Posterior consistency in Bayesian inference with exponential priors

Masoumeh Dashti

University of Sussex

Workshop on *Optimization and Inversion under Uncertainty*

Linz, 12 November 2019

Based on joint work with

S Agapiou (Cyprus), T Helin (LUT, Finland)

The setting

Suppose (indirect) noisy measurements, y, of quantity of interest, u, is available

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

The setting

Suppose (indirect) noisy measurements, y, of quantity of interest, u, is available

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

Examples.

i) $y_j = u(x_j) + \eta, \quad j = 1, \ldots, n, \;\; x_j \in D \subset \mathbb{R}^d$

$u \in C_b(D)$

The setting

Suppose (indirect) noisy measurements, y, of quantity of interest, u, is available

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

Examples.

i) $y_j = u(x_j) + \eta, \quad j = 1, \ldots, n, \;\; x_j \in D \subset \mathbb{R}^d$

$u \in C_b(D)$

ii) $y_j = p(x_j) + \eta, \quad j = 1, \ldots, n \;\; x_j \in D \subset \mathbb{R}^d$

$\nabla \cdot (u \nabla p) = f \;$ in D

$u \in C_b(D)$ with $u > 0$.

Bayesian approach

Consider

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

with $u \in X$, $y \in \mathbb{R}^n$ (X separable Banach spaces),

- prior $u \sim \mu_0$
- statistics of noise is known: $\eta \sim \rho_\eta$

posterior μ^y (*when well-defined**) satisfies

$$\mu^y(\mathrm{d}u) \propto \rho_\eta(y - \mathcal{G}(u))\, \mu_0(\mathrm{d}u)$$

$$\iff \quad \mu^y(A) = \int_A \underbrace{c\, \rho_\eta(y - \mathcal{G}(u))}_{\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u)}\, \mu_0(\mathrm{d}u) \qquad \forall A \in \mathcal{B}(X)$$

Posterior consistency

suppose:

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ with n arbitrarily large
- there exists an underlying truth

$$y = \mathcal{G}(w_0) + \eta$$

Posterior consistency

suppose:

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ with n arbitrarily large
- there exists an underlying truth

$$y = \mathcal{G}(w_0) + \eta$$

Does μ^y concentrate on arbitrarily small neighbourhoods of w_0 as $n \to \infty$ and how fast?

Posterior consistency

suppose:

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ with n arbitrarily large
- there exists an underlying truth

$$y = \mathcal{G}(w_0) + \eta$$

Does μ^y concentrate on arbitrarily small neighbourhoods of w_0 as $n \to \infty$ and how fast?

Simpler: Do modes of μ^y converge to w_0?

Outline

1. MAP estimators and weak posterior consistency
2. Posterior consistency with contraction rates

Outline

1. MAP estimators and weak posterior consistency
2. Posterior consistency with contraction rates

MAP estimates

$\mu(X) = 1$, X a function space

There is no Lebesgue density, modes can be defined topologically:

Any point $\tilde{u} \in X$ satisfying

$$\lim_{\epsilon \to 0} \frac{\sup_{u \in X} \mu(B_\epsilon(u))}{\mu(B_\epsilon(\tilde{u}))} = 1,$$

is a MAP estimator. (MD, LAW, STUART, VOSS '13)

$\exists Z \subset X$ s.t. for $u \in Z$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(u))}{\mu(B_\epsilon(0))} = \mathrm{e}^{-I(u)}$$

- If $X = \mathbb{R}^n$, $Z = \mathbb{R}^n$

  and $I(u) = -\log \rho_\mu(u)$

- For X function space

  Z is a proper dense subset of X with $\mu(Z) = 0$

$\exists\, Z \subset X$ s.t. for $u \in Z$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(u))}{\mu(B_\epsilon(0))} = e^{-I(u)}$$

- If $X = \mathbb{R}^n$, $Z = \mathbb{R}^n$

  and $I(u) = -\log \rho_\mu(u)$

- For X function space

  Z is a proper dense subset of X with $\mu(Z) = 0$

Are modes of μ characterised by minimisers of I?

The Prior

$X \subset L^2$

- $\{\psi_j\}$ orthonormal basis in $L^2(\mathbb{T}^d)$,
- $\xi_j \sim c_p \exp(-\frac{|x|^p}{p})$, $p \geq 1$, i.i.d
- $\{\gamma_j\} \to 0$ positive decreasing sequence

μ_0 law of $(\gamma_j \xi_j)_j$, and $u \sim \mu_0$ satisfies

$$u(x) = \sum_{j \in \mathbb{N}} \gamma_j \, \xi_j \, \psi_j(x)$$

The Prior

$X \subset L^2$

$\{\psi_j\}$ orthonormal basis in $L^2(\mathbb{T}^d)$,

$\xi_j \sim c_p \exp(-\frac{|x|^p}{p})$, $p \geq 1$, i.i.d

$\{\gamma_j\} \to 0$ positive decreasing sequence

μ_0 law of $(\gamma_j \xi_j)_j$, and $u \sim \mu_0$ satisfies

$$u(x) = \sum_{j \in \mathbb{N}} \gamma_j \, \xi_j \, \psi_j(x)$$

Gaussian	Besov (Lassas, Saksman, Siltanen '09)
$p = 2$	$p \geq 1$, γ_j negative powers of j
$\{\psi_j\}$ an orthonormal basis	$\{\psi_j\}$ orthonormal wavelet basis

$p = 1$ sparsity promoting, continuous but not differentiable measure

For $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = c\,\mathrm{e}^{-\Phi(u)}$ with Φ given

$$I(u) = \Phi(u) + \tfrac{1}{p}\|u\|_Z^p,$$

$$Z := \left\{u \in X : \sum_j \left|\tfrac{\langle u,\psi_j\rangle}{\gamma_j}\right|^p < \infty\right\}$$

for $h \in Q := \left\{u \in X : \sum_j \left|\tfrac{\langle u,\psi_j\rangle}{\gamma_j}\right|^2 < \infty\right\}$, $\gamma_j\xi_j \sim \rho_j$

$$\frac{\mathrm{d}\mu_{0,h}}{\mathrm{d}\mu_0}(u) = \lim_{N\to\infty} \prod_{j=1}^{N} \frac{\rho_j(u_i - h_j)}{\rho_j(u_j)}$$

$$= \lim_{N\to\infty} \mathrm{e}^{\sum_{j=1}^{N} -\left|\frac{h_j - u_j}{\gamma_j}\right|^p + \left|\frac{u_j}{\gamma_j}\right|^p} \quad \text{in } L^1_\mu$$

For *locally Lipschitz* Φ, modes of μ are minimisers of I:

- $p = 2$ MD, Law, Stuart, Voss '13 ($Z = Q$)
- $p > 1$ Helin & Burger '15; Lie & Sullivan '18 (differentiable)
- $p = 1$ Agapiou, Burger, MD, Helin '18

Weak posterior consistency

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \rho_\eta(y - \mathcal{G}(u)) =: \mathrm{e}^{-\Phi(u,y)}, \quad \text{suppose:}$$

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ with n arbitrarily large
- there exists an underlying truth, $(y_j = G(w_0) + \eta_j)$

$$y = \mathcal{G}(w_0) + \eta$$

for μ_0 exponential, MAP estimates are

$$u_n := \operatorname*{argmin}_{u \in Z} \Phi(u, y) + \|u\|_Z.$$

$$u_n := \operatorname*{argmin}_{u \in Z} \Phi(u, y) + \|u\|_Z$$

$$= \operatorname*{argmin}_{u \in Z} \left|G(w_0) - G(u)\right|^2 + \frac{2}{n} \sum_{j=1}^{n} \left\langle G(w_0) - G(u), \eta_j \right\rangle + \frac{1}{n} \|u\|_Z$$

Theorem. (AGAPIOU, BURGER, D, HELIN '18)

Assume that

> $G \colon X \to \mathbb{R}_+$ *is locally Lipschitz and*
>
> $w_0 \in Z$.

Then

- $G(u_n) \to G(w_0)$ *in probability.*
- *If* G *is injective* $\|u_n - w_0\|_X \to 0$ *in probability.*

  Otherwise, $\exists u^* \in Z$ *and a subseq of* $\{u_n\}_{n \in \mathbb{N}}$ *such that* $\|u_n - w_0\|_X \to 0$ *in probability. For any such* u^*, $G(u^*) = G(w_0)$.

$$u_n := \operatorname*{argmin}_{u \in Z} \Phi(u, y) + \|u\|_Z$$

$$= \operatorname*{argmin}_{u \in Z} |G(w_0) - G(u)|^2 + \frac{2}{n} \sum_{j=1}^{n} \left\langle G(w_0) - G(u), \eta_j \right\rangle + \frac{1}{n}\|u\|_Z$$

Theorem. (Agapiou, Burger, D, Helin '18)

Assume that

$G\colon X \to \mathbb{R}_+$ *is locally Lipschitz and*

$w_0 \in Z$.

Then

- $G(u_n) \to G(w_0)$ *in probability.*
- *If* G *is injective* $\|u_n - w_0\|_X \to 0$ *in probability.*

  Otherwise, $\exists u^* \in Z$ *and a subseq of* $\{u_n\}_{n\in\mathbb{N}}$ *such that* $\|u_n - w_0\|_X \to 0$ *in probability. For any such* u^*, $G(u^*) = G(w_0)$.

small noise limit similar $y = G(w_0) + \delta_n \eta$

Outline

1. MAP estimators and weak posterior consistency
2. Posterior consistency with contraction rates

Consistency with contraction rates

μ^y is said to contract with rate ϵ_n at w_0 if

$$\mu^y\Big(\{u \in X : \|u - w_0\| \geq C\epsilon_n\}\Big) \to 0 \quad \text{in } \mathbb{P}(y|w_0)\text{-probability}$$

Consistency with contraction rates

μ^y is said to contract with rate ϵ_n at w_0 if

$$\mu^y\Big(\big\{u \in X : \|u - w_0\| \geq C\epsilon_n\big\}\Big) \to 0 \quad \text{in } \mathbb{P}(y|w_0)\text{-probability}$$

GHOSAL, GOSH & VAN DER VAART '00 give sufficient conditions on model and *prior* to ensure this.

Conditions on prior

- μ_0 puts sufficient mass around w_0,
- distribution of mass under μ_0 is '*not too complex*'

model: i.i.d sampling or white noise model

Conditions on prior – Exponential case

AGAPIOU, MD & HELIN '18:

For appropriate ϵ_n* there exists $X_n \subset X$ s.t.

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\tilde{\epsilon}_n, X_n, \|\cdot\|_X) \leq Cn\tilde{\epsilon}_n^2$ (N: min # of balls needed to cover X_n)

$$\mu_0(X \setminus X_n) \leq \mathrm{e}^{-Cn\epsilon_n^2}$$

Conditions on prior – Exponential case

AGAPIOU, MD & HELIN '18:

For appropriate ϵ_n* there exists $X_n \subset X$ s.t.

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\tilde{\epsilon}_n, X_n, \|\cdot\|_X) \leq Cn\tilde{\epsilon}_n^2$ (N: min # of balls needed to cover X_n)

  $\mu_0(X \setminus X_n) \leq \mathrm{e}^{-Cn\epsilon_n^2}$

(∗) ϵ_n satisfies

$$\phi_{w_0}(\epsilon_n) \leq n\epsilon_n^2 \quad \text{with} \quad \phi_w(\epsilon) := \inf_{h \in Z : \|h - w\|_X \leq \epsilon} \frac{1}{p}\|h\|_Z^p - \log \mu_0(\epsilon B_X)$$

(based on VAN DER VAART & VAN ZANTEN '08 for Gaussian)

- for $h \in Z$

$$\mu_0(\epsilon B_X + h) \geq \mathrm{e}^{-\frac{1}{p}\|h\|_Z^p} \mu_0(\epsilon B_X)$$

- By two-level Talagrand's inequality – 1994[1], $\forall M > 0$

$$\mu(A + M^{\frac{p}{2}} B_Q + M B_Z) \geq 1 - \frac{1}{\mu(A)} \exp(-cM^p)$$

$\rightarrow$ choose $X_n = \epsilon B_X + M_n^{\frac{p}{2}} B_Q + M_n B_Z$

with $M_n \propto (n\epsilon_n^2)^{\frac{1}{p}}$

[1] generalised Borell's inequality

Contraction rates

Find the largest ϵ_n s.t. $\phi_{w_0}(\epsilon_n) \le n\epsilon_n^2$

- For White noise model $y_n = \int_0^t u(s)\,\mathrm{d}s + \frac{1}{\sqrt{n}} B_t, \quad t \in [0,1]$

  with truth $w_0 \in B^{\beta}_{qq}$ and prior $B^{\alpha+\frac{1}{p}}_{pp}$-Besov measure

$$c\epsilon_n = \begin{cases} n^{-\frac{\beta}{1+2\beta+p(\alpha-\beta)}}, & \text{if } \beta \le \alpha, \\ n^{-\frac{\alpha}{1+2\alpha}}, & \text{if } \beta > \alpha \end{cases}$$

$$\mu^y\Big(\{u \in X : \|u - w_0\| \ge C\epsilon_n\}\Big) \to 0 \quad \text{in } \mathbb{P}(y|w_0)\text{-probability}$$

Contraction rates

Find the largest ϵ_n s.t. $\phi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$

- For White noise model $y_n = \int_0^t u(s)\,\mathrm{d}s + \frac{1}{\sqrt{n}} B_t, \quad t \in [0,1]$

  with truth $w_0 \in B^{\beta}_{qq}$ and prior $B^{\alpha+\frac{1}{p}}_{pp}$-Besov measure

$$c\epsilon_n = \begin{cases} n^{-\frac{\beta}{1+2\beta+p(\alpha-\beta)}}, & \text{if } \beta \leq \alpha, \\ n^{-\frac{\alpha}{1+2\alpha}}, & \text{if } \beta > \alpha \end{cases}$$

$$\mu^y\Big(\{u \in X : \|u - w_0\| \geq C\epsilon_n\}\Big) \to 0 \quad \text{in } \mathbb{P}(y|w_0)\text{-probability}$$

- Upper bounds on $\mu_0(\epsilon B_X + h)$ enables study of lower bounds on concentration rates:

  $\rightarrow$ work in progress, recently established for $p = 1$

Final remarks

- *Convergence rates for MAPs*

  for Gaussian priors NICKL, VAN DE GEER, WANG '19

- *Posterior contraction* for *nonlinear* forward operator

  VOLLMER '13 – pushforward μ_0 with $\mathcal{G}$: elliptic inverse problem

  NICKL '17 – Bernstein-von Mises theorem: elliptic inverse problem

- *Generalised MAPs* for *discontinuous* priors

  (CLASON, HELIN, KRETSCHMANN, PIIROINEN '19)