Random forward models and log-likelihoods in Bayesian inverse problems

Han Cheng Lie[1,2] **Tim Sullivan**[1,2] Aretha Teckentrup[3,4]

SAMSI Working Group on Probabilistic Numerics
Berlin, DE, 29 January 2018

[1]Free University of Berlin, DE
[2]Zuse Institute Berlin, DE
[3]University of Edinburgh, UK
[4]Alan Turing Institute, London, UK

Overview

- We consider the use of randomised forward models and log-likelihoods within the Bayesian approach to inverse problems.
- Such random approximations to the exact forward model or log-likelihood arise naturally when a computationally expensive model is approximated using a cheaper stochastic surrogate, as in Gaussian process emulation (kriging), or in the field of probabilistic numerical methods.
- We show that the Hellinger distance between the exact and approximate Bayesian posteriors is bounded by moments of the difference between the true and approximate log-likelihoods.
- Example applications of these stability results: randomised misfit models in large data applications and the probabilistic solution of ordinary differential equations.

Lie, Sullivan, and Teckentrup (2017b)
arXiv:1712.05717

Outline

Well-posedness of Bayesian inverse problems

- $(\Omega, \mathcal{F}, \mathbb{P})$ will be an abstract probability space, assumed to be rich enough to serve as a common domain for all random variables of interest.
- $\mathcal{M}_1(\mathcal{U})$ will denote the space of Borel probability measures on a topological space $\mathcal{U}$; in practice, $\mathcal{U}$ will be a separable Banach space.
- When $\mu \in \mathcal{M}_1(\mathcal{U})$, integration of a measurable function (random variable) $f\colon \mathcal{U} \to \mathbb{R}$ will also be denoted by expectation, i.e. $\mathbb{E}_\mu[f] \coloneqq \int_{\mathcal{U}} f(u)\,\mathrm{d}\mu(u)$.
- The space $\mathcal{M}_1(\mathcal{U})$ will be endowed with the Hellinger metric $d_{\mathrm{H}}\colon \mathcal{M}_1(\mathcal{U})^2 \to \mathbb{R}_{\geq 0}$: for probability measures μ and ν on $\mathcal{U}$ that are both absolutely continuous with respect to a common reference measure π, such as $\pi \coloneqq \mu + \nu$,

$$d_{\mathrm{H}}(\mu,\nu)^2 \coloneqq \frac{1}{2}\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(u)} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right|^2 \mathrm{d}\pi(u) = 1 - \int_{\mathcal{U}} \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(u)\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(u)}\,\mathrm{d}\pi(u) = 1 - \mathbb{E}_\nu\left[\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}}\right].$$

- The Hellinger metric defines a topology on $\mathcal{M}_1(\mathcal{U})$ that coincides with the total variation topology (Kraft, 1955), is weaker than the Kullback–Leibler (relative entropy) topology (Pinsker, 1964) and is stronger than the topology of weak convergence of measures.
- The Hellinger metric is useful for uncertainty quantification when assessing the similarity of Bayesian posteriors, since expected values of square-integrable functions are Lipschitz continuous with respect to the Hellinger metric:

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

when $f \in L^2(\mathcal{U}, \mu) \cap L^2(\mathcal{U}, \nu)$. In particular, for bounded f,

$$\left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right| \leq 2\|f\|_\infty d_{\mathrm{H}}(\mu, \nu).$$

- An inverse problem means the recovery of $u \in \mathcal{U}$ from an imperfect observation $y \in \mathcal{Y}$ of $G(u)$, for a known forward operator $G\colon \mathcal{U} \to \mathcal{Y}$.
- In practice, the operator G may arise as the composition $G = O \circ S$ of the solution operator $S\colon \mathcal{U} \to \mathcal{V}$ of a system of ordinary or partial differential equations with an observation operator $O\colon \mathcal{V} \to \mathcal{Y}$, and it is typically the case that $\mathcal{Y} = \mathbb{R}^J$ for some $J \in \mathbb{N}$, whereas $\mathcal{U}$ and $\mathcal{V}$ can have infinite dimension. For simplicity, we assume an additive noise model
$$y = G(u) + \eta, \tag{1}$$
where the statistics but not the realisation of η are known.
- In the strict sense, this inverse problem is ill-posed in the sense that there may be no element u that satisfies (1), or there may be multiple such u that are highly sensitive to the observed data y.
- The Bayesian perspective eases these problems by interpreting u, y, and η all as random variables or fields.

- Through knowledge of the distribution of η, the forward equation (1) defines the conditional distribution of $y|u$.
- After positing a prior probability distribution $\mu_0 \in \mathcal{M}_1(\mathcal{U})$ for u, the Bayesian solution to the inverse problem is nothing other than the posterior distribution for the conditioned random variable $u|y$.
- This posterior measure, which we denote $\mu^y \in \mathcal{M}_1(\mathcal{U})$, is from the Bayesian point of view the proper synthesis of the prior information in μ_0 with the observed data y.
- The same posterior μ^y can also be arrived at via the minimisation of penalised Kullback–Leibler, χ^2, or Dirichlet energies (Dupuis and Ellis, 1997; Jordan and Kinderlehrer, 1996; Ohta and Takatsu, 2011), where the penalisation again expresses compromise between fidelity to the prior and fidelity to the data.

- The rigorous formulation of Bayes' formula for this context requires careful treatment and some further notation (Stuart, 2010).
- The pair (u, y) is assumed to be a well-defined random variable with values in $\mathcal{U} \times \mathcal{Y}$. The marginal distribution of u is the Bayesian prior $\mu_0 \in \mathcal{M}_1(\mathcal{U})$. The observational noise is $\eta \sim \mathbb{Q}_0 \in \mathcal{M}_1(\mathcal{Y})$, independently of u; $y|u \sim \mathbb{Q}_u$, the translate of $\mathbb{Q}_0$ by $G(u)$, which is assumed to be absolutely continuous with respect to $\mathbb{Q}_0$, with
$$\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) = \exp(-\Phi(u; y)).$$
- The function $\Phi\colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ is called the negative log-likelihood or simply potential. In the elementary setting of centred Gaussian noise $\eta \sim \mathcal{N}(0, \Gamma)$ on $\mathcal{Y} = \mathbb{R}^J$, the potential is the non-negative quadratic misfit $\Phi(u; y) = \frac{1}{2}\left\|\Gamma^{-1/2}(y - G(u))\right\|^2$. However, particularly for cases in which $\dim \mathcal{Y} = \infty$, it may be necessary to allow Φ to take negative values and even to be unbounded below (Stuart, 2010, Remark 3.8).

Theorem 1 (Generalised Bayesian formula — (Dashti and Stuart, 2016, Theorem 3.4))

Suppose that $\Phi\colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ *is* $\mu_0 \otimes \mathbb{Q}_0$*-measurable and that*

$$Z(y) \coloneqq \mathbb{E}_{\mu_0}\left[\exp(-\Phi(u;y))\right]$$

satisfies $0 < Z(y) < \infty$ *for* $\mathbb{Q}_0$*-almost all* $y \in \mathcal{Y}$*. Then, for such* y*, the conditional/posterior distribution* μ^y *of* $u|y$ *exists and has the prior density*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{\exp(-\Phi(u;y))}{Z(y)}. \tag{2}$$

For (2) to make sense, it is essential to check that $0 < Z(y) < \infty$. Hereafter, to save space, we regard the data y as fixed, and hence write $\Phi(u)$ in place of $\Phi(u;y)$, Z in place of $Z(y)$, and μ in place of μ^y.

- For a numerical analyst, it is natural to ask about the well-posedness of the Bayesian posterior μ: is it stable when the prior μ_0, the potential Φ, or the observed data y are slightly perturbed, e.g. due to discretisation, truncation, or other numerical errors?
- For example, what is the impact of using an approximate numerical forward operator G_N in place of G, and hence an approximate $\Phi_N \colon \mathcal{U} \to \mathbb{R}$ in place of Φ? Here, we quantify stability in the Hellinger metric d_{H}.
- The stability of Bayesian inverse problems with respect to the prior is a difficult topic (Owhadi et al., 2015; Owhadi and Scovel, 2017) and we will not address it here.

- However, stability with respect to the observed data y and the log-likelihood Φ can be established. Such stability results were proved for Gaussian priors by Stuart (2010) and for more general priors by many contributions since then (Dashti et al., 2012; Hosseini, 2017; Hosseini and Nigam, 2017; Sullivan, 2017).
- Typical approximation theorems for the replacement of the potential Φ by a deterministic approximate potential Φ_N, leading to an approximate posterior μ_N, aim to transfer the convergence rate of the forward problem to the inverse problem, i.e. to prove an implication of the form

$$\left|\Phi(u) - \Phi_N(u)\right| \leq M(\|u\|)\psi(N) \implies d_{\mathrm{H}}(\mu, \mu_N) \leq C\psi(N),$$

  where $M\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is suitably well-behaved, $\psi\colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ quantifies the convergence rate of the forward problem, and C is a constant.
- Following Stuart and Teckentrup (2017), we extend this paradigm and these approximation results to the case in which the approximation Φ_N is a *random* object, and not necessarily Gaussian.

Random log-likelihoods

- In many practical applications, the negative log-likelihood Φ is computationally too expensive or impossible to evaluate exactly, and in computations, one therefore often uses an approximation Φ_N of Φ.
- This leads to an approximation μ_N of the exact posterior distribution μ, and the aim is to show convergence, in a suitable sense, of μ_N to μ as the approximation error $\Phi_N - \Phi$ in the misfit potential tends to zero.
- We focus on random approximations Φ_N, e.g. Gaussian process emulators (Stuart and Teckentrup, 2017); later, we give examples related to randomised misfit models and probabilistic numerical methods.
- Let now $\Phi_N \colon \Omega \times \mathcal{U} \to \mathbb{R}$ be a measurable function that provides a random approximation to $\Phi \colon \mathcal{U} \to \mathbb{R}$, and denote by ν_N the distribution of Φ_N.
- We assume throughout that the randomness in the approximation Φ_N of Φ is independent of that in the parameters being inferred.

- Replacing Φ by Φ_N in (2), we obtain a random approximation μ_N^{sample} of μ:

$$\frac{\mathrm{d}\mu_N^{\text{sample}}}{\mathrm{d}\mu_0}(u) := \frac{\exp(-\Phi_N(u))}{Z_N^{\text{sample}}}, \tag{3}$$

$$Z_N^{\text{sample}} := \mathbb{E}_{\mu_0}\left[\exp(-\Phi_N(\cdot))\right].$$

- Taking the expectation of the random likelihood gives a deterministic approximation:

$$\frac{\mathrm{d}\mu_N^{\text{marginal}}}{\mathrm{d}\mu_0}(u) := \frac{\mathbb{E}_{\nu_N}\left[\exp(-\Phi_N(u))\right]}{\mathbb{E}_{\nu_N}\left[Z_N^{\text{sample}}\right]}. \tag{4}$$

- An alternative deterministic approximation can be obtained by taking the expected value of the density $(Z_N^{\text{sample}})^{-1}e^{-\Phi_N(u)}$ in (3). However, μ_N^{marginal} provides a clear interpretation as the posterior obtained by the approximation of the true data likelihood $e^{-\Phi(u)}$ by $\mathbb{E}_{\nu_N}\left[e^{-\Phi_N(u)}\right]$, and is more amenable to sampling methods such as pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009).

Theorem 2 (Deterministic convergence of the marginal posterior)

Suppose there exist positive scalars C_1, C_2, C_3, that do not depend on N, such that for the Hölder conjugate exponent pairs (p_1, p_1'), (p_2, p_2'), and (p_3, p_3'), we have

- $\min\left\{ \left\| \mathbb{E}_{\nu_N}[e^{-\Phi_N}]^{-1} \right\|_{L^{p_1}_{\mu_0}(\mathcal{U})}, \left\| e^{\Phi} \right\|_{L^{p_1}_{\mu_0}(\mathcal{U})} \right\} \leq C_1(p_1)$;
- $\left\| \mathbb{E}_{\nu_N}\left[\left(e^{-\Phi} + e^{-\Phi_N} \right)^{p_2} \right]^{1/p_2} \right\|_{L^{2p_1' p_3}_{\mu_0}(\mathcal{U})} \leq C_2(p_1, p_2, p_3)$;
- $C_3^{-1} \leq \mathbb{E}_{\nu_N}[Z_N^{\text{sample}}] \leq C_3$.

Then there exists a scalar $C = C(C_1, C_2, C_3, Z) > 0$ that does not depend on N, such that

$$d_{\mathrm{H}}\left(\mu, \mu_N^{\text{marginal}}\right) \leq C \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{p_2'} \right]^{1/p_2'} \right\|_{L^{2p_1' p_3'}_{\mu_0}(\mathcal{U})},$$

$$C(C_1, C_2, C_3, Z) = \left(\frac{C_1(p_1)}{Z} + C_3 \max\left\{ Z^{-3}, C_3^3 \right\} \right) C_2^2(p_1, p_2, p_3).$$

Theorem 3 (Mean-square convergence of the sample posterior)

Suppose there exist positive scalars D_1, D_2, that do not depend on N, such that for Hölder conjugate exponent pairs (q_1, q_1') and (q_2, q_2'), we have

- $\left\| \mathbb{E}_{\nu_N}\left[\left(e^{-\Phi/2} + e^{-\Phi_N/2} \right)^{2q_1} \right]^{1/q_1} \right\|_{L^{q_2}_{\mu_0}(\mathcal{U})} \leq D_1(q_1, q_2);$
- $\left\| \mathbb{E}_{\nu_N}\left[\left(Z_N^{\text{sample}} \max\left\{ Z^{-3}, (Z_N^{\text{sample}})^{-3} \right\} \left(e^{-\Phi} + e^{-\Phi_N} \right)^2 \right)^{q_1} \right]^{1/q_1} \right\|_{L^{q_2}_{\mu_0}(\mathcal{U})} \leq D_2(q_1, q_2).$

Then

$$\mathbb{E}_{\nu_N}\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\text{sample}} \right)^2 \right]^{1/2} \leq (D_1 + D_2) \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{2q_1'} \right]^{1/2q_1'} \right\|_{L^{2q_2'}_{\mu_0}(\mathcal{U})},$$

The assumptions of Theorems 2 and 3 are satisfied when the exact potential Φ and the approximation quality $\Phi_N \approx \Phi$ are suitably well behaved. Recall from (2) that Z is the normalisation constant of μ. Therefore, for μ to be well-defined, we must have that $0 < Z < \infty$. In particular, there exists $0 < C_3 < \infty$ such that $C_3^{-1} < Z < C_3$.

Assumption 4

There exists $C_0 \in \mathbb{R}$ that does not depend on N such that, for all $N \in \mathbb{N}$,

$$\Phi \geq -C_0 \quad \text{and} \quad \nu_N(\{\Phi_N \mid \Phi_N \geq -C_0\}) = 1, \tag{5}$$

and for any $0 < C_3 < +\infty$ with the property that $C_3^{-1} < Z < C_3$, there exists $N^*(C_3) \in \mathbb{N}$ such that, for all $N \geq N^*$,

$$\mathbb{E}_{\mu_0}[\mathbb{E}_{\nu_N}[|\Phi_N - \Phi|]] \leq \frac{1}{2e^{C_0}} \min\left\{Z - \frac{1}{C_3}, C_3 - Z\right\}. \tag{6}$$

Lemma 5

Suppose that Assumption 4 holds with C_0 as in (5) and C_3 and $N^*(C_3)$ as in (6), that $\exp(\Phi) \in L^{p^*}_{\mu_0}(\mathcal{U})$ for some $1 \leq p^* \leq +\infty$ with conjugate exponent $(p^*)'$, and there exists some $C_4 \in \mathbb{R}$ that does not depend on N, such that, for all $N \in \mathbb{N}$,

$$\nu_N(\{\Phi_N \mid \mathbb{E}_{\mu_0}[\Phi_N] \leq C_4\}) = 1. \tag{7}$$

Then the hypotheses of Theorem 2 hold, with

$$p_1 = p^*,\ p_2 = p_3 = +\infty,\ C_1 = \|e^{\Phi}\|_{L^{p^*}_{\mu_0}},\ C_2 = 2e^{C_0},$$

and C_3 as above. Moreover, the hypotheses of Theorem 3 hold, with

$$q_1 = q_2 = \infty,\ D_1 = 4e^{C_0},\ D_2 = 4e^{3C_0} \max\{C_3^{-3}, e^{3C_4}\}.$$

Lemma 6

Suppose that Assumption 4 holds with C_0 as in (5) and C_3 and $N^*(C_3)$ as in (6), and that there exists some $2 < \rho^* < +\infty$ such that $\mathbb{E}_{\nu_N}[\exp(\rho^* \Phi_N)] \in L^1_{\mu_0}$. Then the hypotheses of Theorem 2 hold, with

$$p_1 = \rho^*,\ p_2 = p_3 = +\infty,\ C_1 = \|\mathbb{E}_{\nu_N}[\exp(\rho^* \Phi_N)]\|^{1/\rho^*}_{L^1_{\mu_0}},\ C_2 = 2e^{C_0},$$

and C_3 as above. Moreover, the hypotheses of Theorem 3 hold, with

$$q_1 = \frac{\rho^*}{2},\ q_2 = +\infty,\ D_1 = 4e^{C_0},\ D_2 = 4e^{2C_0}\left(C_3^{-3} e^{C_0} + \|\mathbb{E}_{\nu_N}[e^{\rho^* \Phi_N}]\|^{2/\rho^*}_{L^1_{\mu_0}}\right).$$

Random forward models

In many settings, the potentials Φ and Φ_N have a common form and differ only in the parameter-to-observable map. In this section we shall assume that Φ and Φ_N are quadratic misfits of the form

$$\Phi(u) = \frac{1}{2}\left\|\Gamma^{-1/2}(G(u) - y)\right\|_{\mathcal{Y}}^2 \quad \text{and} \quad \Phi_N(u) = \frac{1}{2}\left\|\Gamma^{-1/2}(G_N(u) - y)\right\|_{\mathcal{Y}}^2, \tag{8}$$

corresponding to centred Gaussian observational noise with symmetric positive-definite covariance Γ. Again, we assume that G is deterministic while G_N is random. In this section, for this setting, we show how the quality of the approximation $G_N \approx G$ transfers to the approximation $\Phi_N \approx \Phi$, and hence to the approximation $\mu_N \approx \mu$ (for either the sample or marginal approximate posterior).

It is easy to ensure that the applicability lemmas for Theorems 2 and 3 apply:

Lemma 7

Let Φ and Φ_N be as in (8)*. Assumption 4 holds if, for some $q, s \geq 1$,*

$$\lim_{N\to\infty} \left\| \mathbb{E}_{\nu_N}\left[\|G_N - G\|^{2q} \right]^{1/q} \right\|_{L^s_{\mu_0}} = 0. \tag{9}$$

We then obtain bounds on the Hellinger error in terms of errors in the forward model, of the following form: for $C, D > 0$ and $r_1, r_2, s_1, s_2 \geq 1$ that do not depend on N,

$$d_{\mathrm{H}}\left(\mu, \mu_N^{\text{marginal}}\right) \leq C \left\| \mathbb{E}_{\nu_N}\left[\|G_N - G\|^{2r_1} \right]^{1/r_1} \right\|^{1/2}_{L^{r_2}_{\mu_0}(\mathcal{U})} \tag{10}$$

$$\mathbb{E}_{\nu_N}\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\text{sample}}\right)^2 \right]^{1/2} \leq D \left\| \mathbb{E}_{\nu_N}\left[\|G_N - G\|^{2s_1} \right]^{1/s_1} \right\|^{1/2}_{L^{s_2}_{\mu_0}(\mathcal{U})}. \tag{11}$$

For brevity and simplicity, the following result uses one pair $q, s \geq 1$ in (9) in order to obtain convergence statements for *both* μ_N^{marginal} and μ_N^{sample}. Of course, one may optimise q and s depending on the measure of interest.

Theorem 8 (Convergence of posteriors for randomised forward models in quadratic potentials)

Let Φ and Φ_N be as in (8).

- *Suppose there exists some $p^* > 1$ with Hölder conjugate $(p^*)'$ such that $\exp(\Phi) \in L^{p^*}_{\mu_0}(\mathcal{U})$, and suppose that (7) holds for some $C_4 \in \mathbb{R}$. If $G_N \to G$ as in (9) with $q = 2$ and $s = 2p^*/(p^* - 1)$, then (10) holds with $r_1 = 1$ and $r_2 = 2p^*/(p^* - 1)$, and (11) holds with $s_1 = 2$ and $s_2 = 2$.*
- *Suppose there exists some $2 < \rho^* < \infty$ such that $\mathbb{E}_{\nu_N}[\exp(\rho^* \Phi_N)] \in L^1_{\mu_0}$. If $G_N \to G$ as in (9) with $q = 2\rho^*/(\rho^* - 2)$ and $s = 2\rho^*/(\rho^* - 1)$, then (10) holds with $r_1 = 1$ and $r_2 = 2\rho^*/(\rho^* - 1)$ and (11) holds with $s_1 = 2\rho^*/(\rho^* - 2)$ and $s_2 = 2$.*

In both cases, μ_N^{marginal} and μ_N^{sample} converge to μ in the appropriate metrics given in (10) and (11) respectively.

Application: random reduction of high-dimensional data

We consider a Monte Carlo approximation Φ_N of a quadratic potential Φ (Nemirovski et al., 2008; Shapiro et al., 2009), further applied and analysed in the MAP estimator context by Le et al. (2017). This approximation is particularly useful when the data $y \in \mathbb{R}^J$ has $J \gg 1$, and is derived in the following way:

$$
\begin{aligned}
\Phi(u) &:= \frac{1}{2}\left\|\Gamma^{-1/2}(y - G(u))\right\|^2 \\
&= \frac{1}{2}\left(\Gamma^{-1/2}(y - G(u))\right)^{\mathrm{T}} \mathbb{E}[\sigma\sigma^{\mathrm{T}}]\left(\Gamma^{-1/2}(y - G(u))\right) && \text{where } \mathbb{E}[\sigma] = 0 \in \mathbb{R}^J,\ \mathbb{E}[\sigma\sigma^{\mathrm{T}}] = I_{J\times J} \\
&= \frac{1}{2}\mathbb{E}\left[\left|\sigma^{\mathrm{T}}\left(\Gamma^{-1/2}(y - G(u))\right)\right|^2\right] \\
&\approx \frac{1}{2N}\sum_{i=1}^{N}\left|\sigma^{(i)^{\mathrm{T}}}\left(\Gamma^{-1/2}(y - G(u))\right)\right|^2 && \text{for i.i.d. } \sigma^{(1)}, \dots, \sigma^{(N)} \overset{\mathrm{d}}{=} \sigma \\
&= \frac{1}{2}\left\|\Sigma_N^{\mathrm{T}}\left(\Gamma^{-1/2}(y - G(u))\right)\right\|^2 && \text{for } \Sigma_N := \frac{1}{\sqrt{N}}[\sigma^{(1)} \cdots \sigma^{(N)}] \in \mathbb{R}^{J\times N} \\
&=: \Phi_N(u).
\end{aligned}
$$

The analysis and numerical studies in Le et al. (2017, Sections 3–4) suggest that a good choice for the $\mathbb{R}^J$-valued random vector σ would be one with independent and identically distributed (i.i.d.) entries from a sub-Gaussian probability distribution. Examples of sub-Gaussian distributions considered include

- the Gaussian distribution: $\sigma_j \sim \mathcal{N}(0,1)$, for $j = 1, \ldots, J$; and
- the ℓ-sparse distribution: for $\ell \in [0,1)$, let $s := \frac{1}{1-\ell} \geq 1$ and set, for $j = 1, \ldots, J$,

$$\sigma_j := \sqrt{s} \begin{cases} 1, & \text{with probability } \frac{1}{2s}, \\ 0, & \text{with probability } \ell = 1 - \frac{1}{s}, \\ -1, & \text{with probability } \frac{1}{2s}. \end{cases}$$

- Le et al. (2017) observe that, for large J and moderate $N \approx 10$, the random potential Φ_N and the original potential Φ are very similar, in particular having approximately the same minimisers and minimum values.
- Statistically, these correspond to the maximum likelihood estimators under Φ and Φ_N being very similar; after weighting by a prior, this corresponds to similarity of maximum a posteriori (MAP) estimators.
- Here, we take a fully Bayesian perpective, and thus the corresponding conjecture is that the deterministic posterior $\mathrm{d}\mu(u) \propto \exp(-\Phi(u))\,\mathrm{d}\mu_0(u)$ is well approximated by the random posterior $\mathrm{d}\mu_N^{\text{sample}}(u) \propto \exp(-\Phi_N(u))\,\mathrm{d}\mu_0(u)$.
- Indeed, via Theorem 3, we have the following convergence result for the case of a sparsifying distribution:

Applying the general results from earlier gives the following transfer of the Monte Carlo convergence rate from the approximation ot Φ to the approximation of μ:

Proposition 9

Suppose that the entries of σ are i.i.d. ℓ-sparse, for some $\ell \in [0,1)$, and that $\Phi \in L^2_{\mu_0}(\mathcal{U})$. Then there exists a constant C, independent of N, such that

$$\left(\mathbb{E}_\sigma\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\text{sample}}\right)^2\right]\right)^{1/2} \leq \frac{C}{\sqrt{N}}.$$

For technical reasons to do with the non-compactness of the support and finiteness of MGFs of maxima, the current proof technique does not work for the Gaussian case.

Application: probabilistic solvers for ODEs

- Random approximate solution of deterministic ODEs is an old idea (Skilling, 1992) that has received renewed attention in recent years (Schober et al., 2014; Conrad et al., 2016; Hennig et al., 2015; Lie et al., 2017a). As random forward models, these probabilistic ODE solvers are amenable to the earlier general analysis.
- We consider an autonomous ODE of the form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} z(t;u) &= f(z(t;u);u), \qquad && \text{for } 0 \le t \le T, \\ z(0;u) &= z_0(u). \end{aligned} \tag{12}$$

  over a fixed time horizon $[0, T]$, where the unknown parameter u will appear in the definition of the initial condition $z_0 = z_0(u)$ or the vector field $f = f_u \colon \mathbb{R}^d \to \mathbb{R}^d$, resulting in the parameter-dependent solution $z(t;u)$.
- Define the deterministic (exact) solution operator

$$S \colon \mathcal{U} \to C([0,T];\mathbb{R}^d), \quad u \mapsto S(u) := (z(t;u))_{t\in[0,T]},$$

  where $(z(t;u))_{t\in[0,T]}$ solves (12). We equip $C([0,T];\mathbb{R}^d)$ with the supremum norm.

- Denote by $\Phi^t \colon \mathbb{R}^d \to \mathbb{R}^d$ the flow map associated to the initial value problem (12), i.e. $\Phi^t(z_0) \coloneqq z(t; u) = S(u)(t)$.
- Fix a constant time step $\tau > 0$ such that $N \coloneqq T/\tau \in \mathbb{N}$, and a time grid

$$t_k \coloneqq k\tau \text{ for } k \in [N] \coloneqq \{0, 1, \dots, N\}. \tag{13}$$

  we shall denote by $z_k \coloneqq z(t_k) \equiv \Phi^\tau(z_{k-1})$ the value of the exact solution to (12) at time t_k. We shall sometimes abuse notation and write $[N] = \{0, 1, \dots, N-1\}$ or $[N] = \{1, 2, \dots, N\}$.
- To a single-step numerical integration method (e.g. a Runge–Kutta method of some order) we shall associate a numerical flow map $\Psi^\tau \colon \mathbb{R}^d \to \mathbb{R}^d$. The numerical flow map approximates the sequence $(z_k)_{k\in[N]}$ by a sequence $(Z'_k)_{k\in[N]}$, where $Z'_k \coloneqq \Psi^\tau(Z'_{k-1})$.

- A fundamental task in numerical analysis is to determine sufficient conditions for convergence of the sequence $(Z'_k)_{k\in[N]}$ to $(z_k)_{k\in[N]}$. The investigations of Conrad et al. (2016) and Lie et al. (2017a) concern a similar task in the context of uncertainty quantification.
- Given $\tau > 0$, consider a collection $(\xi_k)_{k\in[N]}$ of stochastic processes $\xi_k \colon \Omega \times [0,\tau] \to \mathbb{R}^d$. Define a stochastic process $(Z_t)_{t\in[0,T]}$ in terms of a new randomised integrator

$$Z(t_{k+1}; u) \coloneqq \Psi^\tau(Z(t_k; u)) + \xi_k(\tau), \tag{14}$$

where for each $k \in [N]$, $\xi_k \colon \Omega \times [0,\tau] \to \mathbb{R}^d$ is a stochastic process. The stochastic processes $(\xi_k)_{k\in[N]}$ are intended to capture the effect of uncertainties, e.g. those that arise due to properties of the vector field that are not resolved at the resolution given by the time step τ. We extend the definition (14) to continuous time via

$$Z(t; u) \coloneqq \Psi^{t-t_k}(Z(t_k; u)) + \xi_k(t - t_k), \quad \text{for } t_k < t < t_{k+1}. \tag{15}$$

- Note that the $(\xi_k)_{k\in[N]}$ do not depend on the parameter u; this is because, in order to be able to estimate the parameter $u \in \mathcal{U}$, we need our model of the uncertainty to not depend on the parameter that we wish to estimate. From another point of view, since the distribution of the $(\xi_k)_{k\in[N]}$ plays the role of the measure ν_N in the general case, and since ν_N does not depend on $u \in \mathcal{U}$, it follows that the $(\xi_k)_{k\in[N]}$ cannot depend on u either. On the other hand, the map Ψ^τ does depend on the parameter $u \in \mathcal{U}$, because Ψ^τ must involve the vector field $f(\cdot; u)$.
- Define the random solution operator associated to the randomised numerical integrator (15):

$$S_N \colon \mathcal{U} \to C([0,T];\mathbb{R}^d), \quad u \mapsto S_N(u) \coloneqq (Z(t;u))_{t\in[0,T]}.$$

Recall that we equip $C([0,T];\mathbb{R}^d)$ with the supremum norm topology.

- Let $T_J \subset [0, T]$ be a strictly increasing sequence of time points, indexed by a finite, nonempty index set J with cardinality $|J| \in \mathbb{N}$. Define $\mathcal{Y} = \mathbb{R}^{d|J|}$, and equip it with the topology induced by the standard Euclidean inner product. Define the observation operator

$$O \colon C([0, T]; \mathbb{R}^d) \to \mathcal{Y}, \quad \tilde{z} \mapsto O(\tilde{z}) \coloneqq (\tilde{z}(t_j))_{t_j \in T_J},$$

  which projects some $\tilde{z} \in C([0, T]; \mathbb{R}^d)$ to a finite-dimensional vector in $\mathcal{Y}$ constructed by stacking the $\mathbb{R}^d$-valued vectors resulting from evaluating $\tilde{z}$ at the $|J|$ time points in T_J. For any $m \in \mathbb{N}$, the topology on $\mathbb{R}^m$ is thus equivalent to that induced by the ℓ_2 norm $\|\cdot\|_{\ell_2^m}$. Therefore, we take $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_{\ell_2^{d|J|}}$.
- Given the operators S, O, and S_N defined above, we define $\mathcal{Y} \coloneqq (\mathbb{R}^d)^{|J|}$, the forward operators $G, G_N \colon \mathcal{U} \to \mathcal{Y}$ by

$$G \coloneqq O \circ S, \quad G_N \coloneqq O \circ S_N,$$

  and the associated likelihoods are the quadratic misfits given by (8) with some fixed, positive-definite matrix Γ.

- We define the discrete-time error process in terms of the time grid (13) by

$$e_k(u) \equiv e(t_k; u) \coloneqq z(t_k; u) - Z(t_k; u), \quad k \in [N],$$

and the continuous-time error process by

$$e(t; u) \coloneqq z(t; u) - Z(t; u), \quad 0 \leq t \leq T.$$

Since T_J is a proper subset of $[0, T]$, it follows that

$$\|G_N(u) - G(u)\| \equiv \|G_N(u) - G(u)\|_{\mathcal{Y}} \leq \sup_{0 \leq t \leq T} \|e(t; u)\|_{\ell_2^d}.$$

- This completes our formulation of the probabilistic numerical integration of the ODE (12) as a random likelihood model of the general type considered above.
- We now just need to couple a convergence result for the random numerical integrator to the convergence results for the approximate BIP posteriors over u.

Assumption 10 (Assumptions 3.1–3.3, Lie et al. (2017a))

The vector field f admits $0 < \tau^* \leq 1$ and $C_\Phi \geq 1$, such that for $0 < \tau < \tau^*$, the flow map $\Phi^\tau : \mathbb{R}^d \to \mathbb{R}^d$ of the ODE is globally Lipschitz, with

$$\|\Phi^\tau(z_0) - \Phi^\tau(v_0)\| \leq (1 + C_\Phi \tau)\|z_0 - v_0\|, \quad \text{for all } z_0, v_0 \in \mathbb{R}^d.$$

The numerical method Ψ^τ has uniform local truncation error of order $q+1$: for some constant $C_\Psi \geq 1$ that does not depend on τ,

$$\sup_{u \in \mathbb{R}^d} \|\Psi^\tau(u) - \Phi^\tau(u)\| \leq C_\Psi \tau^{q+1}.$$

The stochastic processes $(\xi_k)_{k \in \mathbb{N}}$ admit $p \geq 1$, $R \in \mathbb{N} \cup \{+\infty\}$, and $C_{\xi,R} \geq 1$, independent of k and τ, such that for all $1 \leq r \leq R$ and all $k \in \mathbb{N}$,

$$\mathbb{E}_{\nu_N}\left[\sup_{0<t\leq T/N} \|\xi_k(t)\|^r\right] \leq \left(C_{\xi,R}\left(\frac{T}{N}\right)^{p+1/2}\right)^r.$$

Theorem 11 (Theorem 5.2, Lie et al. (2017a))

Let $n \in \mathbb{N}$, and suppose that Assumptions 10 hold with parameters τ^*, C_Φ, C_Ψ, q, $C_{\xi,R}$, p, and R. Then, for all $T/\tau^* < N$,

$$\mathbb{E}_{\nu_N}\left[\sup_{0\le t\le T}\|e(t;u)\|^n\right] \le 3^{n-1}\left((1+C_\Phi\tau^*)^n\overline{C} + C_\Psi^n(\tau^*)^n + TC_{\xi,R}^n\right)\left(\frac{T}{N}\right)^{n(q\wedge(p-1/2))},$$

where the following scalars do not depend on τ but depend on $u \in \mathcal{U}$:

$$\overline{C} := 2T\max\{(4C_\Psi)^n, (2C_{\xi,R})^n\}\exp(TC_\Phi(n,\tau^*))$$
$$C_\Phi(n,t) := \left[(1+t2^{n-1})^2(1+tC_\Phi)^n - 1\right]t^{-1}.$$

As a corollary, we even get finiteness of the moment generating function of the error.

Theorem 12

Suppose that $\mathcal{U}$ is a compact subset of $\mathbb{R}^m$ for some $m \in \mathbb{N}$, and suppose that $S, S_N \colon \mathcal{U} \to C([0,T];\mathbb{R}^d)$ are continuous maps. Let $2 < \rho^* < \infty$ be arbitrary. Suppose that Assumptions 10 hold with parameters τ^*, C_Φ, C_Ψ, q, $R = +\infty$, $C_{\xi,R}$, and p, and that these parameters depend continuously on u. Then, for $N \in \mathbb{N}$ such that $T/\tau^* < N$, (10) holds for $r_1 = 1$ and $r_2 = 2\rho^*/(\rho^* - 1)$, and (11) holds for $s_1 = 2\rho^*/(\rho^* - 2)$ and $s_2 = 2$.

References

References I

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2): 697–725, 2009. doi:10.1214/07-AOS574.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3): 1139–1160, 2003.

P. R. Conrad, M. Girolami, S. Särkkä, A. M. Stuart, and K. C. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Stat. Comput.*, 2016. doi:10.1007/s11222-016-9671-0.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification*, pages 311–428. 2016. doi:10.1007/978-3-319-11259-6_7-1.

M. Dashti, S. Harris, and A. M. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012. doi:10.3934/ipi.2012.6.183.

P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. doi:10.1002/9781118165904.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proc. A.*, 471(2179): 20150142, 17, 2015. doi:10.1098/rspa.2015.0142.

B. Hosseini. Well-posed Bayesian inverse problems with infinitely-divisible and heavy-tailed prior measures. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):1024–1060, 2017. doi:10.1137/16M1096372.

References II

B. Hosseini and N. Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):436–465, 2017. doi:10.1137/16M1076824.

R. Jordan and D. Kinderlehrer. An extended variational principle. In *Partial Differential Equations and Applications*, volume 177 of *Lecture Notes in Pure and Appl. Math.*, pages 187–200. Dekker, New York, 1996. doi:10.5006/1.3292113.

C. H. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.*, 2:125–141, 1955.

E. B. Le, A. Myers, T. Bui-Thanh, and Q. P. Nguyen. A data-scalable randomized misfit approach for solving large-scale PDE-constrained inverse problems. *Inverse Probl.*, 33(6):065003, 2017. doi:10.1088/1361-6420/aa6cbd.

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations, 2017a. arXiv:1703.03680.

H. C. Lie, T. J. Sullivan, and A. L. Teckentrup. Random forward models and log-likelihoods in Bayesian inverse problems, 2017b. arXiv:1712.05717.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008. doi:10.1137/070704277.

S. Ohta and A. Takatsu. Displacement convexity of generalized relative entropies. *Adv. Math.*, 228(3):1742–1787, 2011. doi:10.1016/j.aim.2011.06.029.

References III

H. Owhadi and C. Scovel. Qualitative robustness in Bayesian inference. *ESAIM Probab. Stat.*, 21:251–274, 2017. doi:10.1051/ps/2017014.

H. Owhadi, C. Scovel, and T. J. Sullivan. Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Stat.*, 9(1):1–79, 2015. doi:10.1214/15-EJS989.

M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1964.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge–Kutta means. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 739–747. Curran Associates, Inc., 2014. http://papers.nips.cc/paper/5451-probabilistic-ode-solvers-with-runge-kutta-means.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, volume 9 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2009. doi:10.1137/1.9780898718751.

J. Skilling. Bayesian solution of ordinary differential equations. In C. R. Smith, G. J. Erickson, and P. O. Neudorfer, editors, *Maximum Entropy and Bayesian Methods*, volume 50 of *Fundamental Theories of Physics*, pages 23–37. Springer, 1992. doi:10.1007/978-94-017-2219-3.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. doi:10.1017/S0962492910000061.

References iv

A. M. Stuart and A. L. Teckentrup. Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Math. Comput.*, 2017. doi:10.1090/mcom/3244.

T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874, 2017. doi:10.3934/ipi.2017040.