int \left|\sqrt{\pi(u)}+\sqrt{\pi'(u)}\right|^2 du\right)^{1/2} \\ &\le d_{\mathrm{H}}(\pi,\pi')\left(\frac{1}{2}\int \left(2\pi(u)+2\pi'(u)\right)du\right)^{1/2} \\ &= \sqrt{2}\, d_{\mathrm{H}}(\pi,\pi'). \end{aligned}$$

Notice that $|\sqrt{\pi(u)}-\sqrt{\pi'(u)}| \le |\sqrt{\pi(u)}+\sqrt{\pi'(u)}|$ since $\sqrt{\pi(u)}, \sqrt{\pi'(u)} \ge 0$.