Well-Posedness of Bayesian Inverse Problems

Stable Priors on Quasi-Banach Spaces

Tim Sullivan[1]

Applied and Computational Mathematics Seminar

University of Edinburgh, UK

5 April 2017

[1]Free University of Berlin & Zuse Institute Berlin, Germany

Overview

Introduction

Bayesian Inverse Problems

Stable Distributions on $\mathbb{R}$ and Quasi-Banach Spaces

Well-Posedness of BIPs

Closing Remarks

Introduction

Introduction I

- The Bayesian perspective on inverse problems has attracted much mathematical attention in recent years (Kaipio and Somersalo, 2005; Stuart, 2010).
- Particular attention has been paid to Bayesian inverse problems (BIPs) in which the parameter to be inferred lies in an infinite-dimensional space, a typical example being a scalar or tensor field u coupled to some observed data y via an ODE or PDE.
- Numerical solution of such infinite-dimensional BIPs must necessarily be performed in an approximate manner on a finite-dimensional subspace, but it is profitable to delay discretisation to the last possible moment.
- Infinite-dimensional well-posedness results and algorithms descend to finite-dimensional subspaces in a discretisation-independent way

INTRODUCTION II

- Careless early discretisation may lead to a sequence of well-posed finite-dimensional BIPs or algorithms whose stability properties degenerate as the discretisation dimension increases.
- Well-posedness results have been established for infinite-dimensional Gaussian priors by Stuart (2010), for Besov priors by Dashti et al. (2012), and for log-concave priors by Hosseini and Nigam (2016).
- Parallel approach of discretisation invariance, introduced by Lehtinen in the 1990s and advanced by e.g. Lassas et al. (2009): finite-dimensional BIP is the primary object, but care is taken to ensure the existence of a well-defined continuum limit.
- A common assumption in these works is some exponential integrability of the prior, and one aim here is to relax this by permitting the prior to be heavy-tailed in the sense of only having finite polynomial moments of order $0 \leq p < \alpha$ for some $\alpha < \infty$.

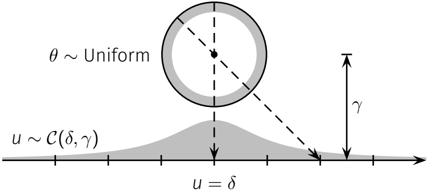
- The Cauchy distribution on $\mathbb{R}$ with location $\delta \in \mathbb{R}$ and width $\gamma > 0$:

$$\frac{\mathrm{d}\mathcal{C}(\delta,\gamma)}{\mathrm{d}u}(u) = \frac{1}{\gamma\pi}\frac{1}{1+((u-\delta)/\gamma)^2}. \tag{1}$$

- $\mathcal{C}(\delta,\gamma)$ is the law of the ratio of two independent Gaussians:

$$\delta + \frac{x}{z} \sim \mathcal{C}(\delta,\gamma) \text{ when } x \sim \mathcal{N}(0,\gamma^2), z \sim \mathcal{N}(0,1) \text{ are independent.}$$

- $\mathcal{C}(\delta,\gamma)$ is radial projection of uniform angular measure onto a line:



Prototype: Cauchy Distributions II

- $\mathcal{C}(\delta, \gamma)$ has no well-defined mean, even though it is ‘obviously’ centred on its median/mode δ, nor indeed polynomial moments of any order greater than $\alpha = 1$.
- Despite this, the Cauchy distribution arises naturally in even quite elementary applications.
- Markkanen et al. (2016) have recently proposed the use of heavy-tailed priors for edge-preserving Bayesian inversion in X-ray tomography, where the seemingly natural choice of a total variation regularisation term cannot be interpreted as a discretisation-invariant Bayesian prior (Lassas and Siltanen, 2004).

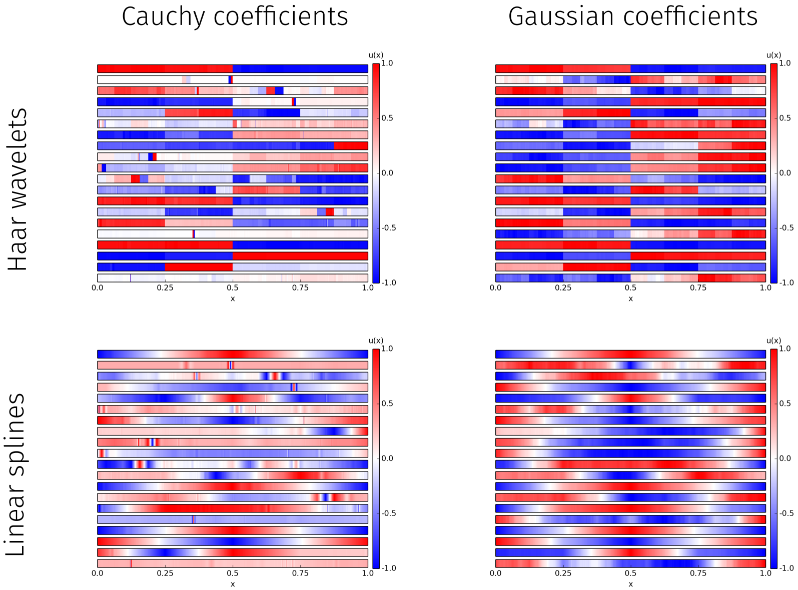
Samples from Heavy-Tailed Laws I

- Using a heavy-tailed prior instead of one with exponentially small tails corresponds to a prior belief that large deviations are not exponentially rare events.
- The information-theoretic difference between the two modelling assumptions can be extreme: the Kullback–Leibler divergences (relative entropies) are

$$D_{\mathrm{KL}}(\mathcal{N}(0,1)\|\mathcal{C}(0,1)) \approx 0.2592 < \infty \quad \text{but} \quad D_{\mathrm{KL}}(\mathcal{C}(0,1)\|\mathcal{N}(0,1)) = \infty.$$

- Thus, the approximation of a heavy-tailed Cauchy prior by a thin-tailed Gaussian prior represents an infinite information loss.
- However, asymmetrically, the 'defensive' adoption of a Cauchy prior in place of a Gaussian one represents a mild information loss, with which one gains access to large deviations that would be exponentially rare in the Gaussian model.

Samples from Heavy-Tailed Laws II



Bayesian Inverse Problems

Inverse Problem

Given spaces $\mathcal{U}$ and $\mathcal{Y}$, and a known forward operator $G \colon \mathcal{U} \to \mathcal{Y}$, recover $u \in \mathcal{U}$ from an imperfect observation $y \in \mathcal{Y}$ of $G(u)$.

- A simple example is an inverse problem with additive noise, e.g.

$$y = G(u) + \eta, \tag{2}$$

  where η is a draw from a $\mathcal{Y}$-valued random variable, e.g. a Gaussian $\eta \sim \mathcal{N}(0, \Gamma)$.
- Crucially, we assume knowledge of the probability distribution of η, but not its exact value.

- Inverse problems are typically ill-posed: no solution, multiple solutions, or solutions highly sensitive to the observed data y.
- There is a long tradition dating back to Tikhonov (1963) and others of addressing such problems using regularisation:

$$u \approx u^{\mathrm{MAP}} := \arg\min \underbrace{\frac{1}{2}\left\|\Gamma^{-1/2}(G(\cdot) - y)\right\|_{\mathcal{Y}}^{2}}_{=:\Phi(\cdot;y)} + R(\cdot).$$

- The Bayesian approach (Kaipio and Somersalo, 2005; Stuart, 2010) is to interpret both u and y as random variables, and relations such as (2) as defining the conditional distribution of y given u.
- Beliefs about u independent of y — e.g. about smoothness — are phrased in the form of a prior distribution $\mu_0 \in \mathcal{M}_1(\mathcal{U})$.

Bayesian Inverse Problem (BIP)

Find the posterior distribution $\mu^y \in \mathcal{M}_1(\mathcal{U})$, i.e. the conditional distribution of $u|y$, or summary statistics such as the conditional mean (CM) estimator or maximum a posteriori (MAP) estimator.

- $\Phi\colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ denotes the misfit or negative log-likelihood:

$$\mathbb{P}[y \in E|u] = \int_E \exp(-\Phi(u;y))\,\mathrm{d}\varrho(y) \Big/ \int_{\mathcal{Y}} \exp(-\Phi(u;y))\,\mathrm{d}\varrho(y)\ ,$$

  where ϱ is some σ-finite reference measure on $\mathcal{Y}$.
- In this setting, the generalised Bayes formula for μ^y is

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

$$Z(y) = \mathbb{E}_{u\sim\mu_0}\big[\exp(-\Phi(u;y))\big].$$

Well-Posedness of the BIP à la Hadamard

μ^y should be well-defined for each $y \in \mathcal{Y}$, i.e. $0 < Z(y) < \infty$, and μ^y should change continuously under

- perturbation of the observed data y to some $y' \in \mathcal{Y}$; and
- perturbation of the likelihood, e.g. approximation of Φ by Φ_N.

- In computational practice we *always* work with $\Phi_N : \mathcal{U}_N \times \mathcal{Y}_N \to \mathbb{R}$ defined on finite-dimensional subspaces $\mathcal{U}_N \subset \mathcal{U}$, $\mathcal{Y}_N \subset \mathcal{Y}$.
- It is not enough to study the discrete problem for fixed N.
- One cautionary example was provided by Lassas and Siltanen (2004): you attempt N-pixel reconstruction of a noisily observed piecewise smooth image, and you allow for edges in the reconstruction $u^{(N)}$ by using the total variation prior $\mu_0(u^{(N)}) \propto \exp\left(-\alpha_N \|u^{(N)}\|_{\mathrm{TV}}\right)$.

Example: Lassas and Siltanen (2004)

$$\mu^y(u^{(N)}) \propto \exp\left(-\frac{1}{2\sigma^2} \| G u^{(N)} - y \|_{\mathcal{Y}}^2 - \alpha_N \sum_{j=1}^{N-1} \left| u_{j+1}^{(N)} - u_j^{(N)} \right| \right)$$

- In one non-trivial scaling of the TV norm prior ($\alpha_N \sim 1$), the MAP estimators converge in bounded variation, but the TV priors diverge, and so do the CM estimators.
- In another non-trivial scaling ($\alpha_N \sim \sqrt{N}$), the posterior converges to a Gaussian random variable, so the CM estimator is not edge-preserving, and the MAP estimator converges to zero.
- In all other scalings, the TV prior distributions diverge and the MAP estimators either diverge or converge to a useless limit.

Stable Distributions on $\mathbb{R}$ and Quasi-Banach Spaces

Stable Distributions

- The family of stable distributions has been studied extensively in the statistical and probabilistic literature.
- A random variable u is stable of order $\alpha \in (0, 2]$ if

$$\sum_{i=1}^{n} u_i \stackrel{\mathrm{d}}{=} n^{1/\alpha} u + d \quad \text{for some } d \in \mathbb{R},$$

  and is strictly stable if this holds with $d = 0$. In terms of the law μ of u and the rescaling $\mu_n(E) := \mu(n^{1/\alpha} E)$,

$$\mu = \underbrace{(\mu_n \star \cdots \star \mu_n)}_{n\text{-fold convolution}}(E + d) \quad \text{for all Borel-measurable } E.$$

- Stability is an appealing property if the aim is to construct Bayesian priors that are 'physically consistent' in the sense of remaining in the same model class regardless of discretisation or coordinate choices, at least when the 'physical quantity' obeys an additive law.

Example

- Suppose that the aim is to model (and later infer, in a Bayesian fashion) the distribution of electrical charge in some domain $\Omega \subseteq \mathbb{R}^3$.
- For computation, Ω is approximated by a triangulation $\mathcal{T}$.
- Consider two elements $T_1, T_2 \in \mathcal{T}$. If $\mathrm{charge}(T_i)$ is stably distributed, then so too is
$$\mathrm{charge}(T_1 \cup T_2) = \mathrm{charge}(T_1) + \mathrm{charge}(T_2).$$
- The charge density $\mathrm{charge}(T_i)/\,\mathrm{volume}(T_i)$ behaves similarly.
- Thus, we remain in the same stable model class if we coarsen or refine the mesh $\mathcal{T}$; this would not be true for an unstable random model of the charge, and this would undesirably complicate computational modelling.

- Real-valued stable random variables are classified by four parameters: u is stably distributed with index of stability $\alpha \in (0, 2]$, skewness $\beta \in [-1, 1]$, scale $\gamma \geq 0$, and location $\delta \in \mathbb{R}$, denoted $u \sim \mathcal{S}(\alpha, \beta, \gamma, \delta; 0)$, if its characteristic function $\mathbb{E}\big[\exp(itu)\big]$ is

$$\begin{cases} \exp\big(i\delta t - \gamma^{\alpha}|t|^{\alpha}[1 + i\beta(\tan\frac{\pi\alpha}{2})(\operatorname{sgn} t)(|\gamma t|^{1-\alpha} - 1)]\big) & \text{if } \alpha \neq 1, \\ \exp\big(i\delta t - \gamma|t|[1 + i\beta\frac{2}{\pi}(\operatorname{sgn} t)\log\gamma|t|]\big) & \text{if } \alpha = 1. \end{cases}$$

  (The convention here is that $0 \log 0 := \lim_{s \searrow 0} s \log s = 0$.)

- If $\gamma = 1$ and $\delta = 0$, then u is standardised and we write $u \sim \mathcal{S}(\alpha, \beta; 0)$.
- A stable random variable u has a smooth Lebesgue density, but exact formulae for this density are not available except in special cases:

$$\mathcal{N}(m, \sigma^2) = \mathcal{S}(2, 0, \sigma/\sqrt{2}, m; 0) \qquad \mathcal{C}(\delta, \gamma) = \mathcal{S}(1, 0, \gamma, \delta; 0).$$

Stable Distributions: Definition and Properties ii

Theorem (Nolan, 2015, Proposition 1.16)

If $u \sim \mathcal{S}(\alpha, \beta, \gamma, \delta; 0)$, then, for $a \neq 0$ and $b \in \mathbb{R}$,

$$au + b \sim \mathcal{S}(\alpha, (\operatorname{sgn} a)\beta, |a|\gamma, a\delta + b; 0).$$

Also, if $u_i \sim \mathcal{S}(\alpha, \beta_i, \gamma_i, \delta_i; 0)$ and $u_2 \sim \mathcal{S}(\alpha, \beta_2, \gamma_2, \delta_2; 0)$ are independent, then $u_1 + u_2 \sim \mathcal{S}(\alpha, \beta, \gamma, \delta; 0)$ with

$$\begin{aligned}
\beta &:= \frac{\beta_1\gamma_1^\alpha + \beta_2\gamma_2^\alpha}{\gamma_1^\alpha + \gamma_2^\alpha}, \\
\gamma^\alpha &:= \gamma_1^\alpha + \gamma_2^\alpha, \\
\delta &:= \begin{cases} \delta_1 + \delta_2 + (\tan \frac{\pi\alpha}{2})(\beta\gamma - \beta_1\gamma_1 - \beta_2\gamma_2), & \text{if } \alpha \neq 1, \\ \delta_1 + \delta_2 + \frac{2}{\pi}(\beta\gamma \log\gamma - \beta_1\gamma_1 \log\gamma_1 - \beta_2\gamma_2 \log\gamma_2), & \text{if } \alpha = 1. \end{cases}
\end{aligned}$$

Stable Distributions: Definition and Properties III

- The stable distributions with $\alpha = 2$ are exactly the Gaussian measures: by Fernique's theorem, Gaussian measures are exponentially integrable, and in particular have polynomial moments of all orders.
- Conversely, for $\alpha \in (0, 2)$, the stable distributions are all heavy-tailed: when $u \sim \mathcal{S}(\alpha, \beta, \gamma, 0; 0)$ with $0 < \alpha < 2$,

$$\mathbb{E}\big[|u|^p\big] = \begin{cases} C_{\alpha,\beta}\gamma^\alpha < \infty, & \text{for } 0 < p < \alpha, \\ \infty, & \text{for } p \geq \alpha. \end{cases}$$

- Asymptotic power laws for probability density function and tail probabilities (Nolan, 2015, Theorem 1.12):

$$\mathbb{P}[u > x] \sim c_\alpha \gamma^\alpha (1 + \beta) x^{-\alpha}, \quad \rho_u(x) \sim c_\alpha \alpha \gamma^\alpha (1 + \beta) x^{-(\alpha+1)} \quad \text{as } x \to \infty.$$

  Similar expressions hold for the behaviour as $x \to -\infty$.

- Now consider the problem of constructing and sampling heavy-tailed stable probability measures on a real quasi-Banach space $\mathcal{U}$:

$$\|u + v\|_{\mathcal{U}} \leq K\big(\|u\|_{\mathcal{U}} + \|v\|_{\mathcal{U}}\big).$$

  E.g. vector spaces ℓ^p, $0 < p \leq \infty$, of summable sequences or a Sobolev space of fields of specified smoothness.
- Supposing that one already has access to a generator of real-valued stable random variables (Chambers et al., 1976), it is natural to try to define a $\mathcal{U}$-valued stable random variable via a random series

$$u := \sum_{n \in \mathbb{N}} u_n \psi_n, \tag{3}$$

  where the ψ_n are a basis for $\mathcal{U}$ and the u_n are $\mathbb{R}$-valued stable random variables.
- The natural question is, when does (3) define a bona fide $\mathcal{U}$-valued random variable?

Defining $\mathcal{U}$-Valued Stable Random Variables

- We assume $\mathcal{U}$ to be a real quasi-Banach space with countable, unconditional, normalised, Schauder basis $(\psi_n)_{n\in\mathbb{N}}$.
- The basis $(\psi_n)_{n\in\mathbb{N}}$ and $q > 0$ are such that the synthesis operator $S_\psi \colon \ell^q \to \mathcal{U}$ defined by $S_\phi \colon (v_n)_{n\in\mathbb{N}} \mapsto \sum_{n\in\mathbb{N}} v_n\psi_n$ is a continuous embedding, i.e.

$$\left\| \sum_{n\in\mathbb{N}} v_n\psi_n \right\|_{\mathcal{U}} \leq C\|\underline{v}\|_{\ell^q}. \tag{4}$$

- If $\mathcal{U}$ is Banach, this assumption holds with $q = 1$ for any basis.
- If inequality (4) can be reversed, possibly with a different constant, then the basis $(\psi_n)_{n\in\mathbb{N}}$ is known as a q-frame for $\mathcal{U}$ (Christensen and Stoeva, 2003).
- The case $q = 2$ is the well-known notion of a Riesz basis.

Almost-Sure Convergence Theorem

Theorem (Well-definedness of $\mathcal{U}$-valued stable RVs)

Let $u := \sum_{n \in \mathbb{N}} u_n \psi_n$ have independent stable coefficients $u_n \sim \mathcal{S}(\alpha, \beta_n, \gamma_n, \delta_n; 0)$, where $\alpha \in (0, 2)$, $\underline{\beta} \subset (-1, 1)$, $\underline{\gamma} \in \ell^\alpha$, $\underline{\delta} \in \ell^q$ and, in addition,

$$[\underline{\gamma}]_{\ell^\alpha \log \ell} := \sum_{n \in \mathbb{N}} \big| \gamma_n^\alpha \log |\gamma_n| \big| < \infty, \qquad \text{if } \alpha = q \text{ or } 2q. \tag{5}$$

Then the series defining u converges in $\mathcal{U}$ a.s., and hence defines a $\mathcal{U}$-valued random variable u.

The proof revolves around Kolmogorov's three series theorem: it suffices to establish summability w.r.t. $n \in \mathbb{N}$, for some $A > 0$, of

$$\begin{aligned} &\mathbb{P}\big[|\gamma_n \hat{u}_n| > A\big] \\ \text{and } &\mathbb{E}\big[|\gamma_n \hat{u}_n|^p \mathbb{1}[|\gamma_n \hat{u}_n| < A]\big] \qquad \text{for } p = 1, 2 \end{aligned}$$

for standardised independent $\hat{u}_n \sim \mathcal{S}(\alpha, \beta_n; 0)$.

Example ($\mathcal{U}$-valued Cauchy random variables)

- If $\mathcal{U}$ is Banach (so $q = 1$) and $u_n \sim \mathcal{C}(0, \gamma_n)$, then

$$\mathbb{P}\big[|\gamma_n u_n| \geq A\big] = 1 - \frac{2}{\pi} \arctan \frac{A}{\gamma_n},$$

$$\mathbb{E}\big[|\gamma_n u_n| \mathbb{1}[|\gamma_n u_n| < A]\big] = \frac{\gamma_n}{\pi} \log\left(1 + \frac{A^2}{\gamma_n^2}\right),$$

$$\mathbb{E}\big[|\gamma_n u_n|^2 \mathbb{1}[|\gamma_n u_n| < A]\big] = \frac{2A\gamma_n}{\pi} + \frac{2\gamma_n^2}{\pi} \arctan \frac{A}{\gamma_n}.$$

- Consistent with Theorem 3, these three series all converge if $[\underline{\gamma}]_{\ell \log \ell}$ is finite, and in particular if $\gamma_n = O(n^{-p})$ for some $p > 1$.
- When this convergence holds, the random series defining u converges a.s. in $\mathcal{U}$, and thereby defines a $\mathcal{U}$-valued Cauchy random variable.

Caution

- Condition (5), requiring that the Orlicz norm-like quantity $[\underline{\gamma}]_{\ell \log \ell}$ be finite, cannot be weakened to just requiring that $\underline{\gamma} \in \ell^1$.
- For example, for $\gamma_n := n^{-1}(\log n)^{-2}$, the integral test reveals that $\sum_{n \geq 2} |\gamma_n| < \infty$ but $\sum_{n \geq 2} |\gamma_n \log \gamma_n| = \infty$; in this situation, summability of the truncated first absolute moments of the coefficients $\gamma_n \hat{u}_n$ is no longer assured.
- However, for polynomial $\underline{\gamma}$, the ℓ^1 and Orlicz criteria *do* coincide: for $\gamma_n = Cn^{-p}$, $\|\underline{\gamma}\|_{\ell^1}$ is finite once $p > 1$, and then $\|\underline{\gamma}\|_{\ell \log \ell}$ is also finite.

Theorem (p^{th}-mean convergence and fractional lower-order moments)

Let $u \sim \mathcal{S}(\alpha, \underline{\beta}, \underline{\gamma}, \underline{\delta}; 0)$ satisfy the assumptions of Theorem 3, and suppose that $(\psi_n)_{n\in\mathbb{N}}$ satisfies (4) for some $q > 0$. Let $0 < p \leq q$ and $p < \alpha$. Then $\sum_{n=1}^{N} u_n \psi_n \to u$ in $L^p(\Omega, \mathbb{P}; \mathcal{U})$ as $N \to \infty$ and, in particular,

$$\|u\|^p_{L^p(\Omega,\mathbb{P};\mathcal{U})} \equiv \mathbb{E}\big[\|u\|^p_{\mathcal{U}}\big] \leq C\|\underline{\gamma}\|_{\ell^\alpha} + C\|\underline{\delta}\|_{\ell^q} < \infty. \tag{6}$$

Example: Wavelet Bases

An α-stable $\mathcal{U}$-valued random variable in a (Riesz) wavelet basis has finite p^{th} moments of all orders $0 < p < \alpha$.

Well-Posedness of BIPs

Assumptions for Well-Posedness I

$\mathcal{U}$ and $\mathcal{Y}$ are real separable quasi-Banach spaces and the misfit function $\Phi\colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ satisfies the following:

(A0) Φ is a locally bounded Carathéodory function, i.e. $\Phi(u; \cdot)$ is continuous for each $u \in \mathcal{U}$, $\Phi(\cdot; y)$ is measurable for each $y \in \mathcal{Y}$, and for every $r > 0$, there exists $M_{0,r} \in \mathbb{R}$ such that, for all $(u, y) \in \mathcal{U} \times \mathcal{Y}$ with $\|u\|_{\mathcal{U}} < r$ and $\|y\|_{\mathcal{Y}} < r$,

$$|\Phi(u; y)| \leq M_{0,r}.$$

(A1) For every $r > 0$, there exists a measurable $M_{1,r}\colon \mathbb{R}_+ \to \mathbb{R}$ such that, for all $(u, y) \in \mathcal{U} \times \mathcal{Y}$ with $\|y\|_{\mathcal{Y}} < r$ and $\|u\|_{\mathcal{U}}$ large enough,

$$\Phi(u; y) \geq M_{1,r}\big(\|u\|_{\mathcal{U}}\big).$$

Assumptions for Well-Posedness II

(A2) For every $r > 0$, there exists a measurable $M_{2,r} \colon \mathbb{R}_+ \to \mathbb{R}_+$ such that, for all $(u, y_1, y_2) \in \mathcal{U} \times \mathcal{Y} \times \mathcal{Y}$ with $\|y_1\|_{\mathcal{Y}} < r$, $\|y_2\|_{\mathcal{Y}} < r$,

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp\big(M_{2,r}(\|u\|_{\mathcal{U}})\big)\|y_1 - y_2\|_{\mathcal{Y}}.$$

Furthermore, for each $N \in \mathbb{N}$, $\Phi_N : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ is an approximation to Φ that satisfies (A0)–(A2) with $M_{i,r}$ independent of N, and such that

(A3) $\Psi \colon \mathbb{N} \to \mathbb{R}_+$ is such that, for every $r > 0$, there exists a measurable $M_{3,r} \colon \mathbb{R}_+ \to \mathbb{R}_+$, such that, for all $(u, y) \in \mathcal{U} \times \mathcal{Y}$ with $\|y\|_{\mathcal{Y}} < r$,

$$|\Phi_N(u; y) - \Phi(u; y)| \leq \exp\big(M_{3,r}(\|u\|_{\mathcal{U}})\big)\Psi(N).$$

Under these assumptions, following the same proof strategies as in Stuart (2010), Dashti and Stuart (2015)...

Theorem (Well-definedness of the Bayesian posterior)

Let μ_0 be a Borel probability measure on $\mathcal{U}$, and let $y \in \mathcal{Y}$. If (A0) *and* (A1) *hold with*

$$S_{1,r} := \mathbb{E}_{u \sim \mu_0}\left[\exp(-M_{1,r}(\|u\|_{\mathcal{U}}))\right] < \infty, \tag{7}$$

then $Z(y) := \mathbb{E}_{u \sim \mu_0}\left[\exp(-\Phi(u; y))\right]$ is strictly positive and finite, and setting

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

defines a Borel probability measure μ^y on $\mathcal{U}$, which is tight in the sense that, for all measurable $E \subseteq \mathcal{U}$,

$$\mu^y(E) = \sup\{\mu^y(K) \mid K \subseteq E \text{ and } K \subset \mathcal{U} \text{ is compact}\}.$$

Theorem (Perturbation of observed data)

Suppose that $r > 0$ is such that (A0)–(A2) hold with

$$S_{1,2,r} := \mathbb{E}_{u \sim \mu_0}\left[\exp(2M_{2,r}(\|u\|_{\mathcal{U}}) - M_{1,r}(\|u\|_{\mathcal{U}}))\right] < \infty. \tag{8}$$

Then there exists a constant C, which may depend on r, $S_{1,2,r}$, and the constants and functions in (A0)–(A2), such that, whenever $\|y\|_{\mathcal{Y}}, \|y'\|_{\mathcal{Y}} < r$,

$$|Z(y) - Z(y')| \leq C\|y - y'\|_{\mathcal{Y}}$$
$$\text{and} \qquad d_{\mathrm{H}}\left(\mu^y, \mu^{y'}\right) \leq C\|y - y'\|_{\mathcal{Y}}.$$

Hellinger Distance

- d_{H} denotes the Hellinger metric on $\mathcal{M}_1(\mathcal{U})$, defined by

$$d_{\mathrm{H}}(\mu,\nu)^2 = \int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(u)} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(u)} \right|^2 \, \mathrm{d}\lambda(u),$$

  where λ is any σ-finite Borel measure on $\mathcal{U}$ with respect to which both μ and ν are absolutely continuous, e.g. $\lambda := \mu + \nu$.
- The Hellinger topology coincides with the total variation topology (Kraft); it is strictly weaker than the Kullback–Leibler / relative entropy topology (Pinkser); all these topologies are strictly stronger than the weak convergence topology.
- Expected values of square-integrable functions are Lipschitz continuous with respect to d_{H}:

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

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

Well-Posedness of the BIP III

Theorem (Perturbation of likelihood)

Let Φ and Φ_N satisfy (A0)–(A3), *and suppose that, for some $r > 0$,*

$$S_{1,3,r} := \mathbb{E}_{u\sim\mu_0}\big[\exp(2M_{3,r}(\|u\|_{\mathcal{U}}) - M_{1,r}(\|u\|_{\mathcal{U}}))\big] < \infty. \tag{9}$$

Then there exists a constant C, which may depend on r, $S_{1,3,r}$, and the constants and functions in (A0)–(A3) *but is independent of N, such that the posteriors μ^y and μ^y_N, arrived at using the same data y with $\|y\|_{\mathcal{Y}} < r$ but the misfit functions Φ and Φ_N respectively, satisfy*

$$d_{\mathrm{H}}\big(\mu^y, \mu^y_N\big) \leq C\Psi(N).$$

The moral here is that the convergence rate of the forward problem (from numerical analysis) transfers to the BIP.

Application to Heavy-Tailed Priors

- It is interesting to note the range of applicability of Theorems 6, 7, and 8 when the prior μ_0 is the probability law of a $\mathcal{U}$-valued α-stable random variable.
- Morally, a BIP with stable prior has a well-defined posterior provided that $\Phi(\cdot; y)$ diverges to $-\infty$ no faster than logarithmically (cf. no faster than polynomially for Gaussian and Besov priors).
- Similarly, (8) is satisfied and gives well-posedness with respect to y if $2M_{2,r}(t) - M_{1,r}(t) \not\to +\infty$ faster than logarithmically, and (9) is satisfied and gives well-posedness with respect to Φ_N if $2M_{3,r}(t) - M_{1,r}(t) \not\to +\infty$ faster than logarithmically.

Closing Remarks

Closing Remarks

- The BIP framework of Stuart (2010) can be extended to allow for infinite-dimensional analogues of stable distributions as priors, sampled using an analogue of the classical Karhunen–Loève expansion, and with an analogous well-posedness theory for the posterior. Complete linear spaces with a weakened triangle inequality are also ok.
- Details at arXiv:1605.05898.
- Next steps:
  - Well-posedness in stronger probability metrics, e.g. Kullback–Leibler?
  - Posterior consistency: does μ^y concentrate on $u^\dagger$ in the limit of infinitely many or infinitely precise observations? If so, at what rate?
  - Connection with applications, e.g. non-smooth image reconstruction à la Markkanen et al. (2016).
  - Point estimators (e.g. MAP estimators) for BIPs with heavy-tailed priors, à la Dashti et al. (2013) and Helin and Burger (2015).

References

J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *J. Amer. Statist. Assoc.*, 71(354):340–344, 1976.

O. Christensen and D. T. Stoeva. p-frames in separable Banach spaces. *Adv. Comput. Math.*, 18(2-4):117–126, 2003. doi:10.1023/A:1021364413257.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems, 2015. arXiv:1302.6989v4.

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.

M. Dashti, K. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Probl.*, 29(9):095017, 27, 2013. doi:10.1088/0266-5611/29/9/095017.

T. Helin and M. Burger. Maximum *a posteriori* probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Probl.*, 31(8):085009, 22, 2015. doi:10.1088/0266-5611/31/8/085009.

B. Hosseini and N. Nigam. Well-posed Bayesian inverse problems: Priors with exponential tails, 2016. arXiv:1604.02575v2.

J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005. doi:10.1007/b138659.

M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563, 2004. doi:10.1088/0266-5611/20/5/013.

M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009. doi:10.3934/ipi.2009.3.87.

M. Markkanen, L. Roininen, J. M. J. Huttunen, and S. Lasanen. Cauchy difference priors for edge-preserving Bayesian inversion with an application to X-ray tomography, 2016. arXiv:1603.06135v1.

J. P. Nolan. *Stable Distributions — Models for Heavy Tailed Data*. Birkhauser, Boston, 2015. In progress, Chapter 1 online at http://academic2.american.edu/ jpnolan.

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

A. N. Tikhonov. On the solution of incorrectly put problems and the regularisation method. In *Outlines Joint Sympos. Partial Differential Equations (Novosibirsk, 1963)*, pages 261–265. Acad. Sci. USSR Siberian Branch, Moscow, 1963.