

# Literature review – Bayesian inverse problems

Giacomo Garegnani

## Setting

Let us first introduce the notation employed in the following. We denote by  $\mathcal{U}$  the parameter space,  $\mathcal{V}$  the solution space and by  $\mathcal{Y}$  the observation space. Moreover, we denote by  $\mathcal{S}: \mathcal{U} \rightarrow \mathcal{V}$  the solution operator and by  $\mathcal{O}: \mathcal{V} \rightarrow \mathcal{Y}$  the observation operator. The forward operator  $\mathcal{G}: \mathcal{U} \rightarrow \mathcal{Y}$  is hence defined as  $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$ . The space of probability distributions over a generic topological space  $\mathcal{X}$  is denoted by  $\mathcal{M}(\mathcal{X})$ .

*Example 1.* Let us clarify the notations above by considering  $f: \mathbb{R}^d \rightarrow \mathbb{R}^d$  and the ODE

$$\begin{aligned} z(t; u)' &= f(t; u), \quad 0 < t \leq T \\ z(0; u) &= z_0(u), \end{aligned}$$

where  $u \in \mathcal{U}$  is a parameter. In this case, we have  $\mathcal{V} = \mathcal{C}([0, T])$  and the solution operator  $\mathcal{S}$  is defined as  $\mathcal{S}: u \mapsto z(t; u)_{t \in [0, T]}$ . Moreover, we could for example observe the solution on a series of time points  $\{t_1, t_2, \dots, t_D\} \subset [0, T]$ , so that  $\mathcal{Y} = \mathbb{R}^{Dd}$  and the observation operator is defined as  $\mathcal{O}: z(t; u)_{t \in [0, T]} \mapsto (z(t_1, u), z(t_2, u), \dots, z(t_D, u))$ .

A Bayesian inverse problem consists in determining the posterior distribution  $\mu^y$  of the parameter  $u$  from corrupted observations  $y$  defined as

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

where  $\eta$  is an additive source of noise. Assuming that  $\eta$  is distributed according to  $\mathbb{Q}_0 \in \mathcal{M}(\mathcal{Y})$ , then  $y|u$  is distributed according to  $\mathbb{Q}_u \in \mathcal{M}(\mathcal{Y})$ , translate of  $\mathbb{Q}_0$  by  $\mathcal{G}(u)$ . Moreover, we assume that  $\mathbb{Q}_u$  and  $\mathbb{Q}_0$  satisfy

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

for some function  $\Phi: \mathcal{U} \times \mathcal{Y} \rightarrow \mathbb{R}$ , called the negative log-likelihood (potential). Given a prior  $\mu_0 \in \mathcal{M}(\mathcal{U})$ , the posterior density is given by Bayes' formula, which in this context reads

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

where  $Z(y) = \mathbb{E}_{\mu_0} \exp(-\Phi(u; y))$  is the normalising constant, assumed to satisfy  $0 < Z(y) < \infty$ . Let us now consider an approximation  $\Phi_N(u; y)$  of  $\Phi(u; y)$  which depends on  $N$  degrees of freedom. This will be associated with an approximated posterior  $\mu_N^y$  defined as

$$\frac{d\mu_N^y}{d\mu_0}(u) = \frac{1}{Z(y)} \exp(-\Phi_N(u; y)). \quad (2)$$

In the following, we employ the Hellinger distance, defined for two measures  $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{X})$  which are both absolutely continuous with respect to a measure  $\Lambda \in \mathcal{M}(\mathcal{X})$  as

$$d_H(\mu_1, \mu_2)^2 := \frac{1}{2} \int_{\mathcal{X}} \left( \sqrt{\frac{d\mu_1}{d\Lambda}}(u) - \sqrt{\frac{d\mu_2}{d\Lambda}}(u) \right)^2 d\Lambda(u).$$

The Hellinger distance is independent on the choice of  $\Lambda$  and is equivalent to the total variation distance.

## Well-posed inverse problems and approximations

In the infinite-dimensional setting, it is relevant to verify whether the generalised Bayes' formula (1) defines a unique and stable posterior measure  $\mu^y$ . The following definition establishes the property of well-posedness with respect to the Hellinger metrics.

**Definition 1** (Definition 1.1. [Hosseini and Nigam, 2017]). Suppose  $\mathcal{U}$  is a Banach space. Then the inverse problem (1) is said to be *Hellinger well-posed* if:

- (i) There exists a unique  $\mu^y$  absolutely continuous with respect to  $\mu_0$  satisfying (1).
- (ii) Given  $\varepsilon > 0$ , there is a constant  $C > 0$  such that if  $\|y - y'\|_{\mathcal{Y}} < C$ , then  $d_H(\mu^y, \mu^{y'}) < \varepsilon$ .

The well-posedness of (1) depends on the function  $\Phi$  and on the choice of the prior measure  $\mu_0$ . In particular, the assumptions that  $\Phi$  has to satisfy are standard and can be found, for example, in Assumption 2.6 of [Stuart, 2010]. Another property which is desired for Bayesian inverse problem is the notion of *consistent approximation*.

**Definition 2** (Definition 1.2. in [Hosseini and Nigam, 2017]). The approximate Bayesian inverse problem (2) is a consistent approximation to (1) if  $d_H(\mu^y, \mu_N^y) \rightarrow 0$  as  $|\Phi(u; y) - \Phi_N(u; y)| \rightarrow 0$ .

In particular, it is desirable that if the approximation of  $\Phi$  by  $\Phi_N$  decays as a function  $\psi(N)$  for  $N \rightarrow \infty$ , then the Hellinger distance has the same rate of convergence.

As far as the choice of the prior measure is concerned, the analysis presented in [Stuart, 2010] proves well-posedness of (1) when  $\mathcal{U}$  is a Banach space and  $\mu_0$  is a Gaussian measure. In this case, proving well-posedness relies on the application of Fernique theorem (Theorem 6.9 in [Stuart, 2010]). Relaxations of the rather limiting Gaussian hypothesis for the prior in literature are summarised below.

- Besov priors in [Dashti et al., 2012].
- Priors with exponential tails in [Hosseini and Nigam, 2017].
- Infinitely divisible and heavy-tailed prior measures in [Hosseini, 2017].
- Heavy-tailed priors defined on quasi-Banach space (relaxation of the hypothesis that  $\mathcal{U}$  is a Banach space) in [Sullivan, 2017]. Examples of quasi-Banach spaces are  $L^p$  spaces for  $0 < p < 1$ .

In all the papers cited above, the authors prove the well-posedness of the posterior measure together with the transfer of convergence rates from the likelihood approximation to the posterior approximation.

In order to sample from the posterior distribution defined in (2), the standard Metropolis-Hastings is not suitable due to a degeneration in the quality of the Markov chain for  $N$  large. Nonetheless, there exist MCMC algorithms for which the acceptance ratio does not degenerate for refined discretisations. Two examples of these algorithms are presented in [Cotter et al., 2013, Rudolf and Sprungk, 2018], where the case of Gaussian priors is studied.

## Random forward models

In [Lie et al., 2017], the authors analyse the case where the approximation  $\Phi_N$  of the (deterministic) negative log-likelihood  $\Phi$  is given by a random variable. In particular, given a measurable space  $(\Omega, \mathcal{F}, P)$  and fixed the observation  $y \in \mathcal{Y}$  the approximation of the potential is given by the random variable  $\Phi_N: \Omega \times \mathcal{U} \rightarrow \mathbb{R}$ . This situation is typical of the approximation given by a probabilistic numerical method. Assume there exists a probability measure  $\nu_N$  on  $\Omega$  such that  $\Phi_N$  is distributed following the distribution  $\nu_N \otimes \mu_0$ . If the formula (2) is applied directly, the posterior  $\mu_N^y$  will be therefore a random variable itself. We will denote this random measure by  $\mu_{N, \text{sample}}^y: \Omega \rightarrow \mathcal{M}(\mathcal{U})$ . This notation means that for each  $\omega \in \Omega$ , the application of (2) will give a different posterior measure.

Conversely, it is possible to obtain a deterministic posterior by marginalising the first argument in  $\Phi_N$ , thus obtaining the posterior  $\mu_{N,\text{marginal}}^y$  defined as

$$\frac{d\mu_{N,\text{marginal}}^y}{d\mu_0}(u) = \frac{\mathbb{E}_{\nu_N}(\exp(-\Phi_N(\cdot, u; y)))}{\mathbb{E}_{\nu_N}(Z(y))},$$

where  $Z(y) = \mathbb{E}_{\mu_0}(\exp(-\Phi_N(\cdot, u; y)))$ . In [Lie et al., 2017], the convergence of both the posteriors  $\mu_{N,\text{sample}}^y$  and  $\mu_{N,\text{marginal}}^y$  are studied (Theorem 3.1 and 3.2 respectively). In particular, it is shown that

$$d_H(\mu^y, \mu_{N,\text{marginal}}^y) \rightarrow 0,$$

and

$$\mathbb{E}_{\nu_N}(d_H(\mu^y, \mu_{N,\text{sample}}^y)^2)^{1/2} \rightarrow 0$$

with the same rate of convergence of  $\Phi_N \rightarrow \Phi$  in some space  $L^p(\Omega)$ . Let us remark that this analysis expands the work [Stuart and Teckentrup, 2018], where only Gaussian process approximations of  $\Phi$  are taken into account.

Sampling from random posteriors requires additional care. In particular

- When sampling from  $\mu_{N,\text{sample}}$ , it is sufficient to sample from a series of posteriors  $\mu_N^{(i)} \sim \mu_{N,\text{sample}}$  with a deterministic algorithm. An approximation of  $\mu^y$  is then obtained with a Monte Carlo average.
- In order to sample from  $\mu_{N,\text{marginal}}^y$ , it is convenient to employ a pseudo-marginal MCMC algorithm. The algorithm is based on an *inner* Monte Carlo average, sampling from  $\Phi_N$  at each iteration of the MCMC algorithm. It has first been introduced in [Andrieu and Roberts, 2009], and its properties have been analysed in [Doucet et al., 2015, Sherlock et al., 2015, Deligiannidis et al., 2017]. In these articles, the optimal choice for the number of inner and outer Monte Carlo samples for minimising the computational cost is studied. Another choice could be the inexact noisy pseudo-marginal algorithm, whose properties are studied in [Medina-Aguayo et al., 2016, Alquier et al., 2016].

## References

- [Alquier et al., 2016] Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Stat. Comput.*, 26(1-2):29–47.
- [Andrieu and Roberts, 2009] Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725.
- [Cotter et al., 2013] Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446.
- [Dashti et al., 2012] Dashti, M., Harris, S., and Stuart, A. (2012). Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200.
- [Deligiannidis et al., 2017] Deligiannidis, G., Doucet, A., and Pitt, M. K. (2017). The correlated pseudo-marginal method. *arXiv preprint arXiv:1712.05717*.
- [Doucet et al., 2015] Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, pages 1 – 19.
- [Hosseini, 2017] Hosseini, B. (2017). Well-posed Bayesian inverse problems with infinitely divisible and heavy-tailed prior measures. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):1024–1060.
- [Hosseini and Nigam, 2017] Hosseini, B. and Nigam, N. (2017). Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):436–465.

- [Lie et al., 2017] Lie, H. C., Sullivan, T., and Teckentrup, A. (2017). Random forward models and log-likelihoods in bayesian inverse problems. *arXiv preprint arXiv:1712.05717*.
- [Medina-Aguayo et al., 2016] Medina-Aguayo, F. J., Lee, A., and Roberts, G. O. (2016). Stability of noisy Metropolis–Hastings. *Stat. Comput.*, 26(6):1187–1211.
- [Rudolf and Sprungk, 2018] Rudolf, D. and Sprungk, B. (2018). On a generalization of the preconditioned Crank-Nicolson metropolis algorithm. *Found. Comput. Math.*, 18(2):309–343.
- [Sherlock et al., 2015] Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275.
- [Stuart, 2010] Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.*, 19:451–559.
- [Stuart and Teckentrup, 2018] Stuart, A. M. and Teckentrup, A. L. (2018). Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Math. Comp.*, 87(310):721–753.
- [Sullivan, 2017] Sullivan, T. J. (2017). Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874.