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 $G\colon \mathbb{R}^d \to \mathbb{R}^k$ 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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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

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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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 $\mathrm{l}(u) \coloneqq \nu\big(y - G(u)\big)$. 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:

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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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

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

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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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 $\sqrt{y}$. ◊

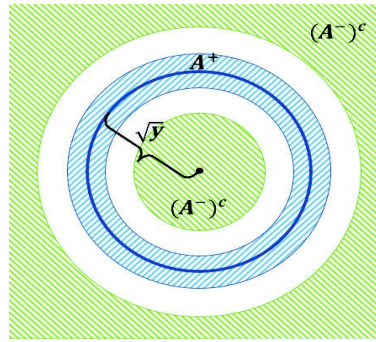

Figure 1.1 The posterior measure concentrates on a circumference with radius $\sqrt{y}$. Here, the blue shadow area is A^+ and the green shadow area is $(A^-)^c$.

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):

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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

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

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

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

$\diamond$

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',*

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

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

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 $\sqrt{\pi(u)}, \sqrt{\pi'(u)} \ge 0$.

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

Thus we have

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

□

The following two lemmas show that if two densities are close in total variation or in Hellinger distance, expectations computed with respect to both densities are also close. In addition, the following lemma also provides a useful characterization of the total variation distance that will be used repeatedly throughout these notes.

Lemma 1.10 *Let f be a function such that $|f|_\infty := \sup_{u\in\mathbb{R}^d} |f(u)| < \infty$. It holds that*

$$\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| \le 2|f|_\infty d_{\mathrm{TV}}(\pi,\pi').$$

Moreover, the following variational characterization of the total variation distance holds:

$$d_{\mathrm{TV}}(\pi,\pi') = \frac{1}{2}\sup_{|f|_\infty\le 1}\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right|. \tag{1.4}$$

Proof For the first part of the lemma, note that

$$\begin{aligned} \left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| &= \left|\int_{\mathbb{R}^d} f(u)\big(\pi(u)-\pi'(u)\big)du\right| \\ &\le 2|f|_\infty \cdot \frac{1}{2}\int_{\mathbb{R}^d} |\pi(u)-\pi'(u)|du \\ &= 2|f|_\infty d_{\mathrm{TV}}(\pi,\pi'). \end{aligned}$$

This in particular shows that, for any f with $|f|_\infty = 1$,

$$d_{\mathrm{TV}}(\pi,\pi') \ge \frac{1}{2}\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right|.$$

Our goal now is to show a choice of f with $|f|_\infty = 1$ that achieves equality. Define $f(u) := \operatorname{sign}\Big(\pi(u)-\pi'(u)\Big)$, so that $f(u)\Big(\pi(u)-\pi'(u)\Big) = |\pi(u)-\pi'(u)|$. Then

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

it holds that $|f|_\infty = 1$, and

$$\begin{aligned} d_{\mathrm{TV}}(\pi, \pi') &= \frac{1}{2} \int_{\mathbb{R}^d} |\pi(u) - \pi'(u)| \, du \\ &= \frac{1}{2} \int_{\mathbb{R}^d} f(u) \Big(\pi(u) - \pi'(u) \Big) \, du \\ &= \frac{1}{2} \Big| \mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f] \Big|. \end{aligned}$$

This completes the proof of the variational characterization. □

Lemma 1.11 *Let f be a function such that $f_2 := \left(\mathbb{E}^{\pi}[|f|^2] + \mathbb{E}^{\pi'}[|f|^2]\right)^{\frac{1}{2}} < \infty$. Then*

$$\Big| \mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f] \Big| \leq 2 f_2 d_{\mathrm{H}}(\pi, \pi').$$

Proof Using the Cauchy–Schwarz inequality gives

$$\begin{aligned} \Big| \mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f] \Big| &= \left| \int_{\mathbb{R}^d} f(u) \Big(\sqrt{\pi(u)} - \sqrt{\pi'(u)} \Big) \Big(\sqrt{\pi(u)} + \sqrt{\pi'(u)} \Big) \, du \right| \\ &\leq \left(\frac{1}{2} \int \Big| \sqrt{\pi(u)} - \sqrt{\pi'(u)} \Big|^2 du \right)^{1/2} \\ &\quad \times \left(2 \int |f(u)|^2 \Big| \sqrt{\pi(u)} + \sqrt{\pi'(u)} \Big|^2 \, du \right)^{1/2} \\ &\leq d_{\mathrm{H}}(\pi, \pi') \left(4 \int |f(u)|^2 \big(\pi(u) + \pi'(u) \big) du \right)^{1/2} \\ &= 2 f_2 \, d_{\mathrm{H}}(\pi, \pi'). \end{aligned}$$

□

***Remark* 1.12** Note that the result for Hellinger only assumes that f is square integrable with respect to π and π'. In contrast, the result for total variation distance assumes that f is bounded, which is a stronger condition. Lemma 1.9 also demonstrates that smallness in the Hellinger metric is a more stringent condition than smallness in total variation. Our aim in the following section is to show well-posedness in some metric on probability densities. The preceding observations suggest that establishing such a result in the Hellinger metric makes a stronger statement than doing so in total variation. ◊

1.3.2 Approximation Theorem

We denote by

$$\mathsf{l}(u) = \nu\big(y - G(u)\big) \quad \text{and} \quad \mathsf{l}_\delta(u) = \nu\big(y - G_\delta(u)\big)$$

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

the likelihoods associated with $G(u)$ and $G_\delta(u)$, so that

$$\pi^y(u) = \frac{1}{Z}\mathsf{l}(u)\rho(u) \quad \text{and} \quad \pi^y_\delta(u) = \frac{1}{Z_\delta}\mathsf{l}_\delta(u)\rho(u),$$

where $Z, Z_\delta > 0$ are the corresponding normalizing constants. Before we proceed to our main result, we make some assumptions.

Assumption 1.13 *There exist $\delta^+ > 0$ and $K_1, K_2 < \infty$ such that, for all $\delta \in (0, \delta^+)$,*

(i) $|\sqrt{\mathsf{l}(u)} - \sqrt{\mathsf{l}_\delta(u)}| \le \varphi(u)\delta$ *for some $\varphi(u)$ such that $\mathbb{E}^\rho[\varphi^2(u)] \le K_1^2$;*
(ii) $\sup_{u\in\mathbb{R}^d}(|\sqrt{\mathsf{l}(u)}| + |\sqrt{\mathsf{l}_\delta(u)}|) \le K_2$.

***Remark* 1.14** Assumption 1.13 only involves conditions on the likelihood l and the approximate likelihood l_δ. Our presentation in this chapter emphasizes the situation in which this approximation is necessitated in order to approximate the forward model G. However, another important scenario which is covered by the theory is approximation due to perturbations of the data y. As an example, we will establish in Chapter 7 a well-posedness result that guarantees stability of Bayesian smoothing under perturbations of the data. More generally, the theoretical framework introduced here is very flexible, and it may be employed to study the stability of many Bayesian formulations of inverse problems and data assimilation under a wide range of perturbations. ◊

Now we state the main result of this section:

Theorem 1.15 (Well-Posedness of Posterior) *Under Assumption 1.13 we have*

$$d_{\mathrm{H}}(\pi^y, \pi^y_\delta) \le c\delta, \quad \delta \in (0, \Delta),$$

for some $\Delta > 0$ and some $c \in (0, +\infty)$ independent of δ.

Notice that this theorem, together with Lemma 1.11, guarantees that expectations computed with respect to π^y and π^y_δ are order δ apart. To prove Theorem 1.15, we first show a lemma which characterizes the normalization factor Z_δ in the small δ limit.

Lemma 1.16 *Under Assumption 1.13 there exist $\Delta > 0$, $c_1, c_2 \in (0, +\infty)$ such that*

$$|Z - Z_\delta| \le c_1\delta \quad \text{and} \quad Z, Z_\delta > c_2, \quad \text{for } \delta \in (0, \Delta).$$

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

Proof Since $Z = \int \mathsf{l}(u)\rho(u)du$ and $Z_\delta = \int \mathsf{l}_\delta(u)\rho(u)du$, we have

$$\begin{aligned}
|Z - Z_\delta| &= \left| \int \big(\mathsf{l}(u) - \mathsf{l}_\delta(u)\big)\rho(u)du \right| \\
&\le \Big(\int \Big| \sqrt{\mathsf{l}(u)} - \sqrt{\mathsf{l}_\delta(u)} \Big|^2 \rho(u)du \Big)^{1/2} \Big(\int \Big| \sqrt{\mathsf{l}(u)} + \sqrt{\mathsf{l}_\delta(u)} \Big|^2 \rho(u)du \Big)^{1/2} \\
&\le \Big(\int \delta^2 \varphi(u)^2 \rho(u)du \Big)^{1/2} \Big(\int K_2^2 \rho(u)du \Big)^{1/2} \\
&\le K_1 K_2 \delta, \quad \delta \in (0, \delta^+).
\end{aligned}$$

Therefore, for $\delta \le \Delta := \min\{\frac{Z}{2K_1K_2}, \delta^+\}$, we have

$$Z_\delta \ge Z - |Z - Z_\delta| \ge \frac{1}{2} Z.$$

The lemma follows by taking $c_1 = K_1K_2$ and $c_2 = \frac{1}{2}Z$. □

Proof of Theorem 1.15 We break the total error into two contributions, one reflecting the difference between Z and Z_δ, and the other the difference between l and l_δ:

$$\begin{aligned}
d_{\mathrm{H}}(\pi^y, \pi^y_\delta) &= \frac{1}{\sqrt{2}} \Big\| \sqrt{\pi^y} - \sqrt{\pi^y_\delta} \Big\|_{L^2} \\
&= \frac{1}{\sqrt{2}} \Big\| \sqrt{\frac{\mathsf{l}\rho}{Z}} - \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} + \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} - \sqrt{\frac{\mathsf{l}_\delta\rho}{Z_\delta}} \Big\|_{L^2} \\
&\le \frac{1}{\sqrt{2}} \Big\| \sqrt{\frac{\mathsf{l}\rho}{Z}} - \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} \Big\|_{L^2} + \frac{1}{\sqrt{2}} \Big\| \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} - \sqrt{\frac{\mathsf{l}_\delta\rho}{Z_\delta}} \Big\|_{L^2}.
\end{aligned}$$

Using Lemma 1.16 we have, for $\delta \in (0, \Delta)$,

$$\begin{aligned}
\Big\| \sqrt{\frac{\mathsf{l}\rho}{Z}} - \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} \Big\|_{L^2} &= \Big| \frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z_\delta}} \Big| \Big(\int \mathsf{l}(u)\rho(u)du \Big)^{1/2} \\
&= \frac{|Z - Z_\delta|}{(\sqrt{Z} + \sqrt{Z_\delta})\sqrt{Z_\delta}} \\
&\le \frac{c_1}{2c_2} \delta,
\end{aligned}$$

and

$$\Big\| \sqrt{\frac{\mathsf{l}\rho}{Z_\delta}} - \sqrt{\frac{\mathsf{l}_\delta\rho}{Z_\delta}} \Big\|_{L^2} = \frac{1}{\sqrt{Z_\delta}} \Big(\int \Big| \sqrt{\mathsf{l}(u)} - \sqrt{\mathsf{l}_\delta(u)} \Big|^2 \rho(u)du \Big)^{1/2} \le \sqrt{\frac{K_1^2}{c_2}} \delta.$$

Therefore

$$d_{\mathrm{H}}(\pi^y, \pi^y_\delta) \le \frac{1}{\sqrt{2}} \frac{c_1}{2c_2} \delta + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1^2}{c_2}} \delta = c\delta,$$

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

with $c = \frac{1}{\sqrt{2}}\frac{c_1}{2c_2} + \frac{K_1}{\sqrt{2c_2}}$, which is independent of δ. □

1.3.3 Example: Well-Posedness for Parameter Estimation in an ODE

Many inverse problems arise from differential equations with unknown input parameters. Here we consider a simple but typical example where $G(u)$ comes from the solution of an ordinary differential equation (ODE), which needs to be solved numerically. Let $x(t)$ be the solution to the initial value problem

$$\frac{dx}{dt} = F(x;u), \quad x(0) = 0, \tag{1.5}$$

where $F\colon \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^k$ is a function such that $F(x;u)$ and the partial Jacobian $D_x F(x;u)$ are uniformly bounded with respect to (x,u), i.e.

$$|F(x;u)|, |D_x F(x;u)| < F_{\max}, \quad \text{for all } (x,u) \in \mathbb{R}^k \times \mathbb{R}^d,$$

for some constant $F_{\max}$, and thus $F(x;u)$ is Lipschitz in x in that, for all $u \in \mathbb{R}^d$,

$$|F(x;u) - F(x';u)| \le F_{\max}|x - x'|, \quad \text{for all } x, x' \in \mathbb{R}^k.$$

Note that $u \in \mathbb{R}^d$ defines parametric dependence of the vector field defining the differential equation.

Now consider the inverse problem setting

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

where

$$G(u) \coloneqq x(1) = x(t)|_{t=1},$$

and $\eta \sim \mathcal{N}(0, \gamma^2 I_k)$. We assume that the exact mapping $G(u)$ is replaced by some numerical approximation $G_\delta(u)$. In particular, $G_\delta(u)$ is given by using the forward Euler method to solve the ODE (1.5). Define $X_0 = 0$ and

$$X_{\ell+1} = X_\ell + \delta F(X_\ell; u), \quad \ell \ge 0,$$

where $\delta = \frac{1}{L}$ for some large integer L. Finally define $G_\delta(u) \coloneqq X_L$.

In what follows, we will prove that $G_\delta(u)$ is uniformly bounded and close to $G(u)$ when δ is small, and that G and G_δ both satisfy the same global bound. Then we will use these results to show that Assumption 1.13 is satisfied. Therefore, we can apply Theorem 1.15 to this example to establish that the approximate posterior π^y_δ, defined by approximate forward model G_δ, is close to the true posterior π^y with exact forward model G.

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

In showing that Assumption 1.13 is satisfied, we use Lemmas 1.17 and 1.18 below. Recall that $\eta \sim \mathcal{N}(0, \gamma^2 I_k)$, and thus

$$\sqrt{\mathrm{I}(u)} = \sqrt{\nu\big(y - G(u)\big)} = \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \exp\Big(-\frac{1}{4\gamma^2}|y - G(u)|^2\Big),$$

$$\sqrt{\mathrm{I}_\delta(u)} = \sqrt{\nu\big(y - G_\delta(u)\big)} = \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \exp\Big(-\frac{1}{4\gamma^2}|y - G_\delta(u)|^2\Big).$$

- For Assumption 1.13(i) notice that the function e^{-w} is Lipschitz for $w > 0$, with Lipschitz constant 1. Therefore we have

$$\begin{aligned}\Big|\sqrt{\mathrm{I}(u)} - \sqrt{\mathrm{I}_\delta(u)}\Big| &\le \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \cdot \frac{1}{4\gamma^2} \cdot \big||y - G(u)|^2 - |y - G_\delta(u)|^2\big| \\ &= \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \cdot \frac{1}{4\gamma^2} \cdot |2y - G(u) - G_\delta(u)||G(u) - G_\delta(u)| \\ &\le \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \cdot \frac{1}{4\gamma^2} \cdot (2|y| + 2F_{\max})c\delta \\ &= \tilde{c}\delta.\end{aligned}$$

  That is, Assumption 1.13(i) is satisfied with $\varphi(u) = \tilde{c}$ and $\int_{\mathbb{R}^d} \varphi^2(u)\rho(u)\,du = \tilde{c}^2 < \infty$.
- Assumption 1.13(ii) is satisfied, since

$$\sqrt{\mathrm{I}(u)} = \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \exp\Big(-\frac{1}{4\gamma^2}|y - G(u)|^2\Big) \le \frac{1}{(2\pi)^{k/4}\gamma^{k/2}},$$

$$\sqrt{\mathrm{I}_\delta(u)} = \frac{1}{(2\pi)^{k/4}\gamma^{k/2}} \exp\Big(-\frac{1}{4\gamma^2}|y - G_\delta(u)|^2\Big) \le \frac{1}{(2\pi)^{k/4}\gamma^{k/2}}.$$

The preceding verification of Assumption 1.13 used the following two lemmas, and the first of these uses the Gronwall inequality which follows them. Define $t_\ell = \ell\delta, x_\ell = x(t_\ell)$. The following lemma gives an estimate on the error generated from using the forward Euler method.

Lemma 1.17 *Let $E_\ell := x_\ell - X_\ell$. Then there is $c < \infty$ independent of δ such that*

$$|E_\ell| \le c\delta, \quad 0 \le \ell \le L.$$

In particular,

$$|G(u) - G_\delta(u)| = |E_L| \le c\delta.$$

Proof For simplicity of exposition, we consider the case $k = 1$; the case $k > 1$ is almost identical, simply requiring the integral form for the remainder term

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

in the Taylor expansion. Using Taylor expansion in the case $k = 1$, there is $\xi_\ell \in [t_\ell, t_{\ell+1}]$ such that

$$\begin{aligned} x_{\ell+1} &= x_\ell + \delta \frac{dx}{dt}(t_\ell) + \frac{\delta^2}{2}\frac{d^2x}{dt^2}(\xi_\ell) \\ &= x_\ell + \delta F(x_\ell; u) + \frac{\delta^2}{2} D_x F\big(x(\xi_\ell); u\big) F\big(x(\xi_\ell); u\big). \end{aligned}$$

Thus we have

$$\begin{aligned} |E_{\ell+1}| &= |x_{\ell+1} - X_{\ell+1}| \\ &= \left| x_\ell - X_\ell + \delta\Big(F(x_\ell; u) - F(X_\ell; u)\Big) + \frac{\delta^2}{2} D_x F\big(x(\xi_\ell); u\big) F\big(x(\xi_\ell); u\big) \right| \\ &\le |x_\ell - X_\ell| + \delta \big|F(x_\ell; u) - F(X_\ell; u)\big| + \frac{\delta^2}{2} \big|D_x F\big(x(\xi_\ell); u\big)\big| \big|F\big(x(\xi_\ell); u\big)\big| \\ &\le |E_\ell| + \delta F_{\max} |E_\ell| + \frac{\delta^2}{2} F_{\max}^2. \end{aligned}$$

Noticing that $|E_0| = 0$, the discrete Gronwall inequality (Theorem 1.19) gives

$$\begin{aligned} |E_\ell| &\le (1 + \delta F_{\max})^\ell |E_0| + \frac{(1 + \delta F_{\max})^\ell - 1}{\delta F_{\max}} \cdot \frac{\delta^2}{2} F_{\max}^2 \\ &\le \left(\left(1 + \frac{F_{\max}}{L}\right)^L - 1 \right) \cdot \frac{F_{\max}\delta}{2} \\ &\le \frac{(e^{F_{\max}} - 1) F_{\max}}{2} \delta. \end{aligned}$$

The lemma follows by taking $c = \frac{(e^{F_{\max}} - 1) F_{\max}}{2}$. □

Lemma 1.18 *For any $u \in \mathbb{R}^d$,*

$$|G(u)|, |G_\delta(u)| \le F_{\max}.$$

Proof For $G(u)$ we use that $F(x; u)$ is uniformly bounded, so that

$$|G(u)| = |x(1)| = \left| \int_0^1 F(x(t); u) dt \right| \le \int_0^1 \big|F(x(t); u)\big| dt \le F_{\max}.$$

As for $G_\delta(u)$, we first notice that

$$|X_{\ell+1}| = |X_\ell + \delta F(X_\ell; u)| \le |X_\ell| + \delta |F(X_\ell; u)| \le |X_\ell| + \delta F_{\max},$$

and by induction

$$|X_\ell| \le |X_0| + \ell \delta F_{\max} = \ell \delta F_{\max}.$$

In particular,

$$|G_\delta(u)| = |X_L| \le L \delta F_{\max} = F_{\max}.$$

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

□

The following discrete Gronwall inequality is used several times in these notes, and is stated and proved here for completeness.

Theorem 1.19 (Discrete Gronwall Inequality) *Let a positive sequence* $\{Z_\ell\}_{\ell=0}^{L}$ *satisfy*

$$Z_{\ell+1} \leq CZ_\ell + D, \qquad \text{for all } \ell = 0, \ldots, L-1$$

for some constants C, D *with* $C > 0$. *Then*

$$Z_\ell \leq \frac{D}{1-C}(1-C^\ell) + Z_0 C^\ell \qquad \text{for all } \ell = 0, \ldots, L, \quad C \neq 1$$

and

$$Z_\ell \leq \ell D + Z_0 \qquad \text{for all } \ell = 0, \ldots, L, \quad C = 1.$$

Proof The proof is by induction. We start with the case $C \neq 1$. The result holds for $\ell = 0$. Assume it is true for $\ell < L$. Then, using the defining inequality,

$$Z_{\ell+1} \leq \frac{CD}{1-C}(1-C^\ell) + Z_0 C^{\ell+1} + D.$$

Rearranging yields

$$Z_{\ell+1} \leq \frac{D}{1-C}(1-C^{\ell+1}) + Z_0 C^{\ell+1}$$

and the result follows by induction.

When $C = 0$ we again note that the result holds for $\ell = 0$. Assume it is true for $\ell < L$. Then, using the defining inequality with $C = 1$,

$$Z_{\ell+1} \leq \ell D + Z_0 + D = (\ell+1)D + Z_0$$

and the result follows by induction. □

1.4 Discussion and Bibliography

Kaipio and Somersalo (2006) provides an introduction to the Bayesian approach to inverse problems, especially in the context of differential equations, and Calvetti and Somersalo (2007) gives an introduction to Bayesian scientific computing. An overview of the subject of Bayesian inverse problems in differential equations, with a perspective informed by the geophysical sciences, is given in Tarantola (2015a) (see, especially, Chapter 5). For non-statistical approaches to inverse problems, we refer to the books Tikhonov and Arsenin (1977), Engl et al.

https://doi.org/10.1017/9781009414319.004 Published online by Cambridge University Press

(1996), Vogel (2002), and the lecture notes Bal (2012) and Miller and Karl (2003).

The subject of Bayesian inverse problems may be developed beyond the specific setting of equation (1.1) to study problems of the form

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

Our emphasis on additive noise η, often assumed to be Gaussian, simplifies some algorithms and enables us to be explicit about some formulae, but is not fundamental in any way. We refer to Dunlop (2019) for well-posedness theory and a study of MAP estimation with multiplicative noise. In addition, the setting of equation (1.1) presupposes that the forward model G is given to us, but in some cases the forward model itself may need to be learned from data.

In Stuart (2010) the Bayesian approach to regularization is reviewed, developing a function space viewpoint on the subject; a similar development of this approach is described in Lasanen (2012a,b). A well-posedness theory and some algorithmic approaches which are used when adopting the Bayesian approach to inverse problems are introduced. The function space viewpoint on the subject is developed in more detail in the lecture notes of Dashti and Stuart (2017). An early application of this function space methodology to a large-scale applied inverse problem, taken from the geophysical sciences, may be found in Martin et al. (2012). Lieberman et al. (2010) demonstrates the potential for the use of dimension reduction techniques from control theory within statistical inverse problems.

We refer to Gibbs and Su (2002) for further study on the subject of metrics, and other distance-like functions, on probability measures. The first published paper to discuss stability and well-posedness of the Bayesian inverse problem was Marzouk and Xiu (2009), in which the Kullback–Leibler divergence (see Chapter 4) is employed. Related results on stability and well-posedness, but using other distances and divergences, may be found in Latz (2020). The articles Stuart (2010) and Dashti and Stuart (2017) study well-posedness of Bayesian inverse problems in the Hellinger metric, with respect to perturbations in the data; Cotter et al. (2010) and Harlim et al. (2020) consider stability of the posterior distribution with respect to numerical approximation of partial differential equations appearing in the forward model. Hosseini and Nigam (2017); Hosseini (2017) discuss generalizations of the well-posedness theory to various classes of specific non-Gaussian priors. On the other hand, Iglesias et al. (2014b) contains an interesting set of examples where the Meta Theorem stated in this chapter fails in the sense that, whilst well-posedness holds, the posterior is Hölder with exponent less than 1, rather than Lipschitz, with respect to perturbations.

The Bayesian approach to inverse problems builds on, and benefits from, the vast literature on Bayesian statistics. Fienberg (2006) provides a historical overview of the development and popularization of Bayesian statistics, starting with the introduction of Bayes' formula (Bayes, 1763) and emphasizing the leading role of Savage (1972) in axiomatizing and popularizing the subjective view of probability pioneered by De Finetti (2017). We refer to Gelman et al. (2013) for a recent and comprehensive textbook on Bayesian methodology. See Nickl (2022) for an overview of Bayesian inversion and, in particular, statistical consistency results in this context.

A topic of debate in Bayesian statistics, and specifically in the Bayesian approach to inverse problems, is how to construct prior probability measures from available prior information, which is typically not described probabilistically. Owhadi et al. (2015a,b) demonstrate that this is an important question: different priors, both consistent with available prior information, can lead to wildly different Bayesian inference when computing posterior expectations: what the authors term *Bayesian brittleness.* Arguably, this issue may be dealt with through application of the scientific method: a given prior and likelihood are postulated, and posterior predictions are made; data acquired after making posterior predictions may then be used to evaluate the Bayesian probabilistic model employed, and in particular the prior and likelihood and, if necessary, modify it.

The body of work on Bayesian brittleness builds on related analysis in the context of forward uncertainty quantification (Owhadi et al., 2013), a topic concerned with propagating uncertainty on parameters through a model into predictions. The subject of uncertainty quantification, both the forward and inverse varieties, is overviewed in Sullivan (2015) and Smith (2013).