Mathematical Analysis of Statistical Inference

and Why it Matters

Tim Sullivan[1]

SFB 805 Colloquium

TU Darmstadt, DE

31 May 2017

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

Introduction

- Inverse problems — the recovery of parameters in a mathematical model that 'best match' some observations — are ubiquitous in applied mathematics.
- The Bayesian probabilistic perspective on blending models and data is arguably a great success story of late 20$^{\text{th}}$–early 21$^{\text{st}}$ Century mathematics — e.g. numerical weather prediction.
- This perspective on inverse problems has attracted much *mathematical* attention in recent years (Kaipio and Somersalo, 2005; Stuart, 2010). It has led to algorithmic improvements as well as theoretical understanding — and highlighted serious open questions.

- Particular mathematical attention has fallen on Bayesian inverse problems (BIPs) in which the parameter to be inferred lies in an infinite-dimensional space $\mathcal{U}$, e.g. 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 $\mathcal{U}^{(n)}$, but it is profitable to delay discretisation to the last possible moment.
- 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.

A Little History

Prehistory

Bayes (1763) and Laplace (1812, 1814) lay the foundations of inverse probability.

Late 1980s–Early 1990s

Physicists notice that well-posed inferences degenerate in high finite dimension — resolution (in)dependence.

Late 1990s–Early 2000s

The Finnish school, e.g. Lassas and Siltanen (2004), formulate discretisation invariance of finite-dimensional problems.

Since 2010

Stuart (2010) advocates direct study of BIPs on function spaces.

Forward Problem

Given spaces $\mathcal{U}$ and $\mathcal{Y}$, a forward operator $G\colon \mathcal{U} \to \mathcal{Y}$, and $u \in \mathcal{U}$, find $y := G(u)$.

Inverse Problem

Given spaces $\mathcal{U}$ and $\mathcal{Y}$, a forward operator $G\colon \mathcal{U} \to \mathcal{Y}$, and $y \in \mathcal{Y}$, find $u \in \mathcal{U}$ such that $G(u) = y$.

- The distinction between forward and inverse problems is somewhat subjective, since many 'forward' problems involve inversion, e.g. of a square matrix, or a differential operator, etc.
- In practice, the forward model G is only an approximation to reality, and the observed data is imperfect or corrupted.

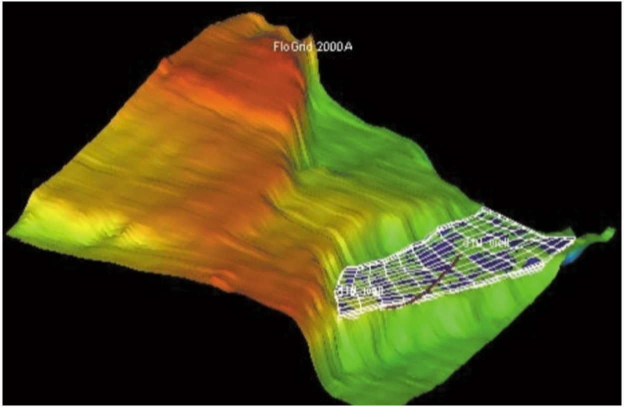
Inverse Problem (revised)

Given spaces $\mathcal{U}$ and $\mathcal{Y}$, and a 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.

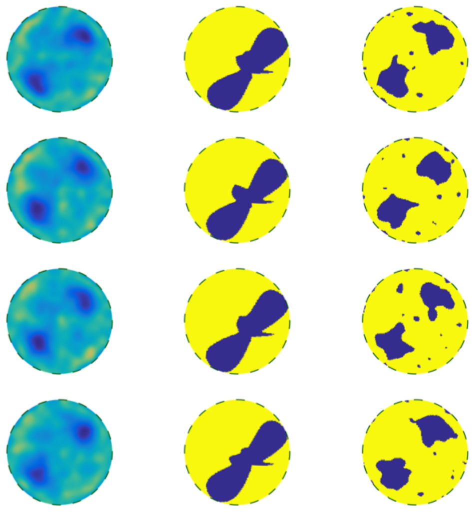
$$y = G(u) + \eta,$$

  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.



Lan et al. (2016)

Darcy flow inverse problem: recover u such that $-\nabla \cdot (u \nabla p) = f$ in $D \subset \mathbb{R}^3$, plus boundary conditions, from pointwise measurements of p and f.



Dunlop and Stuart (2016a)

Electrical impedance tomography: recover conductivity field σ

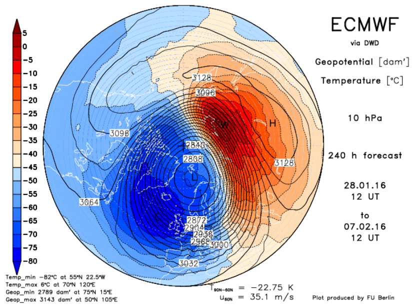
$$
\begin{aligned}
-\nabla \cdot (\sigma \nabla v) &= 0 && \text{in } D \\
\sigma \frac{\partial v}{\partial n} &= 0 && \text{on } \partial D \setminus \bigcup_j e_j
\end{aligned}
$$

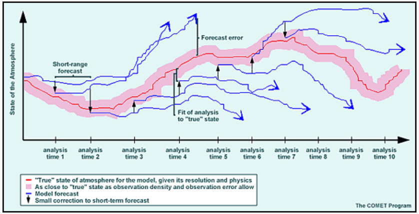
from voltage and current on boundary electrodes e_j:

$$
\begin{aligned}
\int_{e_j} \sigma \frac{\partial v}{\partial n} &= I_j \\
v + z_j \sigma \frac{\partial v}{\partial n} &= V_j && \text{on } e_j.
\end{aligned}
$$

Numerical weather prediction:

- recover pressure, temperature, humidity, velocity fields etc. from meteorological observations
- reconcile them with numerical solution of the Navier–Stokes equations etc.
- predict into the future
- within a tight computational and time budget





ECMWF / DWD / NOAA / NASA

Bayesian Inverse Problems

- Prototypical ill-posed problem from linear algebra: given $A \in \mathbb{R}^{m \times n}$ and $y \in \mathcal{Y} = \mathbb{R}^m$, find $u \in \mathcal{U} = \mathbb{R}^n$ such that

$$y = Au.$$

- More often: recover u from $y := Au + \eta$.
- If η is centred with covariance matrix $\Gamma \in \mathbb{R}^{m \times m}$, then the Gauss–Markov theorem says that (in the sense of minimum variance, and minimum expected squared error) the best estimator of u minimises the weighted misfit $\Phi \colon \mathbb{R}^n \to \mathbb{R}$,

$$\Phi(u) := \frac{1}{2}\|Au - y\|_\Gamma^2 = \frac{1}{2}\left\|\Gamma^{-1/2}(Au - y)\right\|_2^2.$$

- A least-squares solution will exist, but may be non-unique and depend very sensitively upon y through ill-conditioning of $A^{*}\Gamma^{-1}A$.
- The classical way of simultaneously enforcing uniqueness, stabilising the problem, and encoding prior beliefs about what a 'good guess' for u is to regularise the problem:

$$\text{minimise } \Phi(u) + R(u)$$

  - classical Tikhonov (1963) regularisation: $R(u) = \frac{1}{2}\|u\|_2^2$
  - weighted Tikhonov / ridge regression: $R(u) = \frac{1}{2}\left\|C_0^{-1/2}(u - u_0)\right\|_2^2$
  - LASSO: $R(u) = \frac{1}{2}\left\|C_0^{-1/2}(u - u_0)\right\|_1$
- Bayesian probabilistic interpretation: regularisations encode priors μ_0 for u with Lebesgue densities $\frac{\mathrm{d}\mu_0}{\mathrm{d}\Lambda}(u) \propto \exp(-R(u))$.

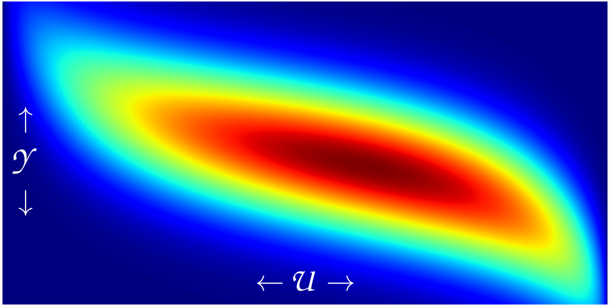
- The variational approach does not easily select among multiple minimisers. Why prefer one with small Hessian to one with large Hessian?
- In the Bayesian formulation of the inverse problem (BIP) (Kaipio and Somersalo, 2005; Stuart, 2010):
  - u is a $\mathcal{U}$-valued random variable, initially distributed according to a prior probability distribution μ_0 on $\mathcal{U}$;
  - the forward map G and the structure of the observational noise determine a probability distribution for $y|u$;
  - and the solution is the posterior distribution μ^y of $u|y$.
- 'Lifting' the inverse problem to a BIP resolves several well-posedness issues, but also raises new challenges in the definition, analysis, and access of μ^y.

Bayesian Modelling Setup

Prior measure μ_0 on $\mathcal{U}$:

Prior measure μ_0 on $\mathcal{U}$:

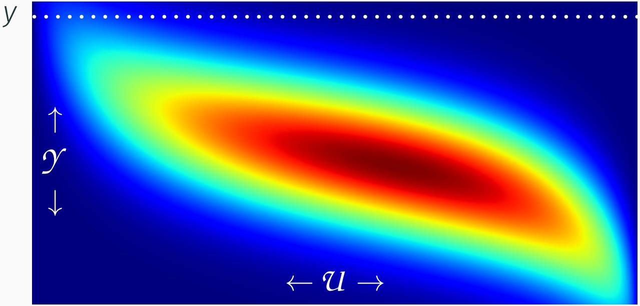
Joint measure μ on $\mathcal{U} \times \mathcal{Y}$:



Bayesian Modelling Setup

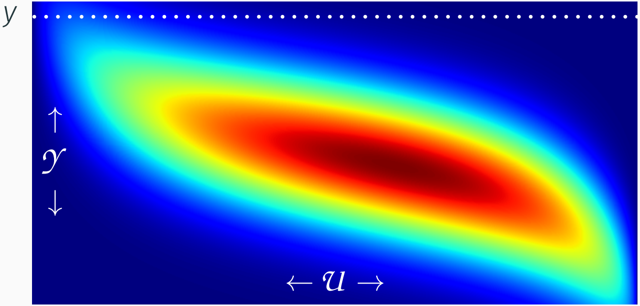
Prior measure μ_0 on $\mathcal{U}$:

Joint measure μ on $\mathcal{U} \times \mathcal{Y}$:



Prior measure μ_0 on $\mathcal{U}$:

Joint measure μ on $\mathcal{U} \times \mathcal{Y}$:



Posterior measure $\mu^y := \mu_0(\cdot|y) \propto \mu|_{\mathcal{U}\times\{y\}}$ on $\mathcal{U}$:

Theorem (Bayes's rule in discrete form)

If u and y assume only finitely many values with probabilities $0 \leq p(u) \leq 1$ etc., then

$$p(u|y) = \frac{p(y|u)p(u)}{\sum_{u'=1}^{n} p(y|u')p(u')}.$$

$$p(u|y) = \frac{p(y|u)p(u)}{\sum_{u'=1}^{n} p(y|u')p(u')}.$$

Theorem (Bayes's rule with Lebesgue densities)

If u and y have positive joint Lebesgue density $\rho(u, y)$ on finite- dimensional space $\mathcal{U} \times \mathcal{Y}$, then

$$\rho(u|y) = \frac{\rho(y|u)\rho(u)}{\int_{\mathcal{U}} \rho(y|u')\rho(u')\,\mathrm{d}u'}.$$

$$p(u|y) = \frac{p(y|u)p(u)}{\sum_{u'=1}^{n} p(y|u')p(u')}$$

$$\rho(u|y) = \frac{\rho(y|u)\rho(u)}{\int_{\mathcal{U}} \rho(y|u')\rho(u')\,\mathrm{d}u'}$$

Theorem (General Bayes's rule: Stuart (2010))

If G is continuous, $\rho(y)$ has full support, and $\mu_0(\mathcal{U}) = 1$, then the posterior μ^y is well-defined, given in terms of its probability density (Radon–Nikodym derivative) with respect to the prior μ_0 by

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

where the misfit potential $\Phi(u;y)$ differs from $-\log\rho(y - G(u))$ by an additive function of y alone.

- Subjective belief: classically, μ_0 really is a *belief* about u before seeing y.
- Regularity: in PDE inverse problems, μ_0 describes the smoothness of u, e.g. on $D \subset \mathbb{R}^d$, $\mathcal{N}(0, (-\Delta)^{-s})$ with $s > d/2$ charges $C(\bar{D}; \mathbb{R})$ with mass 1.
- Physical prediction: in data assimilation / numerical weather prediction (Reich and Cotter, 2015; Law et al., 2015), the prior is a propagation of past analysed states.

Questions to Ask

- Existence and uniqueness: does the posterior really exist as a probability measure μ^y on $\mathcal{U}$?
- Well-posedness: does the posterior μ^y depend in a 'nice' way upon the problem setup, e.g. errors or approximations in the observed data y, the potential Φ, the prior μ_0?
- Consistency: in the limit as the observational errors go to zero (or number of observations $\to \infty$), does the posterior concentrate all its mass on u, at least modulo non-injectivity of G?
- Computation: can μ^y be efficiently sampled (MCMC etc.), summarised or approximated (posterior mean and variance, MAP points, etc.)?

Why be Nonparametric? Why the Function Spaces?

- In computational practice, BIPs have to be discretised: we seek an approximate solution in a finite-dimensional $\mathcal{U}^{(n)} \subset \mathcal{U}$.
- Unfortunately, it is *not* enough to study the just the finite-dimensional problem.
- Analogy with numerical analysis of PDE: discretised wave equation is controllable and has no finite speed of light.
- A cautionary example was provided by Lassas and Siltanen (2004): you attempt n-pixel reconstruction of a piecewise smooth image, from linear observations with additive $\mathcal{N}(0, \sigma^2 I)$ noise, and allow for edges in the reconstruction $u^{(n)} \in \mathcal{U}^{(n)} \cong \mathbb{R}^n$ by using the discrete total variation prior

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}\Lambda}(u^{(n)}) \propto \exp\bigl(-\alpha_n \|u^{(n)}\|_{\mathrm{TV}}\bigr).$$

Total Variation Priors: Lassas and Siltanen (2004)

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\Lambda}(u^{(n)}) \propto \exp\left(-\frac{1}{2\sigma^2}\|Gu^{(n)} - y\|_{\mathcal{Y}}^2 - \alpha_n \sum_{j=1}^{n-1} |u_{j+1}^{(n)} - u_j^{(n)}|\right)$$

- 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.

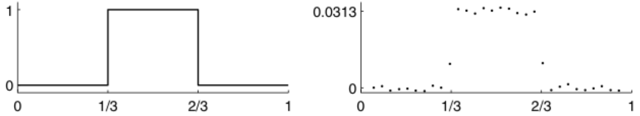


Figure 2. Left: simulated intensity distribution $u(t)$. Right: simulated noisy measurement $\hat{m}$. The dots are plotted at centre points of pixels.

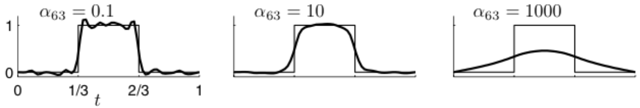


Figure 3. Gaussian MAP estimates with three different choices of α_{63}. The function $u(t)$ is plotted with a thin line. Left: too small α_{63}. Middle: satisfactory solution. Right: too large α_{63}.

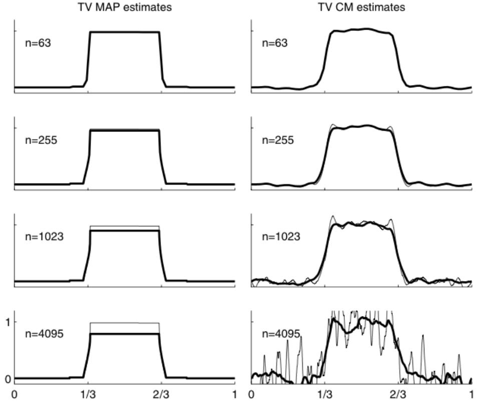


Figure 4. In all the plots in this figure, the coordinate axis limits are the same to allow easy comparison. Left column: MAP estimates for the TV prior with parameter $\alpha_n = 135$ (thin line) and $\alpha_n = 16.875\sqrt{n+1}$ (thick line). Right column: CM estimates for the TV prior with parameter $\alpha_n = 135$ (thin line) and $\alpha_n = 16.875\sqrt{n+1}$ (thick line).

Lassas and Siltanen (2004)

Why be Nonparametric? Why the Function Spaces?

- The infinite-dimensional point of view is also very useful for constructing dimensionally-robust sampling schemes, e.g. the Crank–Nicolson proposal of Cotter et al. (2013) and its variants.
- The definition and analysis of MAP estimators (points of maximum μ^y-probability) in the absence of Lebesgue measure is also mathematically challenging, but yields yields similar fruits for robust finite-dimensional computation (Dashti et al., 2013; Helin and Burger, 2015; Dunlop and Stuart, 2016b).

Well-Posedness of BIPs

Ill- and Well-Posedness of Inverse Problems

- Recall that inverse problems are typically ill-posed: there is no $u \in \mathcal{U}$ such that $G(u) = y$, or there are multiple such u, or it/they depend very sensitively upon the observed data y.
- Regularisation typically enforces existence; the extent to which it enforces uniqueness and robustness depends is problem- dependent.
- One advantage of the Bayesian approach is that the solution is a probability measure: the posterior μ^y can be shown to be exist, be unique, and stable under perturbation.
- Stability in what sense...?

The following well-posedness results work for, among others,

- Gaussian priors (Stuart, 2010)
  - Gaussian prior, linear forward model, quadratic misfit $\implies$ Gaussian posterior, via simple linear algebra (Schur complements).
- Besov priors (Lassas et al., 2009; Dashti et al., 2012).
- Stable priors (Sullivan, 2016) and infinitely-divisible heavy-tailed priors (Hosseini, 2016).
- Hierarchical priors (Agapiou et al., 2014; Dunlop et al., 2016): *careful* adaptation of parameters in the above.

The Hellinger distance between probability measures μ and ν on $\mathcal{U}$ is

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

where r is any measure with respect to which both μ and ν are absolutely continuous, e.g. $r = \mu + \nu$.

Lemma (Hellinger controls second moments)

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

when $f \in L^2(\mathcal{U},\mu) \cap L^2(\mathcal{U},\nu)$ and, in particular,

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

Other Distances on Probability Measures

- The Lévy–Prokhorov distance metrises (for separable $\mathcal{U}$) the topology of weak convergence:

$$d_{\mathrm{LP}}(\mu, \nu) := \inf\left\{\varepsilon > 0 \middle| \forall A \in \mathcal{B}(\mathcal{U}), \mu(A) \leq \nu(A^{\varepsilon}) + \varepsilon \;\&\; \nu(A) \leq \mu(A^{\varepsilon}) + \varepsilon\right\}.$$

$$d_{\mathrm{LP}}(\mu_n, \mu) \to 0 \iff \textstyle\int_{\mathcal{U}} f \, \mathrm{d}\mu_n \to \int_{\mathcal{U}} f \, \mathrm{d}\mu \text{ for all bounded continuous } f$$

- The Hellinger metric is topologically equivalent to the total variation metric:

$$d_{\mathrm{TV}}(\mu, \nu) := \frac{1}{2} \int_{\mathcal{U}} \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu = \sup_{A \in \mathcal{B}(\mathcal{U})} \bigl| \mu(A) - \nu(A) \bigr|.$$

- The relative entropy distance or Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(\mu \| \nu) := \int_{\mathcal{U}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\nu.$$

Theorem (Stuart (2010); Sullivan (2016); Dashti and Stuart (2017))

Suppose that Φ is locally bounded, Carathéodory (i.e. continuous in y and measurable in u), bounded below by $\Phi(u; y) \geq M_1(\|u\|_{\mathcal{U}})$ with $\exp(-M_1(\|\cdot\|_{\mathcal{U}})) \in L^1(\mathcal{U}, \mu_0)$. Then

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

$$Z(y) := \int_{\mathcal{U}} \exp(-\Phi(u; y))\, \mathrm{d}\mu_0(u) \in (0, \infty)$$

does indeed define a Borel probability measure on $\mathcal{U}$, which is Radon if μ_0 is Radon, and μ^y really is the posterior distribution of $u \sim \mu_0$ conditioned upon the data y.

- The spaces $\mathcal{U}$ and $\mathcal{Y}$ should be separable (but see Hosseini (2016)), complete, and normed (but see Sullivan (2016)).
- For simplicity, the above theorem and its sequels are actually slight misstatements: more precise formulations would be local to data y with $\|y\|_{\mathcal{Y}} \leq r$.
- The lower bound on Φ is allowed to tend to $-\infty$ as $\|u\|_{\mathcal{U}} \to \infty$ or $\|y\|_{\mathcal{Y}} \to \infty$. Indeed this is essential if $\dim \mathcal{Y} = \infty$ or if one is preparing for the limit $\dim \mathcal{Y} \to \infty$.
- For Gaussian priors this means that quadratic blowup is allowed (Stuart, 2010); for α-stable priors only logarithmic blowup is allowed (Sullivan, 2016).

Theorem (Stuart (2010); Sullivan (2016); Dashti and Stuart (2017))

Suppose that Φ satisfies the previous assumptions and

$$\left|\Phi(u;y) - \Phi(u;y')\right| \leq \exp(M_2(\|u\|_{\mathcal{U}}))\left\|y - y'\right\|_{\mathcal{Y}}$$

with $\exp(2M_2(\|\cdot\|_{\mathcal{U}}) - M_1(\|\cdot\|_{\mathcal{U}})) \in L^1(\mathcal{U}, \mu_0)$*. Then there exists* $C \geq 0$ *such that*

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

Moral

(Local) Lipschitz dependence of the potential Φ upon the data transfers to the BIP.

Theorem (Stuart (2010); Sullivan (2016); Dashti and Stuart (2017))

Suppose that Φ and Φ_n satisfy the previous assumptions uniformly in n and

$$\left|\Phi_n(u;y) - \Phi(u;y)\right| \leq \exp(M_3(\|u\|_{\mathcal{U}}))\psi(n)$$

with $\exp(2M_3(\|\cdot\|_{\mathcal{U}}) - M_1(\|\cdot\|_{\mathcal{U}})) \in L^1(\mathcal{U}, \mu_0)$. *Then there exists $C \geq 0$ such that*

$$d_{\mathrm{H}}\left(\mu^y, \mu^{y,n}\right) \leq C\psi(n).$$

Moral

The convergence rate of the forward problem / potential, as expressed by $\psi(n)$, transfers to the BIP.

More Consequences of Approximation

- Even the numerical posterior $\mu^{y,n}$ will have to be approximated, e.g. by MCMC sampling, an ensemble approximation, or a Gaussian fit.
- The bias and variance in this approximation should be kept at the same order as the error $\mu^{y,n} - \mu^{y}$.
- It is mathematical analysis that allows the correct tradeoffs among these sources of error/uncertainty, and hence the appropriate allocation of resources.

Brittleness of BIPs

Well-Posedness w.r.t. Prior?

- As seen above, under reasonable assumptions, BIPs are well-posed in the sense that the posterior μ^y is Lipschitz in the Hellinger metric with respect to the data y and uniform changes to the likelihood Φ.
- What about simultaneous perturbations of the prior μ_0 and likelihood model, i.e. the full Bayesian model μ on $\mathcal{U} \times \mathcal{Y}$?
- When the model is well-specified and $\dim \mathcal{U} < \infty$, we have the Bernstein–von Mises theorem: μ^y concentrates on 'the truth' as number of samples $\to \infty$ or data noise $\to 0$.
- Freedman (1963, 1965) showed that this can fail when $\dim \mathcal{U} = \infty$; even limiting inferences can be model-dependent!
- The situation is especially bad if the model is misspecified, i.e. doesn't cover 'the truth'.

$$d_{\mathrm{H}}(\mu^y, \tilde{\mu}^y) \leq C \cdot d_{\mathrm{H}}(\mu, \tilde{\mu})?$$

Brittleness under Misspecification

$$d_{\mathrm{H}}(\mu^y, \cancel{\tilde{\mu}^y) \leq C \cdot} d_{\mathrm{H}}(\mu, \tilde{\mu})?$$

Theorem (Owhadi et al. (2015a,b))

For any misspecified model μ *on 'general' spaces* $\mathcal{U}$ *and* $\mathcal{Y}$, *for any* $Q \colon \mathcal{U} \to \mathbb{R}$, *any any* $\operatorname{ess\,inf}_{\mu_0} Q < q < \operatorname{ess\,sup}_{\mu_0} Q$, *there is another model* $\tilde{\mu}$ *as close as you like to* μ *so that*

$$\mathbb{E}_{\tilde{\mu}^y}[Q] \approx q$$

for all sufficiently finely observed data y.

Moral

Closeness is measured in the Hellinger, total variation, Lévy–Prokhorov, or common-moments topologies. BIPs are ill-posed in these topologies, because small changes to the model give you any posterior value you want, by slightly (de)emphasising particular parts of the data.

Should we worry about this?

- Maybe... The explicit examples in Owhadi et al. (2015a,b) have a very slow convergence rate. Approximate priors arising from numerical inversion of differential operators appear to have the 'wrong' kind of error bounds.
- Yes! Koskela et al. (2017), Kennedy et al. (2017), and Kurakin et al. (2017) observe brittleness-like phenomena 'in the wild'.
- No! Just stay away from high-precision data, e.g. by coarsening à la Miller and Dunson (2015), or be more careful with the geometry of credible/confidence regions (Castillo and Nickl, 2013, 2014), e.g. adaptive confidence regions (Szabó et al., 2015a,b).

Closing Remarks

Conclusions and Open Problems

- Mathematical analysis reveals the well- and ill-posedness of Bayesian inference procedures, which lie at the heart of many modern applications in physical and now social sciences.
- This quantitative analysis allows tradeoff of errors and resources.
- Open topic: fundamental limits on the robustness and consistency of general procedures.
- Interpreting (forward) numerical tasks as Bayesian inference tasks leads to Bayesian probabilistic numerical methods for linear algebra, quadrature, optimisation, ODEs and PDEs (Hennig et al., 2015; Cockayne et al., 2017).

S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart. Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):511–544, 2014. doi:10.1137/130944229.

T. Bayes. *An Essay towards solving a Problem in the Doctrine of Chances*. 1763.

I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4): 1999–2028, 2013. doi:10.1214/13-AOS1133.

I. Castillo and R. Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5): 1941–1969, 2014. doi:10.1214/14-AOS1246.

J. Cockayne, C. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods, 2017. arXiv::1702.03673.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 2013. doi:10.1214/13-STS421.

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*. Springer, 2017. arXiv:1302.6989.

M. Dashti, S. Harris, and A. 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. J. H. 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.

M. M. Dunlop and A. M. Stuart. The Bayesian formulation of EIT: analysis and algorithms. *Inverse Probl. Imaging*, 10(4): 1007–1036, 2016a. doi:10.3934/ipi.2016030.

M. M. Dunlop and A. M. Stuart. MAP estimators for piecewise continuous inversion. *Inverse Probl.*, 32:105003, 2016b. doi:10.1088/0266-5611/32/10/105003.

M. M. Dunlop, M. A. Iglesias, and A. M. Stuart. Hierarchical Bayesian level set inversion. *Stat. Comput.*, 2016. doi:10.1007/s11222-016-9704-8.

D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963. doi:10.1214/aoms/1177703871.

D. A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.*, 36:454–456, 1965. doi:10.1214/aoms/1177700155.

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.

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, 2016. arXiv:1609.07532.

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.

L. A. Kennedy, D. J. Navarro, A. Perfors, and N. Briggs. Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis. *Behav. Res. Meth.*, pages 1–16, 2017. doi:10.3758/s13428-017-0854-1.

J. Koskela, P. A. Jenkins, and D. Spanò. Bayesian non-parametric inference for Λ-coalescents: posterior consistency and a parametric method. *Bernoulli*, 2017. To appear, arXiv:1512.00982.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world, 2017. arXiv:1607.02533.

S. Lan, T. Bui-Thanh, M. Christie, and M. Girolami. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. *J. Comput. Phys.*, 308:81–101, 2016. doi:10.1016/j.jcp.2015.12.032.

P.-S. Laplace. *Théorie analytique des probabilités*. 1812.

P.-S. Laplace. *Essai philosophique sur les probabilités*. 1814.

M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.*, 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.

K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*, volume 62 of *Texts in Applied Mathematics*. Springer, 2015. doi:10.1007/978-3-319-20325-6.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening, 2015. arXiv:1506.06101.

H. Owhadi, C. Scovel, and T. J. Sullivan. On the brittleness of Bayesian inference. *SIAM Rev.*, 57(4):566–582, 2015a. doi:10.1137/130938633.

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, 2015b. doi:10.1214/15-EJS989.

S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, New York, 2015. doi:10.1017/CBO9781107706804.

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

T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors, 2016. arXiv:1605.05898.

B. Szabó, A. W. van der Vaart, and J. H. van Zanten. Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plan. Inference*, 166:36–51, 2015a. doi:10.1016/j.jspi.2014.06.005.

B. Szabó, A. W. van der Vaart, and J. H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428, 2015b. doi:10.1214/14-AOS1270.

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.