Bayesian Inversion: Algorithms

Andrew M Stuart[1]

[1] Mathematics Institute and
Centre for Scientific Computing
University of Warwick

Woudshoten Lectures 2013
October 4th 2013

Work funded by EPSRC, ERC and ONR

http://homepages.warwick.ac.uk/∼masdr/

- A.M. Stuart. *Inverse problems: a Bayesian perspective.* Acta Numerica **19**(2010). `∼masdr/BOOKCHAPTERS/stuart15c.pdf`
- M. Dashti, K.J.H. Law, A.M. Stuart and J. Voss. *MAP estimators and posterior consistency* .... Inverse Problems, **29**(2013), 095017. `arxiv:1303.4795`.
- F. Pinski, G. Simpson, A.M. Stuart and H. Weber. *Kullback-Leibler approximation for probability measures on infinite dimensional spaces.* In preparation.
- S.L. Cotter , G.O. Roberts, A.M. Stuart and D. White. *MCMC methods for functions* .... Statistical Science **28**(2013). `arxiv:1202.0709`.
- M. Hairer, A.M.Stuart and S. Vollmer. *Spectral gaps for a Metropolis-Hastings algorithm* .... `arxiv:1112.1392`.

Outline

Outline

The Setting

- Probability measure μ on Hilbert space H.
- Reference measure μ_0 (often a prior).
- μ related to μ_0 by (often Bayes' Theorem)

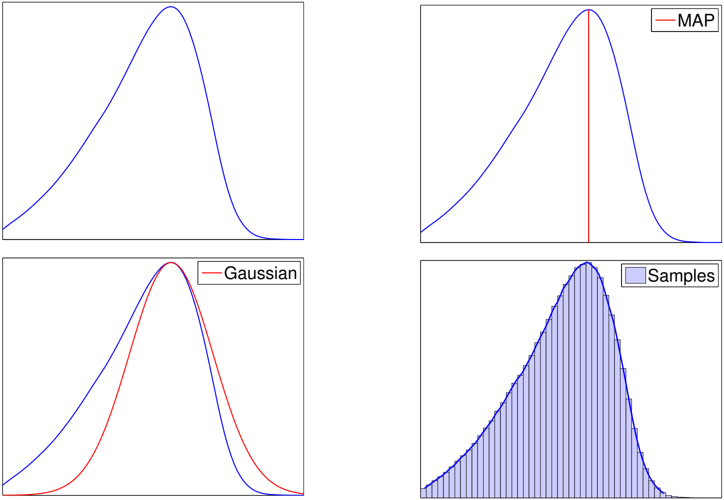
$$\frac{d\mu}{d\mu_0}(u) = \frac{1}{Z_\mu}\exp\Bigl(-\Phi(u)\Bigr).$$

- Another way of saying the same thing:

$$\mathbb{E}^{\mu} f(u) = \frac{1}{Z_\mu}\mathbb{E}^{\mu_0}\Bigl(\exp\bigl(-\Phi(u)\bigr) f(u)\Bigr).$$

- How do we get information from μ if we know μ_0 and Φ?

The Talk In One Picture



The Assumptions

- $\mu_0 = N(0, C_0)$ a centred Gaussian measure on H.
- $\mu_0(X) = 1$; X (Banach) continuously embedded in H.
- Let $E = \mathcal{D}(C_0^{-\frac{1}{2}})$ (Cameron-Martin space).
- Then $E \subset X \subseteq H$. E (Hilbert) compactly embedded in X.
- The function $\Phi \in C(X; \mathbb{R}^+)$.
- For all u, v with $\|u\|_X \leq r$, $\|v\|_X \leq r$ there are $M_i(r)$:

$$|\Phi(u)| \leq M_1(r);$$
$$|\Phi(u) - \Phi(v)| \leq M_2(r)\|u - v\|.$$

Outline

Probability Maximizers and Tikhonov Regularization

Define the Tikhonov-regularized LSQ functional $I : E \to \mathbb{R}^+$ by

$$I(u) := \frac{1}{2}\|C_0^{-\frac{1}{2}}u\|^2 + \Phi(u).$$

Let $B^\delta(z)$ be a ball of radius δ in X centred at $z \in E = \mathcal{D}(C_0^{-\frac{1}{2}})$.

Theorem

(Dashti, Law, S and Voss, 2013). *The probability measure μ and functional I are related by*

$$\lim_{\delta \to 0} \frac{\mu\big(B^\delta(z_1)\big)}{\mu\big(B^\delta(z_2)\big)} = \exp\big(I(z_2) - I(z_1)\big).$$

Thus probability maximizers are minimizers of the regularized Tikhonov functional I.

Existence of Probability Maximizers

The minimization is well-defined:

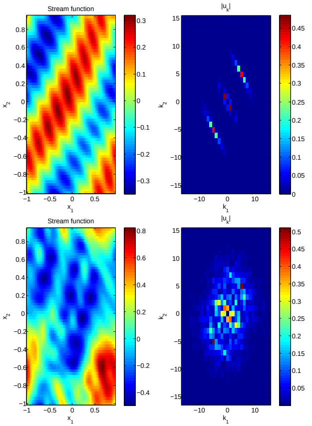
Theorem

(S, Acta Numerica, 2010). $\exists\, \overline{u} \in E$:

$$I(\overline{u}) = \bar{I} := \inf\{I(u) : u \in E\}.$$

Furthermore, if $\{u_n\}$ is a minimizing sequence satisfying $I(u_n) \to \bar{I}$ then there is a subsequence $\{u_{n'}\}$ that converges strongly to $\overline{u}$ in E.

Example: Navier-Stokes Inversion for Initial Condition



- Incompressible NSE on $\Omega_T = \mathbb{T}^2 \times (0, \infty)$:

$$\begin{aligned} \partial_t v - \nu \triangle v + v \cdot \nabla v + \nabla p = f & \quad \text{in } \Omega_T, \\ \nabla \cdot v = 0 & \quad \text{in } \Omega_T, \\ v|_{t=0} = u & \quad \text{in } \mathbb{T}^2. \end{aligned}$$

- $y_{j,k} = v(x_j, t_k) + \eta_{j,k}, \quad \eta_{j,k} \sim N(0, \sigma^2 I_{2\times 2}).$
- $y = \mathcal{G}(u) + \eta, \quad \eta \sim N(0, \sigma^2 I).$
- $C_0 = (-\triangle_{\text{stokes}})^{-2}; \; \Phi = \frac{1}{10^3 \sigma^2} |y - \mathcal{G}(u)|^2.$

Example: Navier-Stokes Inversion for Initial Condition

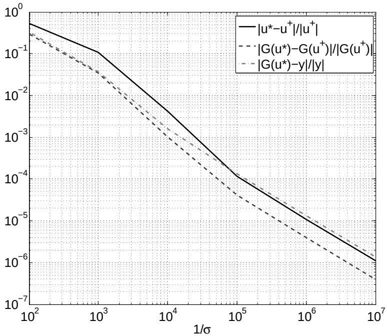


Figure: MAP estimator $u^\star$; Truth $u^\dagger$

Outline

The Objective Functional

Recall $\mu_0 = N(0, C_0)$ and $\mu(du) \propto \exp\big(-\Phi(u)\big)\mu_0(du)$. Let $\mathcal{A}$ denote a set of simple measures on H (usually Gaussian).

Problem

Find $\nu \in \mathcal{A}$ that minimizes $I(\nu) := D_{\mathrm{KL}}(\nu \| \mu)$.

Here D_{KL} =Kullbach-Leibler divergence = relative entropy

$$D_{\mathrm{KL}}(\nu\|\mu) = \begin{cases} \int_{\mathcal{H}} \frac{d\nu}{d\mu}(x) \log\left(\frac{d\nu}{d\mu}(x)\right)\mu(dx) & \text{if } \nu \ll \mu \\ +\infty & \text{else.} \end{cases}$$

We note, for intuition, the inequality:

$$d_{Hell}(\nu, \mu)^2 \leq 2D_{\mathrm{KL}}(\nu\|\mu).$$

Existence of Minimizers

The minimization is well-defined:

Theorem

(Pinski, Simpson, S, Weber, 2013) If $\mathcal{A}$ is closed under weak convergence and there is $\nu \in \mathcal{A}$ with $I(\nu) < \infty$ then $\exists \overline{\nu} \in \mathcal{A}$ such that

$$I(\overline{\nu}) = \overline{I} := \inf\{I(\nu) : \nu \in \mathcal{A}\}.$$

Furthermore, if $\{\nu_n\}$ is a minimizing sequence satisfying $I(\nu_n) \to \overline{I}$ then there is a subsequence $\{\nu_{n'}\}$ that converges to $\overline{\nu}$ in the Hellinger metric:

$$d_{Hell}(\nu_n, \nu) \to 0.$$

Example: $\mathcal{A} := \mathcal{G} = \{\text{Gaussian measures on } \mathcal{H}\}$.

Parameterization of $\mathcal{G}$

Gaussian case equivalent to $\nu = N\Big(m, \big(C_0^{-1} + \Gamma\big)^{-1}\Big)$.

$$J(m,\Gamma) = \begin{cases} D_{\mathrm{KL}}(\nu \| \mu) & \text{if } (m,\Gamma) \in E \times \mathcal{HS}(E, E^*) \\ +\infty & \text{else.} \end{cases}$$

Theorem

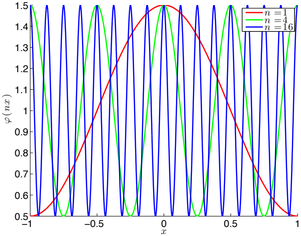
(Pinski, Simpson, S, Weber, 2013) $\exists\, (\overline{m}, \overline{\Gamma}) \in E \times \mathcal{HS}(E, E^*)$:

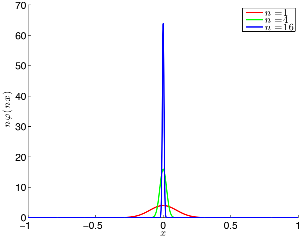
$$J(\overline{m}, \overline{\Gamma}) = \overline{J} := \inf\{J(m,\Gamma) : (m,\Gamma) \in E \times \mathcal{HS}(E, E^*)\}.$$

Furthermore, if $\{(m_n, \Gamma_n)\}$ *is a minimizing sequence satisfying* $J(m_n, \Gamma_n) \to \overline{J}$ *then there is a subsequence* $\{(m_{n'}, \Gamma_{n'})\}$ *such that*

$$\|m_{n'} - \overline{m}\|_E + \|\Gamma_{n'} - \overline{\Gamma}\|_{\mathcal{HS}(E,E^*)} \to 0.$$

Cautionary Examples





- μ_0 is Brownian bridge on $[-1,1]$:
- $\nu_n := N\Big(0, \big(C_0^{-1} + \Gamma_n\big)^{-1}\Big)$.
- Either:

$$(\Gamma_n u)(x) = \varphi(nx)u(x), \varphi \in C^\infty_{\text{per}}, \text{ mean } \overline{\varphi}$$

$$\nu_n \Rightarrow \nu := N\Big(0, \big(C_0^{-1} + \overline{\varphi} Id\big)^{-1}\Big)$$

- Or:

$$(\Gamma_n u)(x) = n\varphi(nx)u(x), \varphi \in C^\infty_0, \|\varphi\|_{L^1} = 1$$

$$\nu_n \Rightarrow \nu := N\Big(0, \big(C_0^{-1} + \delta_0 Id\big)^{-1}\Big).$$

Regularization of J I

Let $(S, \|\cdot\|_S)$ be compact in $\mathcal{L}(E, E^*)$.

$$J_\delta(m,\Gamma) = \begin{cases} J(m,\Gamma) + \delta\|\Gamma\|_S^2 & \text{if } (m,\Gamma) \in E \times S \\ +\infty & \text{else.} \end{cases}$$

Theorem

(Pinski, Simpson, S, Weber, 2013) $\exists\, (\overline{m}, \overline{\Gamma}) \in E \times S$:

$$J_\delta(\overline{m}, \overline{\Gamma}) = \overline{J}_\delta := \inf\{J_\delta(m,\Gamma) : (m,\Gamma) \in E \times S\}.$$

Furthermore, if $\{\nu_n(m_n, \Gamma_n)\}$ *is a minimizing sequence satisfying* $J_\delta(m_n, \Gamma_n) \to \overline{J}_\delta$ *then there is a subsequence* $\{(m_{n'}, \Gamma_{n'})\}$ *such that*

$$d_{Hell}(\nu_{n'}, \nu) + \|\Gamma_{n'} - \overline{\Gamma}\|_S \to 0.$$

Regularization of J II

- Let $H = L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$, $C_0 = (-\triangle)^{-\alpha}$ with $\alpha > d/2$.
- Choose $(\Gamma u)(x) = B(x)u(x)$ and $S = H^r$, $r > 0$.
- Thus $\nu = N(m, C)$, $\quad C^{-1} = C_0^{-1} + B(\cdot)Id$ for potential B.

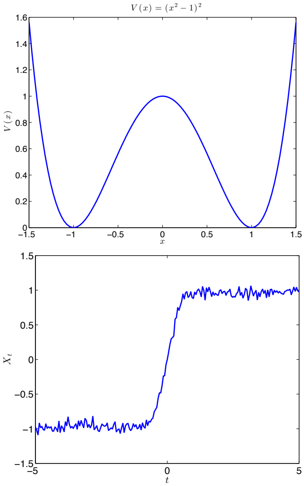
$$J_\delta(m, B) = \begin{cases} J(m, B) + \delta\|B\|_{H^r}^2 & \text{if } (m, B) \in H^\alpha \times H^r \\ +\infty & \text{else.} \end{cases}$$

Theorem

(Pinski, Simpson, S, Weber, 2013) $\exists\, (\overline{m}, \overline{B}) \in H^\alpha \times H^r$:

$$J_\delta(\overline{m}, \overline{B}) = \overline{J}_\delta := \inf\{J_\delta(m, B) : (m, B) \in H^\alpha \times H^r\}.$$

Example: Conditioned Diffusion in a Double Well



- Consider the conditioned diffusion

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\epsilon}dW_t s, .$$
$$X_{-T} = x_- < 0, X_{+T} = x_+ > 0.$$

- For $x_- = -x_+$, by symmetry, we can study paths satisfying $X_0 = 0$, $X_T = x_+$
- Path space distribution approximated by $N(m(t), C)$, with

$$C^{-1} = \frac{1}{2\epsilon}\left(-\frac{d^2}{dt^2} + BI\right)$$

- B is either a constant or $B = B(t)$
- m and B obtained by minimizing D_{KL}

Stochastic Root Finding & Optimization

Robbins-Monro with Derivatives and Iterate Averaging

Functions Estimated Via Sampling

Assume $f(x)$ (the target function) can be estimated via $F(y; x)$ as

$$f(x) = \mathbb{E}^Y[F(Y; x)], \quad f'(x) = \mathbb{E}^Y[\partial_x F(Y; x)]$$

Iteration Scheme (See, for instance, Asmussen & Glynn)

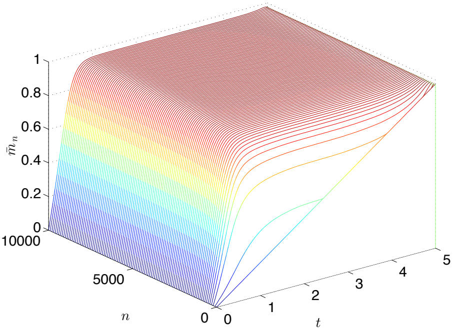
$$x_{n+1} = x_n - a_n \Big(\frac{1}{M} \sum_i^M \partial_x F(Y_i; x_n) \Big)^{-1} \Big(\frac{1}{M} \sum_i^M F(Y_i; x_n) \Big),$$

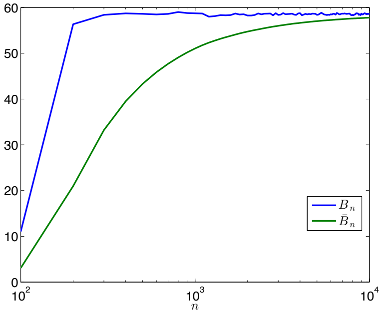
with $a_n \sim n^{-\gamma}, \gamma \in (\frac{1}{2}, 1)$ and Y_i i.i.d. Also let $\bar{x}_n \equiv \sum_{j=1}^n x_i$. Then $\bar{x}_n \to x_\star$, with $f(x_\star) = 0$, in distribution at rate $n^{-1/2}$.

Numerical Results With Constant Potential B

$T = 5$, $\epsilon = .15$,

10^4 Iterations, 10^3 Samples per Iteration, 10^2 Points in $(0, T)$ per Sample

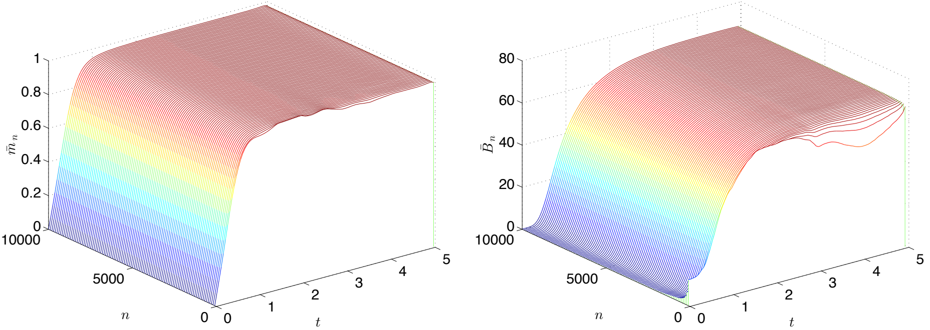




Numerical Results With Variable Potential B

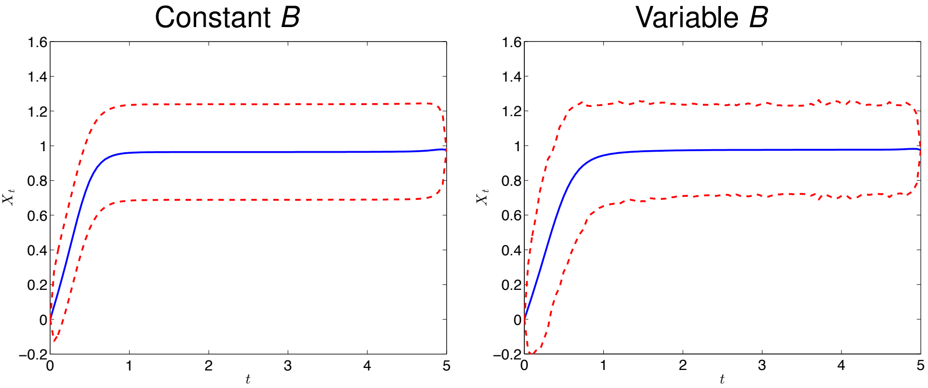
$T = 5$, $\epsilon = .15$, $\delta = \frac{1}{2} \times 10^{-4}$, $r = 1$ so H^1 regularization

10^4 Iterations, 10^4 Samples per Iteration, 10^2 Points in $(0, T)$ per Sample



Model Comparison

95% Confidence Intervals about the Mean Path



Outline

MCMC

- MCMC: create an ergodic Markov chain $u^{(k)}$ which is invariant for approximate target μ (or μ^N the approximation on $\mathbb{R}^N$) so that

$$\frac{1}{K}\sum_{k=1}^{K} f(u^{(k)}) \to \mathbb{E}^{\mu} f$$

.

- Recall $\mu_0 = N(0, C_0)$ and $\mu(du) \propto \exp\big(-\Phi(u)\big)\mu_0(du)$.
- Recall the Tikhonov functional $I(u) = \frac{1}{2}\|C_0^{-\frac{1}{2}}u\|^2 + \Phi(u)$.

Standard Random Walk Algorithm

Metropolis, Rosenbluth, Teller and Teller, J. Chem. Phys. 1953.

- Set $k = 0$ and Pick $u^{(0)}$.
- Propose $v^{(k)} = u^{(k)} + \beta\xi^{(k)}, \quad \xi^{(k)} \sim N(0, C_0)$.
- Set $u^{(k+1)} = v^{(k)}$ with proability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \to k + 1$.

Here $a(u, v) = \min\{1, \exp\big(I(u) - I(v)\big)\}$.

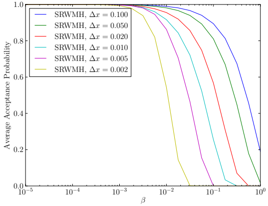
New Random Walk Algorithm

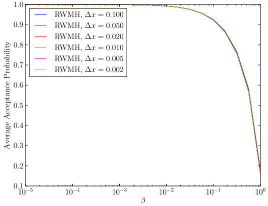
Cotter, Roberts, S and White, Stat. Sci. 2013.

- Set $k = 0$ and Pick $u^{(0)}$.
- Propose $v^{(k)} = \sqrt{(1-\beta^2)}u^{(k)} + \beta\xi^{(k)}, \quad \xi^{(k)} \sim N(0, C_0)$.
- Set $u^{(k+1)} = v^{(k)}$ with proability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \to k+1$.

Here $a(u, v) = \min\{1, \exp\big(\Phi(u) - \Phi(v)\big)\}$.

Example: Navier-Stokes Inversion for Forcing





- Incompressible NSE on $\Omega_T = \mathbb{T}^2 \times (0, \infty)$:

$$
\begin{aligned}
\partial_t v - \nu \triangle v + v \cdot \nabla v + \nabla p &= u && \text{in } \Omega_T, \\
\nabla \cdot v &= 0 && \text{in } \Omega_T, \\
v|_{t=0} &= v_0 && \text{in } \mathbb{T}^2.
\end{aligned}
$$

- $y_{j,k} = v(x_j, t_k) + \xi_{j,k}, \quad \xi_{j,k} \sim N(0, \sigma^2 I_{2\times 2})$.
- $y = \mathcal{G}(u) + \xi, \quad \xi \sim N(0, \sigma^2 I)$.
- Prior OU process; $\Phi = \frac{1}{\sigma^2}|y - \mathcal{G}(u)|^2$.

Spectral Gaps

Theorem

*(*Hairer, S, Vollmer, arXiv 2012.*)*

- *For the standard Random walk algorithm the spectral gap is bounded above by* $C\,N^{-\frac{1}{2}}$.
- *For the new Random walk algorithm the spectral gap is bounded below independently of dimension.*

Outline

What We Have Shown

We have shown that:

- **Common Structure:** A range of problems require extracting information from a probability measure on a Hilbert space, having density with respect to a Gaussian.
- **Algorithmic Approaches** We have laid the foundations of a range of computational methods related to this task.
- **MAP Estimators** Maximum a posteriori estimators can be defined on Hilbert space; there is a link to Tikhonov regularization.
- **Kullback-Leibler Approximation** Kullback-Leibler approximation can be defined on Hilbert space and finding the closest Gaussian results in a well-defined problem in the calculus of variations.
- **Sampling** MCMC methods can be defined on Hilbert space. Results in new algorithms robust to discretization.

http://homepages.warwick.ac.uk/∼masdr/

- A.M. Stuart. *Inverse problems: a Bayesian perspective.* Acta Numerica **19**(2010).
- M. Dashti, K.J.H. Law, A.M. Stuart and J. Voss. *MAP estimators and posterior consistency* . . . . Inverse Problems, **29**(2013), 095017. `arxiv:1303.4795`.
- F. Pinski, G. Simpson, A.M. Stuart and H. Weber. *Kullback-Leibler approximation for probability measures on infinite dimensional spaces.* In preparation.
- S.L. Cotter , G.O. Roberts, A.M. Stuart and D. White. *MCMC methods for Functions* . . . . Statistical Science **28**(2013). `arxiv:1202.0709`.
- M. Hairer, A.M.Stuart and S. Vollmer. *Spectral gaps for a Metropolis-Hastings algorithm* . . . . `arxiv:1112.1392`.