MAP estimators and posterior consistency for Bayesian inverse problems for functions

Masoumeh Dashti

University of Sussex

Workshop on Gaussian Processes

Saint-Étienne October 2018

Based on joint works with

S Agapiou (Cyprus), M Burger (Münster), T Helin (Helsinki)

Inverse problems

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

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

Inverse problems

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

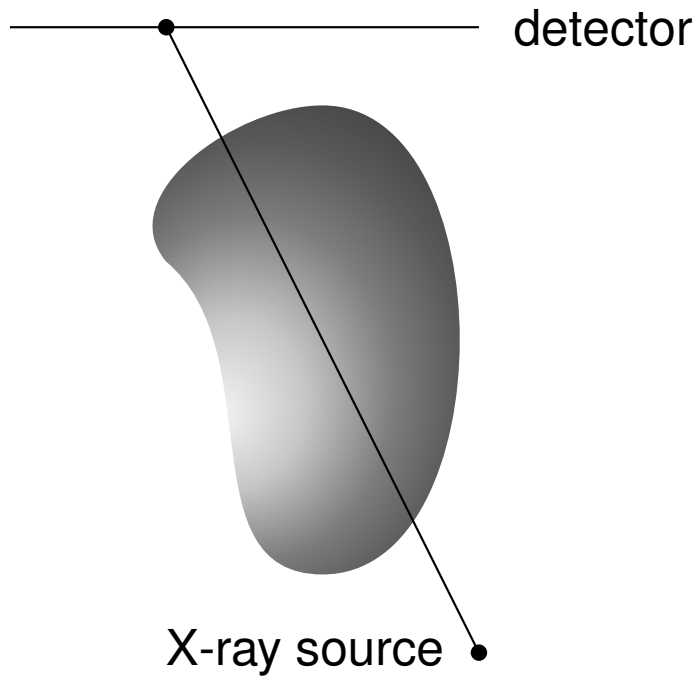
$$y = \mathcal{G}(u) + \eta$$

Example. X-ray imaging

$$y_1 = \int_L c\, u(x)\, \mathrm{d}S(x) + \eta$$

u density, $x \in D \subset \mathbb{R}^d$

η noise



Bayesian approach

Consider

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

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

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

solution: posterior $\text{Prob}(\text{u} \,|\, \text{data}) \propto \text{Prob}(\text{data} \,|\, u)\, \text{Prob}(u)$

Bayesian approach

Consider

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

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

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

solution: posterior $\text{Prob}(\text{u} \,|\, \text{data}) \propto \text{Prob}(\text{data} \,|\, u)\, \text{Prob}(u)$

---

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

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

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

Posterior consistency

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

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

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

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

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

Does μ^y concentrate on arbitrarily small neighbourhoods of $u^\dagger$ as $n \to \infty$?

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

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

Does μ^y concentrate on arbitrarily small neighbourhoods of $u^\dagger$ as $n \to \infty$?

Simpler: Do modes of μ^y converge to $u^\dagger$?

Outline

Outline

MAP estimates

$(X, \|\cdot\|)$ *a separable Banach space;* $\mu(X) = 1$

Let $B^\delta(z)$ be a ball of radius δ and centre z in X.

- Fix δ and find z^δ such that $B^\delta(z^\delta)$ has maximal probability
- Look at the 'limit' of $\{z^\delta\}_\delta$ as δ shrinks to zero

MAP estimates

$(X, \|\cdot\|)$ *a separable Banach space;* $\mu(X) = 1$

Let $B^\delta(z)$ be a ball of radius δ and centre z in X.

- Fix δ and find z^δ such that $B^\delta(z^\delta)$ has maximal probability
- Look at the 'limit' of $\{z^\delta\}_\delta$ as δ shrinks to zero

Definition (D., Law, Stuart, Voss 2013)

Let

$$M^\delta = \sup_{z \in X} \mu(B^\delta(z)).$$

Any point $\tilde{z} \in X$ *satisfying*

$$\lim_{\delta \to 0} \frac{M^\delta}{\mu(B^\delta(\tilde{z}))} = 1,$$

is a mode (MAP estimator) of μ*.*

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(v))}{\mu(B_\epsilon(u))} \stackrel{?}{=} e^{f(u,v)} \qquad (*)$$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(v))}{\mu(B_\epsilon(u))} \stackrel{?}{=} \mathrm{e}^{f(u,v)} \qquad (*)$$

- If $X = \mathbb{R}^K$ and $\mu_0 \sim \mathrm{e}^{-W(u)}$,

  $f(u,v) = I(u) - I(v)$ with $I(u) = \Phi(u) + W(u)$

$$\lim_{\epsilon\to 0} \frac{\mu(B_\epsilon(v))}{\mu(B_\epsilon(u))} \stackrel{?}{=} \mathrm{e}^{f(u,v)} \qquad (*)$$

- If $X = \mathbb{R}^K$ and $\mu_0 \sim \mathrm{e}^{-W(u)}$,

  $f(u,v) = I(u) - I(v)$ with $I(u) = \Phi(u) + W(u)$

- For ∞-d X, usually
  - $\exists\, F \subset X$ s.t. $f(u,v) = I(u) - I(v)$ for $u, v \in F$
  - $\exists\, E \subset F \subset X$

    s.t. $(*)$ makes sense for $u, v \in X$ and $u - v \in E$

$$\lim_{\epsilon\to 0} \frac{\mu(B_\epsilon(v))}{\mu(B_\epsilon(u))} \stackrel{?}{=} \mathrm{e}^{f(u,v)} \qquad (*)$$

- If $X = \mathbb{R}^K$ and $\mu_0 \sim \mathrm{e}^{-W(u)}$,

  $f(u,v) = I(u) - I(v)$ with $I(u) = \Phi(u) + W(u)$

- For ∞-d X, usually
  - $\exists\, F \subset X$ s.t. $f(u,v) = I(u) - I(v)$ for $u, v \in F$
  - $\exists\, E \subset F \subset X$

    s.t. $(*)$ makes sense for $u, v \in X$ and $u - v \in E$

Example: $\mu_0 \sim \mathcal{N}(0, \mathcal{C}_0)$ on a Hilbert space X

$$I(u) = \Phi(u) + \tfrac{1}{2}\|u\|_F^2 \quad \text{with} \quad F = \mathcal{C}_0^{\frac{1}{2}} X \text{ and } \mu(F) = 0$$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(v))}{\mu(B_\epsilon(u))} \stackrel{?}{=} \mathrm{e}^{f(u,v)} \qquad (*)$$

- If $X = \mathbb{R}^K$ and $\mu_0 \sim \mathrm{e}^{-W(u)}$,

  $f(u,v) = I(u) - I(v)$ with $I(u) = \Phi(u) + W(u)$

- For ∞-d X, usually
  - $\exists\, F \subset X$ s.t. $f(u,v) = I(u) - I(v)$ for $u, v \in F$
  - $\exists\, E \subset F \subset X$

    s.t. $(*)$ makes sense for $u, v \in X$ and $u - v \in E$

Example: $\mu_0 \sim \mathcal{N}(0, \mathcal{C}_0)$ on a Hilbert space X

$$I(u) = \Phi(u) + \tfrac{1}{2}\|u\|_F^2 \quad \text{with} \quad F = \mathcal{C}_0^{\frac{1}{2}} X \text{ and } \mu(F) = 0$$

Do minimisers of I characterise MAP estimators?

Gaussian prior

For μ satisfying $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \propto \exp(-\Phi(u))$, on Banach space X,

with $\mu_0 \sim \mathcal{N}(0, \mathcal{C}_0)$

and locally Lipschitz $\Phi : X \to \mathbb{R}_+$

MAP estimators are characterised by the minimisers of I

$$I(u) = \Phi(u) + \frac{1}{2}\|u\|_F^2$$

with F: Cameron-Martin space of μ_0. (D., Law, Stuart, Voss 2013)

Weak MAP estimates

Definition (T. Helin & M. Burger 2015)

Let E be a dense subspace of X. We call a point $\hat{u} \in X$, a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} \leq 1,$$

for all $h \in E$.

Weak MAP estimates

Definition (T. Helin & M. Burger 2015)

Let E be a dense subspace of X. We call a point $\hat{u} \in X$, a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} \leq 1,$$

for all $h \in E$.

T. Helin & M. Burger 2015: *weak MAPs coincide with minimisers of I for smooth enough convex measures*

H.C. Lie & T.J. Sullivan 2018: *for μ nonatomic Borel probability measure on X, weak and strong modes coincide.*

Exponential priors

Define μ_0 through Karhunen-Loève expansion of its draws:

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

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

ξ_j i.i.d random variables,

$\{\gamma_j\}$ decreasing sequence determining smoothness of u

Exponential priors

Define μ_0 through Karhunen-Loève expansion of its draws:

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

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

ξ_j i.i.d random variables,

$\{\gamma_j\}$ decreasing sequence determining smoothness of u

Gaussian	Besov (Lassas, Saksman, Siltanen 09)
$\xi_j \sim c \exp(-\frac{1}{2}\lvert x\rvert^2)$	$\xi_j \sim c_p \exp(-\lvert x\rvert^p), \quad p \geq 1$
$\{\psi_j\}$ an orthonormal basis	$\{\psi_j\}$ orthonormal wavelet basis

Example:

B^s_{11}-Besov prior: draws of μ_0 can be expressed as

$$u(x) = \sum_{j \in \mathbb{N}} \gamma_j \, \xi_j \, \psi_j(x), \qquad \xi_j \sim \frac{1}{2} \mathrm{e}^{-|x|}, \quad \gamma_j = j^{\frac{-s}{d} + \frac{1}{2}}$$

Example:

B^s_{11}-Besov prior: draws of μ_0 can be expressed as

$$u(x) = \sum_{j \in \mathbb{N}} \gamma_j \, \xi_j \, \psi_j(x), \qquad \xi_j \sim \frac{1}{2} e^{-|x|}, \quad \gamma_j = j^{\frac{-s}{d} + \frac{1}{2}}$$

- useful when $u^\dagger$ expected to be smooth with a few local irregularities – promotes sparsity
- $I(u) = \Phi(u) + \|u\|_{B^s_{11}}$

Example:

B^s_{11}-Besov prior: draws of μ_0 can be expressed as

$$u(x) = \sum_{j \in \mathbb{N}} \gamma_j \, \xi_j \, \psi_j(x), \qquad \xi_j \sim \frac{1}{2} \mathrm{e}^{-|x|}, \quad \gamma_j = j^{\frac{-s}{d} + \frac{1}{2}}$$

- useful when $u^\dagger$ expected to be smooth with a few local irregularities – promotes sparsity
- $I(u) = \Phi(u) + \|u\|_{B^s_{11}}$

  not smooth enough for Helin and Burger (2015). But continuity of f_h

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(u+h))}{\mu(B_\epsilon(u))} =: f_h(u) \quad \text{in } X$$

  is enough for equivalence of MAPs and minimisers of I (*Agapiou, Burger, D, Helin 2018*)

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

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

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

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

for μ_0 exponential, MAP estimates are

$$u_n := \operatorname*{argmin}_{u \in F} \Phi(u) + \|u\|_F^p.$$

Theorem. *Assume that*

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

$u^\dagger \in F$.

Then

- $\mathcal{G}(u_n) \to \mathcal{G}(u^\dagger)$ *in probability.*
- *If* $\mathcal{G}$ *is injective* $\|u_n - u^\dagger\|_X \to 0$ *in probability.*

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

Outline

Posterior consistency – general idea

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

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

Posterior consistency – general idea

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

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

This depends on * (*Ghosal & van der Vaart 2007*)

- if μ_0 puts sufficient mass around w_0,
- how μ_0 '*distributes*' the mass in space X

Posterior consistency – general idea

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

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

This depends on * (*Ghosal & van der Vaart 2007*)

- if μ_0 puts sufficient mass around w_0,
- how μ_0 '*distributes*' the mass in space X

(*) *Assuming that data are either*

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

Posterior consistency – Gaussian prior

van der Vaart & van Zanten (2008):

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

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq Cn\epsilon_n^2$

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

Posterior consistency – Gaussian prior

van der Vaart & van Zanten (2008):

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

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq Cn\epsilon_n^2$

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

It then follows that

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

Posterior consistency – Gaussian prior

van der Vaart & van Zanten (2008):

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

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq Cn\epsilon_n^2$

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

It then follows that

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

$(*)$ ϵ_n satisfies

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

- $\phi_w(\epsilon) = \inf_{h \in F : \|h - w\|_X \leq \epsilon} \frac{1}{2} \|h\|_F^2 \; - \; \log \mu_0(\epsilon B_X)$

- $\phi_w(\epsilon) = \inf_{h \in F : \|h - w\|_X \leq \epsilon} \frac{1}{2} \|h\|_F^2 \; - \; \log \mu_0(\epsilon B_X)$

$$\phi_0(\epsilon) = -\log \mu_0(\epsilon B_X) \to \infty \text{ as } \epsilon \to 0$$

- $\phi_w(\epsilon) = \inf_{h \in F : \|h-w\|_X \leq \epsilon} \frac{1}{2}\|h\|_F^2 \; - \; \log \mu_0(\epsilon B_X)$

  $\phi_0(\epsilon) = -\log \mu_0(\epsilon B_X) \to \infty$ as $\epsilon \to 0$

  $\phi_w(\epsilon)$ is a measure of probability of ϵ-balls around w

- $\phi_w(\epsilon) = \inf_{h \in F : \|h - w\|_X \leq \epsilon} \frac{1}{2} \|h\|_F^2 \; - \; \log \mu_0(\epsilon B_X)$

  $\phi_0(\epsilon) = -\log \mu_0(\epsilon B_X) \to \infty$ as $\epsilon \to 0$

  $\phi_w(\epsilon)$ is a measure of probability of ϵ-balls around w

- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq C n \epsilon_n^2$

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

  $$\mu(A + r\,B_F) \geq 1 - \frac{1}{\mu(A)} \exp(-cr^2)$$

  (Borell's inequality)

Posterior consistency – Exponential priors

Agapiou, D & Helin (2018):

For $\epsilon_n > 0$ with

$$\phi_w(\epsilon_n) = \inf_{h \in F : \|h - w_0\|_X \leq \epsilon_n} \frac{1}{2}\|h\|_F^p - \log \mu_0(\epsilon_n B_X) \leq n\epsilon_n^2$$

there exists $X_n \subset X$ s.t.

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq Cg(\epsilon_n)$

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

Posterior consistency – Exponential priors

Agapiou, D & Helin (2018):

For $\epsilon_n > 0$ with

$$\phi_w(\epsilon_n) = \inf_{h \in F : \|h - w_0\|_X \leq \epsilon_n} \frac{1}{2}\|h\|_F^p - \log \mu_0(\epsilon_n B_X) \leq n\epsilon_n^2$$

there exists $X_n \subset X$ s.t.

- $\mu_0(\|u - w_0\|_X < 2\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}$
- $\log N(\epsilon_n, X_n, \|\cdot\|_X) \leq Cg(\epsilon_n)$

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

It then follows that

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

F not a Hilbert space
and proper subset of space of admissible shifts Q

Generalisation of Borell's inequality

$$\mu(A + r^{\frac{p}{2}} B_Q + rB_Z) \geq 1 - \frac{1}{\mu(A)} \exp(-cr^p)$$

(two-level Talagrand's inequality – 1994)

Contraction rates ϵ_n satisfy

$$\phi_w(\epsilon_n) = \inf_{h \in F : \|h - w_0\|_X \leq \epsilon_n} \frac{1}{2}\|h\|_F^p - \log \mu_0(\epsilon_n B_X) \leq n\epsilon_n^2$$

Contraction rates ϵ_n satisfy

$$\phi_w(\epsilon_n) = \inf_{h \in F : \|h - w_0\|_X \leq \epsilon_n} \frac{1}{2}\|h\|_F^p - \log \mu_0(\epsilon_n B_X) \leq n\epsilon_n^2$$

- 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 $w_0 \in B_{qq}^{\beta}$ and $B_{pp}^{\alpha+\frac{1}{p}}$ prior measure

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

Agapiou, D & Helin 2018

Final remarks

- *MAP estimates*
  Generalised MAPs for non-continuous priors
  (*Clason, Helin, Kretschmann, Piiroinen 2018*)

- *Posterior consistency* with contraction rates
  For exp. priors, with distance function L^2 type settled
  Besov norm?
  Upper bound for small ball probabilities?