Introduction to Bayesian inverse problems

Let us consider a model

$$y = G(u),$$

where $u \in X_1$ and $G : X_1 \to X_2$. The inverse problem consists of finding u given y. Problems that may occur -

1) There maybe many solutions u corresponding to a single observation y.
2) The solution maybe very sensitive to the observation. Small error in observing y may cause large change in estimated value of u.
3) The error in observation may throw the observation y out of the range of G.

- We restrict ourselves to $X_1 = \mathcal{H}_1$, $X_2 = \mathcal{H}_2$ where $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable Hilbert spaces.
- G is compact and injective. Only the second and third problem are relevant here.

Our focus: Statistical, mainly Bayesian approaches to the inverse problem.
Bayesian approach to inverse problems is still new for infinite dimensions and even some basic questions in this setup are unanswered.

Let us consider the case of noisy observations

$$y = G(u) + \frac{1}{\sqrt{n}}\eta,$$

where $\eta \sim N(0, \zeta)$ is the Gaussian noise and n is the parameter controlling the intensity of noise.

- The noise may throw the observation y out of the range of G almost surely. Infact, the commonly used white noise throws the observation out of $\mathcal{H}_2$ almost surely.
- y maybe seen as element of a Banach space, which is possibly an extension of $\mathcal{H}_2$.

Definition

Estimators of u are functions of the observation y with values in $\mathcal{H}_1$.

Main concerns about estimators are **well-posedness** and **consistency**.
Setting u_0 as the **true solution**, define

- $y|u_0 \sim N(G(u_0), \frac{\zeta}{n}) \equiv \mathbb{Q}_{u_0,n}$.
- $\xi_n^{\hat{u},u_0} \equiv \|\hat{u}(y) - u_0\|_{L^2(\mathbb{Q}_{u_0,n})}$.

Definition

An estimator $\hat{u}(y)$ is said to be consistent if

$$\xi_n^{\hat{u},u_0} \to 0 \qquad \text{as } n \to \infty$$

Question: Rate of consistency ?

Definition (Minimax Rates)

Minimax rates for the model over a set S is given by

$$\xi_n(S) = \min_{\hat{u}} \max_{u_0 \in S} \xi_n^{\hat{u}, u_0}$$

Remark: Minimax rates put a bound on how quickly a posterior can approximate the true solution as noise goes to zero.

Let $\{e_i, \rho_i^2\}$ be the eigenpair of $G^T G$. The ill-posedness of the model is characterised as

- when $\rho_i^2 \approx i^{-2\alpha}$, the problem is said to be mildly ill posed, e.g - Deconvolution problems.
- when $\rho_i^2 \approx \exp(-i^\beta)$, the problem is said to be severely ill posed, e.g - Heat equation.

The sets on which the minimax rate is estimated are Sobolev balls in the basis $\{e_i\}$.

$$H^r(R) = \{ u : \Sigma (i^r \langle u, e_i \rangle)^2 \le R \}$$

Minimax Rates for $\begin{cases} \text{mildly ill posed problem:} & n^{-\frac{r}{1+2\alpha+2r}} \\ \text{severely ill posed problem:} & (\log n)^{-\frac{r}{\beta}} \end{cases}$

Ref. L. Cavalier «Nonparametric statistical inverse problems»

Next we establish the preliminaries for inverse problems (in Bayesian approach)

Reference:

① Andrew M. Stuart, Uncertainty Quantification in Bayesian inversion. ICM 2014

② Richard Nickl, Statistical Inverse Problems and PDEs: Progress and Challenges, ICM 2022.

Recall the model:

$$y = G(u) + \frac{1}{\sqrt{n}}\eta$$

Introduce **Prior** - $u \sim N(0, \frac{\mathcal{C}}{R_n^2}) \equiv \mu_n$

- The solution in the bayesian setup is given by the conditional random variable $u|y \sim \mu_n^y$.
- The prior allows us to incorporate any prior notions we might have about the behaviour of the true solution u_0.
- Functionals of posterior can serve as point estimators.

Posterior density - Bayes' theorem now gives the posterior density as

$$\frac{d\mu_n^y}{d\mu_n} = \frac{\exp\left(-\Phi(y,u)\right)}{\int_{\mathcal{H}_1} \exp\left(-\Phi(y,u)\right) d\mu_n}$$

- *The denominator is positive and finite for almost all y (Tonelli's theorem).*

Wellposedness captures the notion that the solution(posterior in this case) varies continuously with observation. This requires metrics on the relevant spaces.

Definition

Given two probability measures μ and ν and a third probability measure λ such that μ and ν has densities with respect to λ, then the **Hellinger distance** is

$$d(\mu,\nu) \equiv \sqrt{\int \left(\frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda}\right)^2 d\lambda}$$

Definition

Posterior for a model is said to be **wellposed** if there exists a Banach space $\{Y, \|.\|_Y\}$ such that observations $y \in Y$ almost surely and $y \to \mu_n^y$ is a continuous function from Y to the space of probability measures on $\mathcal{H}_1$.

Remark :

Other Distances on probability spaces are also considered, some specific distances will loose our restrictions towards the problem . ref. Jona Latz , « Bayesian Inverse Problems are usually well-posed »

$$
\begin{array}{rcl}
\textbf{Kullback–Leibler}\text{ well-posed} & & \textbf{(Lipschitz, Hellinger)}\text{ well-posed} \\
\Downarrow & & \Downarrow \\
\textbf{Total variation}\text{ well-posed} & \Leftrightarrow & \textbf{Hellinger}\text{ well-posed} \\
 & & \Downarrow \\
\textbf{Wasserstein}(p)\text{ well-posed} & \Rightarrow & \textbf{Weakly}\text{ well-posed}
\end{array}
$$

Fig. 7 *Relations between concepts of well-posedness. Here, $A \Rightarrow B$ means that* (BIP) *being A-well-posed implies that it is also B-well-posed.*

- Stuart(2010) : Well-posedness for Bayesian models on separable Banach spaces under certain sufficient technical conditions on the potential $\Phi(u, y)$ and the gaussian prior.
- Agapiou, Larsson and Stuart(2013) : The above result is used in context of our model to show its wellposedness. To satisfy the conditions, the authors have put extra conditions on the operators involved.

Besides consistency, we also consider the posterior contraction rate.

- The posterior μ_n^y is a random measure on $\mathcal{H}_1$ with randomness coming from y.
- Assuming a true solution u_0, the distribution of y is $N(G(u_0), \frac{\zeta}{n})$.
- Posterior is a good representation of the solution only if it concentrates around the true solution in some appropriate fashion as the noise goes to 0.

The random variable $X_n(\xi, y) \equiv \mu_n^y\{u : \|u - u_0\| > \xi\}$ quantifies the measure that the posterior assigns outside a ξ-ball of the true solution u_0.

Definition

The posterior is said to be **consistent** when

$$X_n(\xi, y) \to 0$$

in probability as $n \to \infty$ for all $\xi > 0$ and $u_0 \in \mathcal{H}_1$.

Contraction rates quantify how quickly the posterior converges to the true solution.

Definition

$\xi_n \to 0$ (as $n \to \infty$) is said to be a **contraction rate** for the posterior at u_0 if

$$X_n(\xi_n, y) \to 0$$

in probability as $n \to \infty$

As with minimax rates, contraction rates which are common over Sobolev balls are of interest. The discussion on contraction rates takes two main directions.

- Try to get contraction rates for a large class of priors
- Try to improve the rates by changing the parameters of the prior depending on the level of noise and even the observation y.

We've given the general framework of contraction rate, next we dig into the (technical) details towards the proof of posterior contraction rate.
Ref. Aad van der Vaart, « Fundamentals of Nonparametric Bayesian inference ».

(First we give a general generic result, next we dig into several papers to illustrate the results).

Definition 8.1 (Contraction rate) A sequence ϵ_n is a *posterior contraction rate* at the parameter θ_0 with respect to the semimetric d if $\Pi_n(\theta: d(\theta, \theta_0) \geq M_n\epsilon_n | X^{(n)}) \to 0$ in

$P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$. If all experiments share the same probability space and the convergence to zero takes place almost surely $[P_{\theta_0}^{(\infty)}]$, then ϵ_n is said to be a *posterior contraction rate in the strong sense*.

We defined "*a*" rather than *the* rate of contraction, and hence logically any rate slower than a contraction rate is also a contraction rate. Naturally we are interested in a fastest decreasing sequence ϵ_n, but in general this may not exist or may be hard to establish. Thus our rate is an upper bound for a targeted rate, and generally we are happy if our rate is equal to or close to an "optimal" rate. With an abuse of terminology we often make statements like "ϵ_n is *the* rate of contraction."

Since this definition is difficult to check directly, we introduce the theorem below. (Ref. Convergence rates of posterior distributions, Theorem 2.1)

THEOREM 2.1. *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathscr{P}_n \subset \mathscr{P}$, we have*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \le n\varepsilon_n^2, \tag{2.2}$$

$$\Pi_n(\mathscr{P} \setminus \mathscr{P}_n) \le \exp\big(-n\varepsilon_n^2(C+4)\big), \tag{2.3}$$

$$\Pi_n\Big(P\colon -P_0\Big(\log \frac{p}{p_0}\Big) \le \varepsilon_n^2, P_0\Big(\log \frac{p}{p_0}\Big)^2 \le \varepsilon_n^2\Big) \ge \exp\big(-n\varepsilon_n^2 C\big). \tag{2.4}$$

Then for sufficiently large M, we have that $\Pi_n(P\colon d(P, P_0) \ge M\varepsilon_n | X_1, \dots, X_n) \to 0$ in P_0^n-probability.

Transfer these 3 conditions in the PDE version:
(Ref. Richard Nickl, « Bayesian Non-linear Statistical Inverse Problems »).

Theorem 1.3.2. *Let Π_N be a sequence of prior Borel probability measures on some Borel subset $\Theta \subset L^2_\zeta(\mathcal{Z}, W)$, and let $\Pi_N(\cdot|(Y_i, X_i)_{i=1}^N) = \Pi_N(\cdot|D_N)$ be the resulting posterior distribution (1.12) arising from observations in model (1.9) with forward map $\mathcal{G}\colon \Theta \to L^2_\lambda(\mathcal{X}, V)$. Assume that for some fixed $\theta_0 \in \Theta$, envelope constants $U = U_N \ge 1$, $\|\mathcal{G}(\theta_0)\|_\infty \le U$, and a sequence $\delta_N \to 0$ such that $N\delta_N^2 \ge 1$, $\sqrt{N}\delta_N / U_N \to \infty$ as $N \to \infty$, the sets*

$$\mathcal{B}_N := \big\{\theta \in \Theta : d_\mathcal{G}(\theta, \theta_0) \le \delta_N, \ \|\mathcal{G}(\theta)\|_\infty \le U\big\} \tag{1.23}$$

satisfy, for all N large enough,

$$\Pi_N(\mathcal{B}_N) \ge e^{-AN\delta_N^2} \quad \text{for some } A > 0. \tag{1.24}$$

Further, assume that there exists a sequence of Borel sets $\Theta_N \subset \Theta$ for which

$$\Pi_N(\Theta_N^c) \le e^{-BN\delta_N^2} \quad \text{for some } B > A + 2, \tag{1.25}$$

and such that, for all $\bar{m} > 0$ large enough,

$$\log N(\Theta_N, d_\mathcal{G}, \bar{m}\delta_N) \le N\delta_N^2. \tag{1.26}$$

Then, for all $0 < b < B - A - 2$, we can choose $L = L(B, \bar{m}, b)$ large enough such that

$$P_{\theta_0}^N\big(\Pi_N\big(\theta \in \Theta_N, d_\mathcal{G}(\theta, \theta_0) \le L\delta_N C_{V_N}^{-1} | D_N\big) \le 1 - e^{-bN\delta_N^2}\big) \to 0 \tag{1.27}$$

as $N \to \infty$, where the constant C_V is as after (1.20), with

$$V = V_N = \max\big(U, \sup_{\theta \in \Theta_N} \|\mathcal{G}(\theta)\|_\infty\big).$$

To obtain these results $\Rightarrow$ 2 explicit conditions towards PDE Structure.

Condition 2.1.1. Consider a parameter space $\Theta \subseteq L^2_\zeta(\mathcal{Z}, W)$ and measurable map $\mathcal{G}: \Theta \to L^2_\lambda(\mathcal{X}, V)$. Let $\kappa \geq 0$. Suppose for all $M > 0$ and some normed linear subspace $(\mathcal{R}, \|\cdot\|_\mathcal{R})$ of $L^2_\zeta(\mathcal{Z}, W)$, there exist finite constants $U \geq 1$ and $L > 0$ (that may depend on M) such that

$$\sup_{\theta \in \Theta \cap B_{\mathcal{R}(M)}} \sup_{x \in \mathcal{X}} |\mathcal{G}(\theta)(x)|_V \leq U \tag{2.3}$$

and

$$\|\mathcal{G}(\theta_1) - \mathcal{G}(\theta_2)\|_{L^2_\lambda(\mathcal{X},V)} \leq L\|\theta_1 - \theta_2\|_{(H^\kappa(\mathcal{Z}))^*} \quad \text{for all } \theta_1, \theta_2 \in \Theta \cap B_\mathcal{R}(M). \tag{2.4}$$

Condition 2.1.4 (Stability estimate). Let $B_\mathcal{R}(M)$ be as in (2.2) for the regularity space $\mathcal{R}$. For some $\eta > 0$ and all $M > 0$, suppose there exists a constant $L' = L'_\mathcal{G}$ such that for all $\delta > 0$ small enough,

$$\sup\left\{\|\theta - \theta_0\|_{L^2_\zeta} : \theta \in \Theta \cap B_\mathcal{R}(M),\ \|\mathcal{G}(\theta) - \mathcal{G}(\theta_0)\|_{L^2_\lambda} \leq \delta\right\} \leq L'\delta^\eta. \tag{2.13}$$

For the Schrödinger equation this condition can be checked (see Exercise 2.4.1) using the preceding ideas. Condition 2.1.4 is relevant in non-linear inverse problems more generally as only a 'stability inequality' (2.13) is required rather than an inversion formula for $\mathcal{G}^{-1}$ such as (2.11).

Theorem 2.3.1. *Suppose the Gaussian process prior* Π_N *and forward map* $\mathcal{G}$ *satisfy the conditions of Theorem* 2.2.2 *with* Θ_N *as in* (2.20)*, some regularisation norm* $\mathcal{R}$*,* $\delta_N = N^{-(\alpha+\kappa)/(2\alpha+2\kappa+d)}$ *and some* $\kappa \geq 0$*. If* $\Pi(\cdot|(Y_i, X_i)_{i=1}^N) = \Pi(\cdot|D_N)$ *is the posterior distribution from* (1.12) *arising from observations in the model* (1.9)*, then for all* $b > 0$*, we can choose* m *large enough such that as* $N \to \infty$*,*

$$P_{\theta_0}^N\left(\Pi_N\left(\theta \in \Theta_N, \|\mathcal{G}(\theta) - \mathcal{G}(\theta_0)\|_{L^2_\lambda(\mathcal{X},V)} \leq m\delta_N | D_N\right) \leq 1 - e^{-bN\delta_N^2}\right) \to 0.$$

Moreover, assuming also Condition 2.1.4 *for the present choice of* $\mathcal{R}$ *and some* $\eta, L' > 0$*, we deduce that*

$$P_{\theta_0}^N\left(\Pi_N\left(\theta \in \Theta_N, \|\theta - \theta_0\|_{L^2_\zeta} \leq L'(m\delta_N)^\eta | D_N\right) \leq 1 - e^{-bN\delta_N^2}\right) = o(1). \quad (2.26)$$

Proof. From Theorems 1.3.2 and 2.2.2 we deduce directly the contraction rate (2.27) on the 'forward level'. Next Condition 2.1.4 implies the set inclusion

$$\left\{\theta \in \Theta_N, \|\mathcal{G}(\theta) - \mathcal{G}(\theta_0)\|_{L^2_\lambda(\mathcal{X},V)} \leq m\delta_N\right\} \subset \left\{\|\theta - \theta_0\|_{L^2_\zeta} \leq L'(m\delta_N)^\eta\right\},$$

so that (2.26) follows directly from (2.27). ■

(Analysis of specific PDEs)