Analysis of the Gibbs Sampler for Hierarchical Inverse Problems

Sergios Agapiou

Joint work with Johnathan M. Bardsley (Montana), Omiros Papaspiliopoulos (Barcelona) and Andrew M. Stuart (Warwick)

THE UNIVERSITY OF WARWICK

Overview

We consider the problem of determining an unknown function u from an indirect noisy observation y,

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

η additive noise, $\eta \sim N(0, \lambda^{-1}I)$. Parameter $\lambda^{-1} > 0$ models the noise level and $\mathcal{G} : X \to Y$ is the forward operator, X separable Hilbert, $Y = X$ or $\mathbb{R}^M$.

Bayesian approach:

- Gaussian prior on the unknown, $u \sim N(0, \delta^{-1}\mathcal{C}_0)$;
- $\delta^{-1} > 0$ models amplitude of prior, trace-class operator $\mathcal{C}_0$ models covariance structure;
- endow parameter δ with a hyper-prior, $\delta \sim \mathbb{P}(\delta)$.

Stuart '10, regularity conditions on $\mathcal{G}$ securing that conditionally on δ, posterior is well defined and $\mathbb{P}(u|y, \delta) \ll \mathbb{P}(u|\delta)$. We assume that this abs ctty property holds.

Aim: efficiently sample the joint posterior $\mathbb{P}(u, \delta|y) \propto \mathbb{P}(y|u, \delta)\mathbb{P}(u|\delta)\mathbb{P}(\delta)$.

Natural to use Gibbs sampler to probe $\mathbb{P}(u, \delta|y)$ and there is a range of possible parametrizations. To implement need to discretize unknown in $\mathbb{R}^N$. We study mixing of Gibbs sampler as $N \to \infty$ using function space intuition. High-dimensional analysis of Gibbs sampler is difficult, so we concentrate on linear conjugate setting to analyze.

Numerical simulations are presented which corroborate our theory and intuition.

Gibbs Samplers and Intuition

The most natural parametrization of the Gibbs sampler is the so-called *Centered Algorithm* (CA), used in *Bardsley '12*:

Centered Algorithm
- update $u^{(k+1)}|y, \delta^{(k)}$
- update $\delta^{(k+1)}|y, u^{(k+1)}$

Intuition: in continuum limit

- δ almost sure property of $u|\delta$;
- Abs ctty $\Rightarrow$ δ almost sure property of $u|y, \delta$;
- 2nd step, CA estimates δ pretending to know u; in fact knows $u^{(k+1)}|y, \delta^{(k)}$;
- $\delta^{(k+1)}|y, u^{(k+1)}$ point mass on $\delta^{(k)}$.

Intuition independent of $\dim(Y)$ and $\mathbb{P}(\delta)$. Expect that for large N, CA slows down.

Bad mixing of CA due to strong dependence of $u|\delta$ and δ as $N \to \infty$. Motivated by *Papaspiliopoulos et al. '07* we break dependence: $u = \delta^{-\frac{1}{2}}v$, $v \sim N(0, \mathcal{C}_0)$, $\delta \sim \mathbb{P}(\delta)$.

The resulting algorithm is termed the *Non-Centered Algorithm* (NCA) and is expected to be robust wrt N.

Non-Centered Algorithm
- update $u^{(k+1)}|y, \delta^{(k)}$, compute $v^{(k+1)} = (\delta^{(k)})^{\frac{1}{2}}u^{(k+1)}$
- update $\delta^{(k+1)}|y, v^{(k+1)}$

Linear Gaussian-Gamma Case

In order to analyze CA, we restrict to a linear conjugate setting.

$\mathcal{G} = K : X \to Y$ linear bounded, $\delta \sim \text{Ga}(\alpha, \beta)$.

- Discretize unknown in $\mathbb{R}^N$, data in $\mathbb{R}^M$, approximate operators $K, \mathcal{C}_0, I$.
- *Bayes'* theorem gives density of posterior on $\mathbb{R}^N \times \mathbb{R}$

$$p(u, \delta|y) \propto \delta^{\alpha + \frac{N}{2} - 1} \exp\Big(-\beta\delta - \frac{\lambda}{2}\|Ku - y\|^2 - \frac{\delta}{2}\|\mathcal{C}_0^{-\frac{1}{2}}u\|^2\Big)$$

- Conditional conjugacy

$$u|y, \delta \sim N(m, \mathcal{C})$$
$$\delta|y, u \sim \text{Ga}(\alpha + \frac{N}{2}, \beta + \frac{1}{2}\|\mathcal{C}_0^{-\frac{1}{2}}u\|^2)$$

Use understanding developed in *Agapiou et al. '13* to analyze norm in δ-draw.

Lemma:

$$\frac{1}{2}\|\mathcal{C}_0^{-\frac{1}{2}}u^{(k+1)}\|^2 = (\delta^{(k)})^{-1}\frac{N}{2} + (\delta^{(k)})^{-1}\sqrt{\frac{N}{2}}W_{1,N} + F_N(\delta^{(k)}).$$

i) Approximate $u^{(k+1)}$ with draw from $N(0, (\delta^{(k)})^{-1}\mathcal{C}_0)$;

ii) Get norm of white noise $N(0, (\delta^{(k)})^{-1}I_N)$, 1st and 2nd term given by LLN and CLT;

iii) $F_N(\delta^{(k)})$ well controlled correction term since $u^{(k+1)}$ drawn from $\mathbb{P}(u|y, \delta^{(k)}) \ll \mathbb{P}(u|\delta^{(k)})$.

Combined with property of gamma distribution, $\text{Ga}(\alpha + \frac{N}{2}, \beta + \mu^{-1}\frac{N}{2}) \simeq \text{Dirac}(\mu)$ for large N, gives $\delta^{(k+1)} \simeq \delta^{(k)}$ for large N.

Main Result - Large N behaviour of CA

Theorem: For $N \to \infty$, for any $\delta > 0$ we have y almost surely

$$\frac{N}{2}\mathbb{E}\left[\delta_N^{(k+1)} - \delta_N^{(k)} | \delta_N^{(k)} = \delta\right] = (\alpha + 1)\delta - f(\delta; y)\delta^2 + o(1)$$

$$\frac{N}{2}\text{Var}\left[\delta_N^{(k+1)} - \delta_N^{(k)} | \delta_N^{(k)} = \delta\right] = 2\delta^2 + \mathcal{O}(N^{-\frac{1}{2}}).$$

All expectations taken wrt the randomness in the algorithm.

Looks like numerical discretization of

$$d\delta = \Big((\alpha + 1)\delta - f(\delta; y)\delta^2\Big)\,dt + \sqrt{2}\delta\,dW$$

with time-step $2N^{-1}$; hence $\mathcal{O}(N)$ steps to sample posterior.

Simulation Results - Discrete Observations

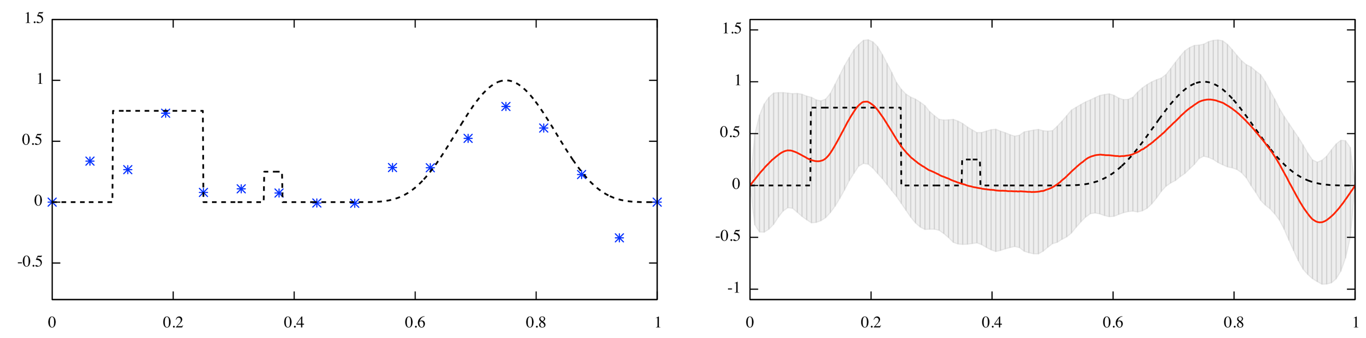

Figure: True solution and discrete noisy data (left) and true solution, sample mean and credibility bounds using NCA (right). $K = P_{15} \circ (I + c(-\Delta))^{-1}$, $\mathcal{C}_0 = (-\Delta)^{-1}$, $\lambda = 100$, $N = 1023$

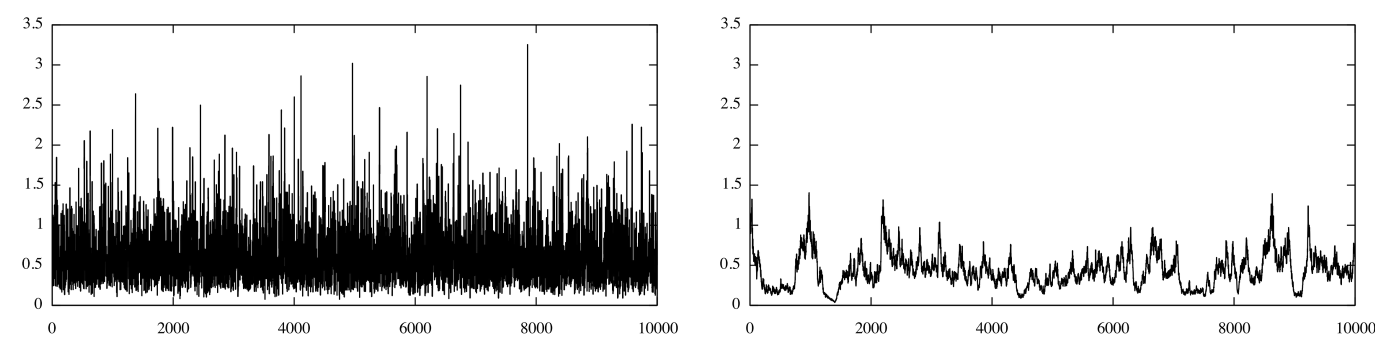

Figure: CA δ-chains for $N = 15$ (left) and $N = 1023$ (right)

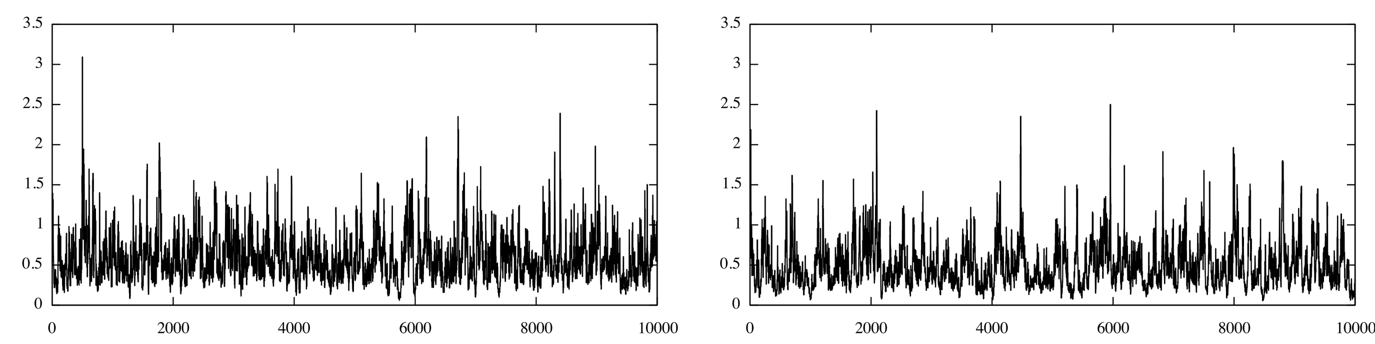

Figure: NCA δ-chains for $N = 15$ (left) and $N = 1023$ (right)

Conclusions - Further Work

- CA easy to implement, deteriorates for large dimension.
- NCA easy to implement, robust wrt dimension.

Unfortunately NCA deteriorates for small noise, λ large:

δ and v a posteriori dependent via data; for exact data we have strong dependence.

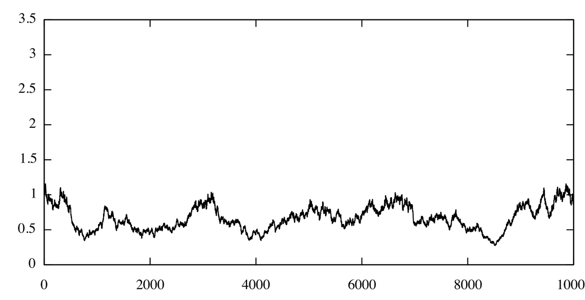

Figure: NCA δ-chain for small Noise $\lambda = 100^2$

- For linear case can integrate out u from likelihood. Marginal algorithm optimal but expensive. In nonlinear setting marginalization intractable, approximate by importance sampling leading to *Pseudo-Marginal approach*, *Filippone and Girolami '13*.
- Intuition applies to other parameters of the Gaussian prior, for example decay rate of the spectrum of $\mathcal{C}_0$ which relates to the regularity of draws.

References

S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, A. M. Stuart,
Analysis of the Gibbs Sampler for Hierarchical Inverse Problems, arXiv:1311.1138.

S. Agapiou, S. Larsson, A. M. Stuart,
Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems, Stoch. Proc. Appl., 123(10) 2013.

O. Papaspiliopoulos, G. O. Roberts, M. Sköld
A general framework for the parametrization of hierarchical models, Statist. Sci., 22(1), 2007.

J. M. Bardsley
MCMC-based image reconstruction with uncertainty quantification, SIAM J. Sci. Comput. 34, 2012.