Analysis of the Gibbs Sampler for Hierarchical Inverse Problems

Sergios Agapiou

Mathematics Institute, University of Warwick

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

MCQMC2014, KU Leuven, 6-11 April 2014

Outline

Hierarchical Inverse Problems with Gaussian Noise and Prior

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

- $\mathcal{G} : X \to Y$ continuous, X separable Hilbert, $Y = X$ or $\mathbb{R}^M$.
- $\eta \sim N(0, \lambda^{-1} I)$.
- Prior $u \sim N(0, \delta^{-1}\mathcal{C}_\theta)$, $\delta^{-1} > 0$ amplitude, $\sigma(\mathcal{C}_\theta) \asymp \{j^{-2\theta-1}\}_{j\in\mathbb{N}}$, $\theta > 0$ regularity.
- Fix θ. Hyper-prior $\delta \sim \mathbb{P}(\delta)$. (can also fix $\delta = 1$, hyper-prior on θ)
- Stuart '10 conditions on $\mathcal{G}$

$$\mathbb{P}(u|y, \delta) \ll \mathbb{P}(u|\delta).$$

Sampling the Posterior

AIM: efficiently sample $\mathbb{P}(u, \delta|y) \propto \mathbb{P}(u|y, \delta)\mathbb{P}(\delta|y)$

Natural to use Metropolis within Gibbs algorithm; range of possible parametrizations.

Centered MwG Algorithm

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

Bardsley '12

Individually each Metropolis step well understood, study interplay.

Sampling the Posterior - Intuition

Think ∞-dim, discretize unknown in $\mathbb{R}^N$: study mixing as $N \to \infty$.

Centered MwG Algortihm: 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 know $u^{(k+1)}|y,\delta^{(k)}$.
- $\delta^{(k+1)}|y,u^{(k+1)}$ point mass on $\delta^{(k)}$.

$$\delta_N^{(k+1)} \simeq \delta_N^{(k)}, \text{ for large } N.$$

- Intuition independent of $\dim(Y)$ and $\mathbb{P}(\delta)$.

Sampling the Posterior - Intuition

- Bad mixing of CA due to strong dependence of $u|\delta$ and δ as $N \to \infty$.
- Break dependence: $u = \delta^{-\frac{1}{2}} v$, $v \sim N(0, \mathcal{C}_\theta)$, $\delta \sim \mathbb{P}(\delta)$.

Non-Centered MwG Algorithm

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

Roberts and Stramer '01, Papaspiliopoulos, Roberts and Sköld '07

- NCA robust wrt N.

Sampling the Posterior - Hierarchical Regularity

Intuition applies also for sampling $\mathbb{P}(u, \theta | y)$ when $\theta \sim \mathbb{P}(\theta)$ and $\delta = 1$ fixed.

- θ almost sure property of $u|\theta$.
- Conditions on $\mathcal{G}$ secure $\mathbb{P}(u|y,\theta) \ll \mathbb{P}(u|\theta)$.
- CA deteriorates as $N \to \infty$.
- Reparametrize, $u = \mathcal{C}_\theta^{\frac{1}{2}} v$, $v \sim N(0, I)$, $\theta \sim \mathbb{P}(\theta)$.
- NCA robust wrt N.

Analysis of GS difficult, no convergence results in limit $N \to \infty$ in nontrivial settings.

Linear Setting - Conditional Conjugacy

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

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

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

- Conditional conjugacy

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

- Use understanding developed in A, Larsson, Stuart '13 to analyze.

Linear Inverse Problem - Analysis

- Assume prior regular enough, $\mathbb{P}(u|y,\delta) = N(m,C) \ll \mathbb{P}(u|\delta) = N(0,\delta^{-1}\mathcal{C}_\theta)$ in limit.
- Use consistent discretizations of operators.

Lemma

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

i) 1st and 2nd terms LLN and CLT terms if $u^{(k+1)}$ drawn from prior $\mathbb{P}(u|\delta^{(k)})$;

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

- Property: $\mathrm{Ga}(\alpha+\frac{N}{2},\beta+\mu^{-1}\frac{N}{2}) \simeq \mathrm{Dirac}(\mu)$, for large N.
- Combine $\mathbb{P}(\delta^{(k+1)}|y,u^{(k+1)}) \simeq Ga(\alpha+\frac{N}{2},\beta+(\delta^{(k)})^{-1}\frac{N}{2}) \simeq \mathrm{Dirac}(\delta^{(k)})$, for large N.

Result - δ Evolves Slowly

Theorem (A, Bardsley, Papaspiliopoulos, Stuart '13)

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

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

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

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.

Marginal Algorithm

In linear case can analytically find $y|\delta$ by integrating out u from likelihood.

Marginal algorithm

- update $u^{(k+1)}|y, \delta^{(k)}$;
- draw $\delta^{(k+1)}|y$.

MA optimal (in principle i.i.d δ samples) robust in N, use as gold standard.

Hierarchical Amplitude, $X = L^2[0,1]$, $Y = \mathbb{R}^{15}$

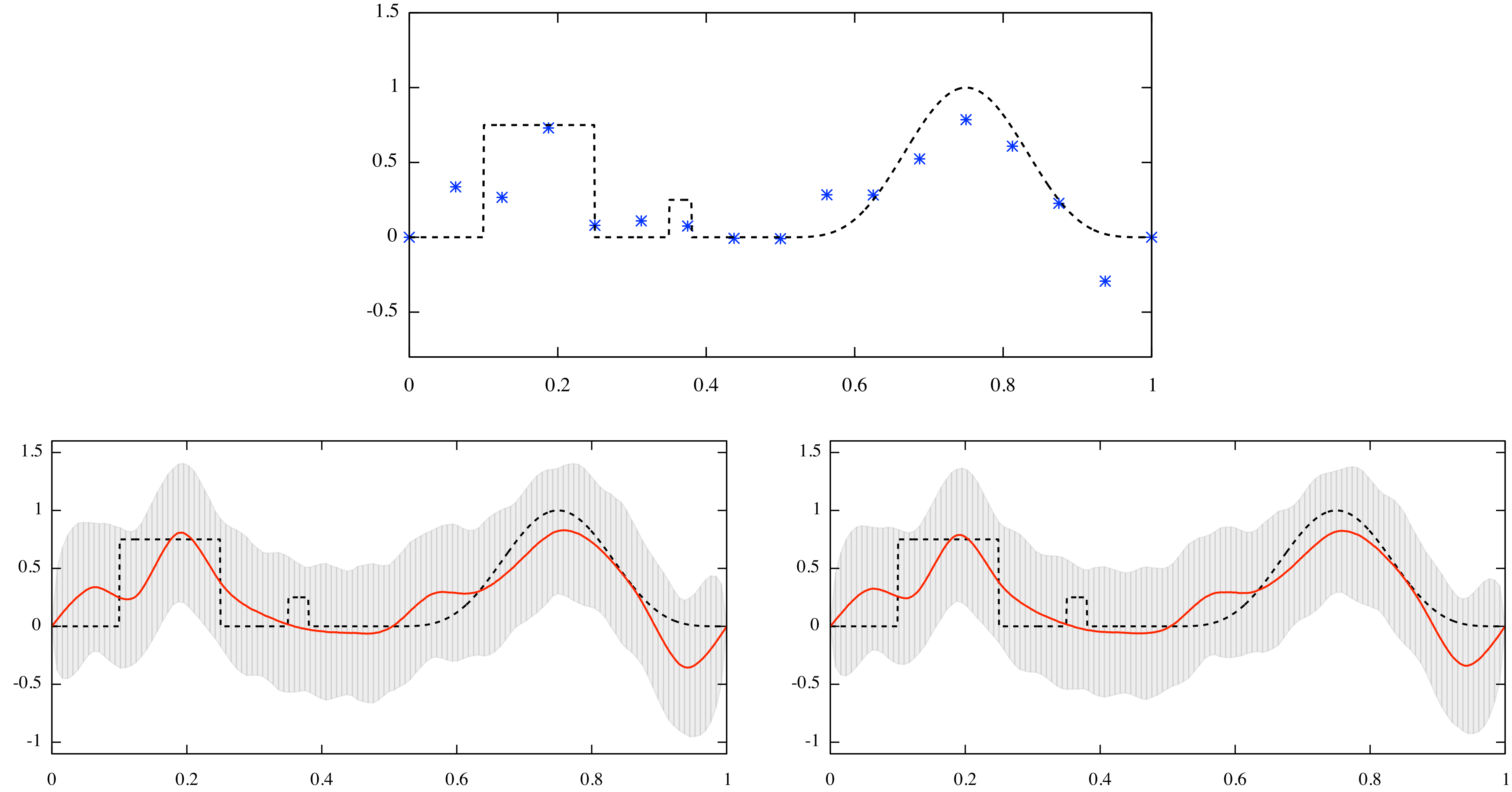
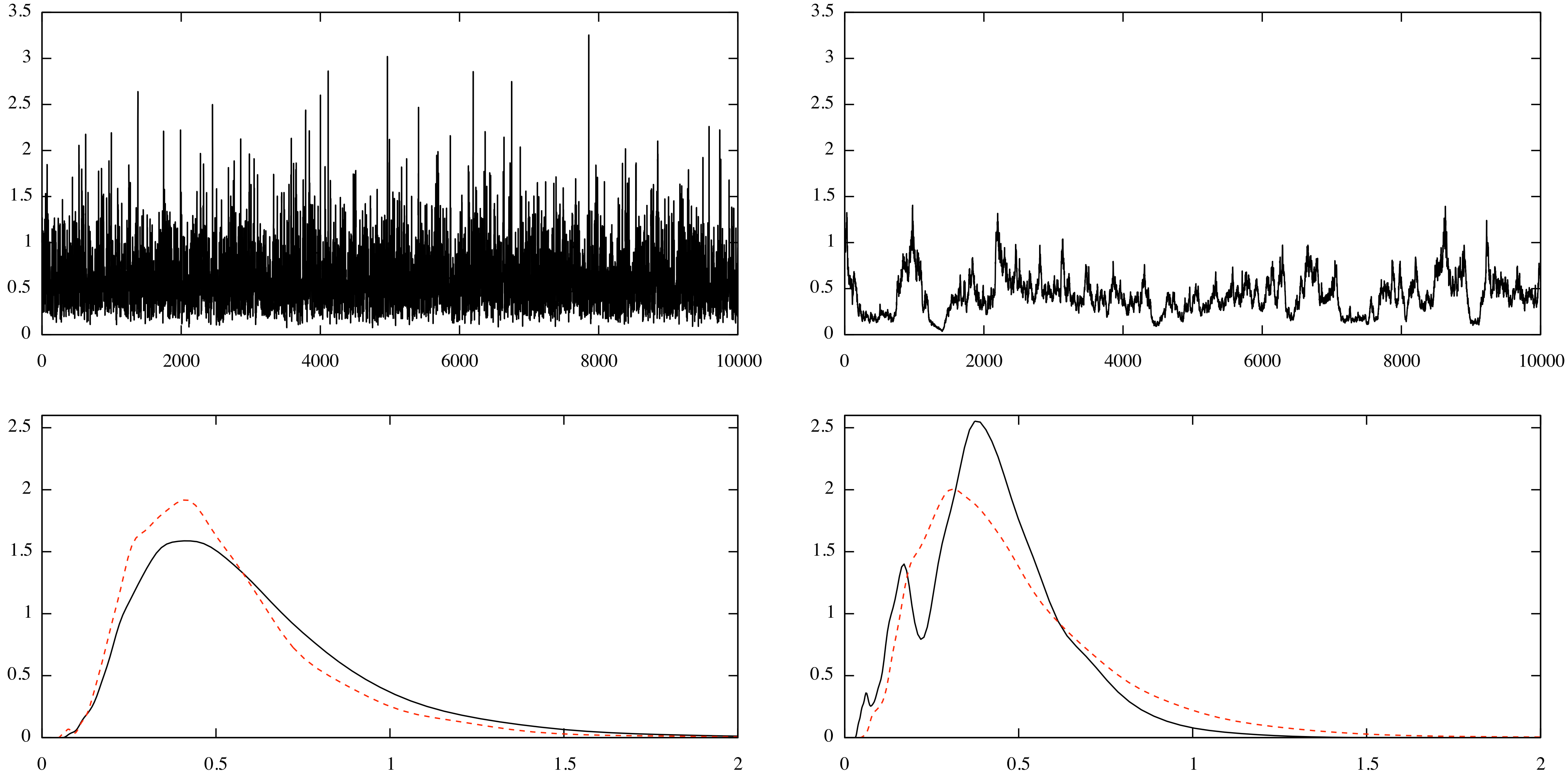
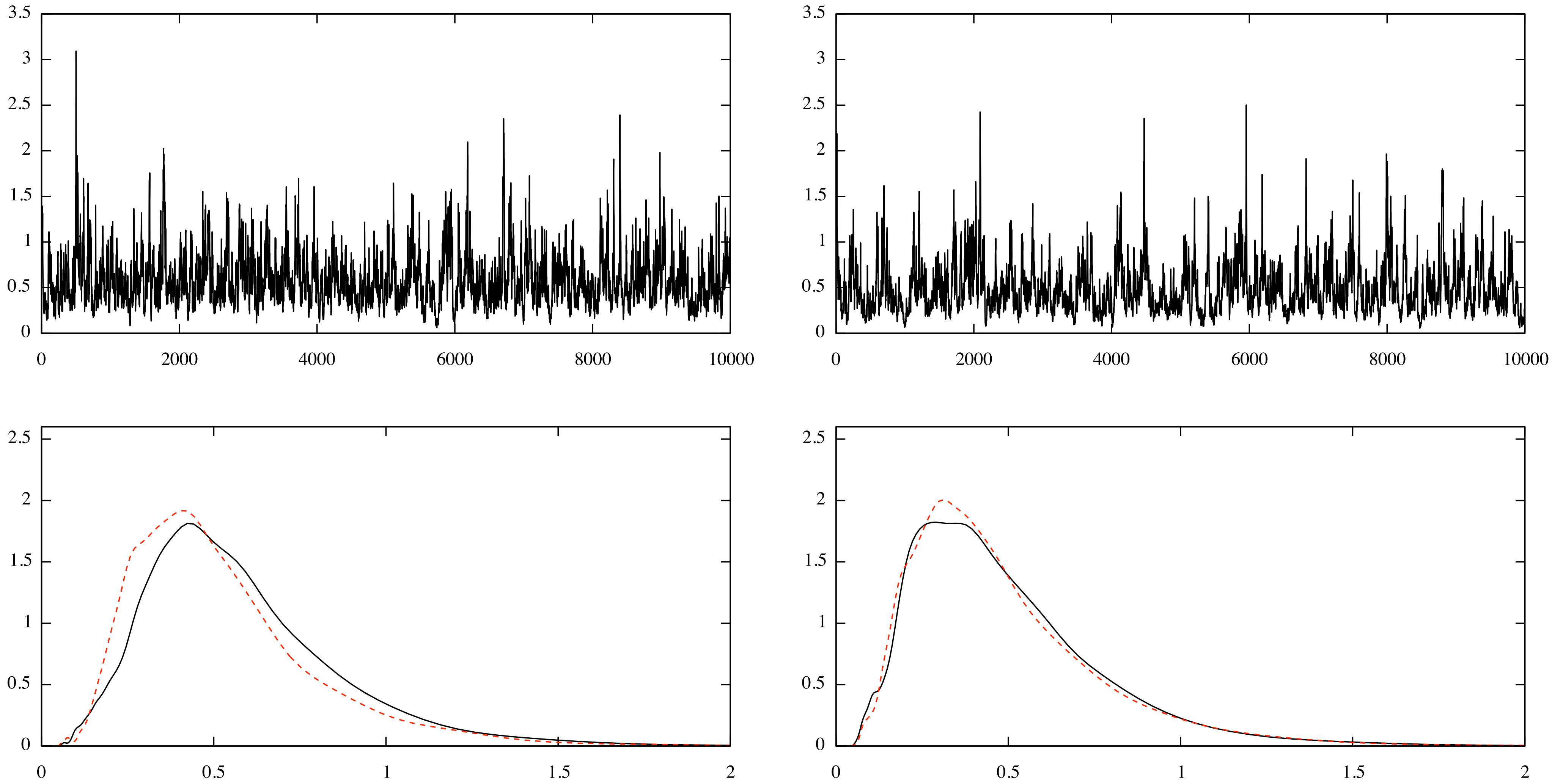


Figure: $K = P_{15} \circ (I + c(-\Delta))^{-1}$, $\mathcal{C}_\theta = (-\Delta)^{-1}$, $\delta \sim \mathrm{Ga}(1, 10^{-4})$, $\lambda = 100$, $N = 1023$

Centered Algorithm, $N = 15, 1023$



Non-Centered Algorithm, $N = 15, 1023$



Conclusions

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

δ (resp. θ) and v a posteriori dependent via data; for exact data strong dependence.

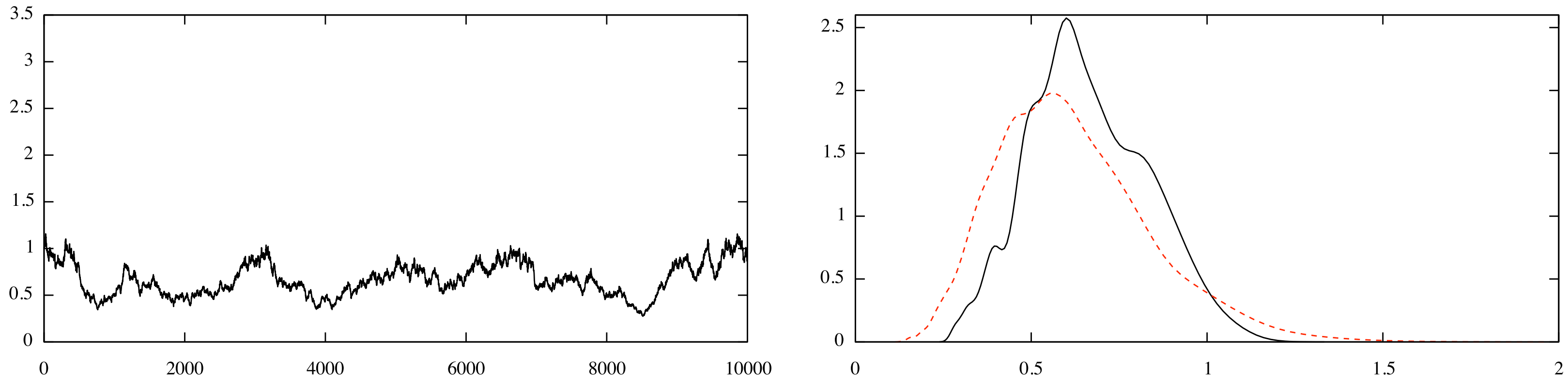


Figure: Non-Centered Algorithm Amplitude, Small Noise $\lambda = 100^2$

Further Work

More research needed in small noise limit.

- MA optimal but in nonlinear setting marginalization is intractable, approximate using importance sampling leading to Pseudo-Marginal approach, Andrieu and Roberts '09, Filippone and Girolami '13.
- Interweaving method of Yu and Meng '11, PNCA of Papaspiliopoulos Roberts and Sköld '03 do not improve significantly.
- Random truncation of prior and use of Reversible Jump algorithm promising.

Further analysis of GS in high dimensions, extending theory to nonlinear setting, other parameters in prior, other priors...

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

- S. Agapiou, J. Bardsley, O. Papaspiliopoulos and A. M. Stuart, *Analysis of the Gibbs sampler for hierarchical inverse problems*, http://arxiv.org/abs/1311.1138
- S. Agapiou, S. Larsson and A. M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stoch. Proc. Appl., 123(10) 2013
- A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer. 19 (2010).
- O. Papaspiliopoulos, G. O. Roberts, and 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)
- M. Filippone, M. Girolami, *Pseudo-Marginal Bayesian inference for Gaussian processes*, http://arxiv.org/abs/1310.0740

Hierarchical Regularity, $X = Y = L^2[0,1]$

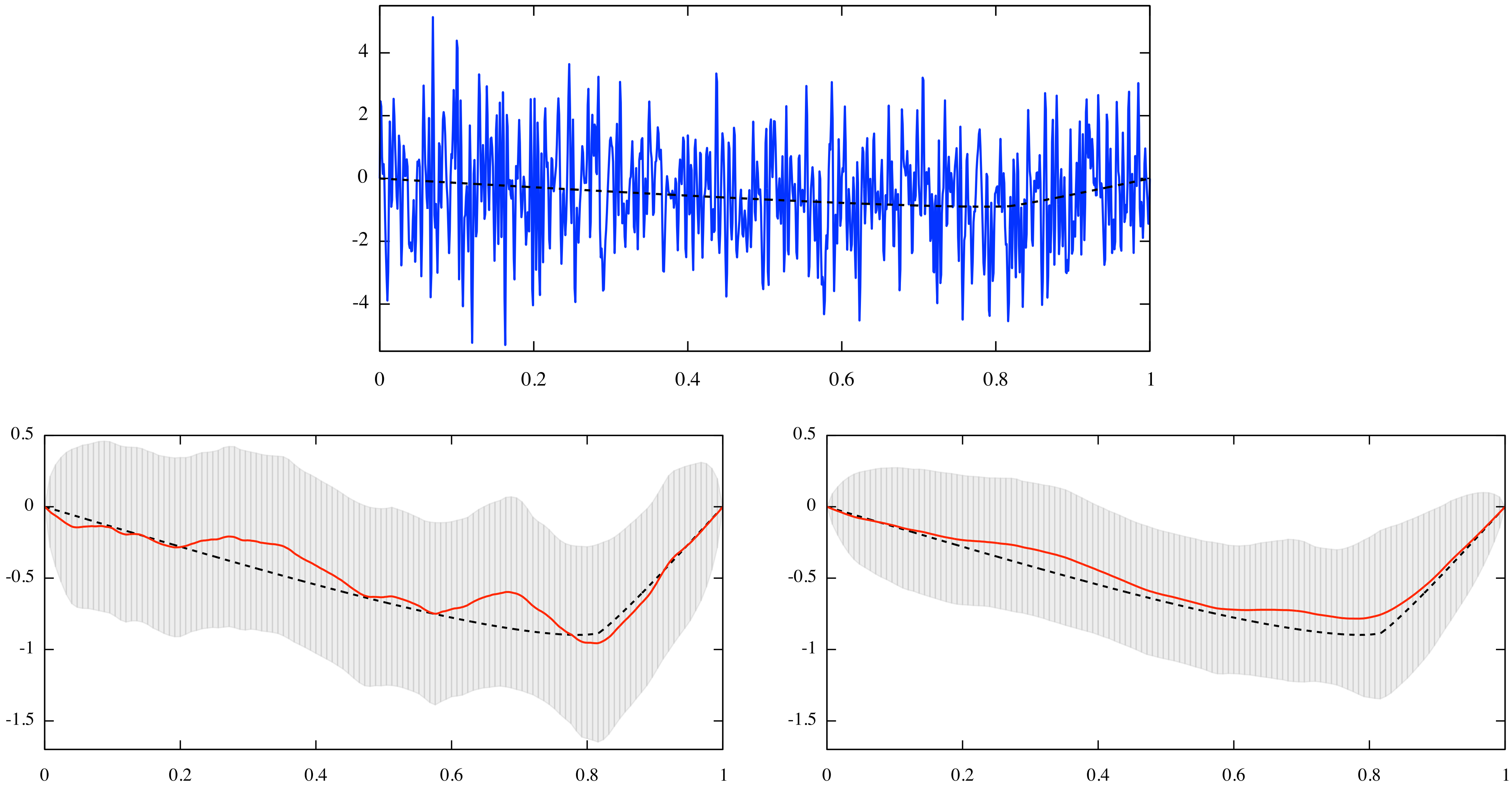
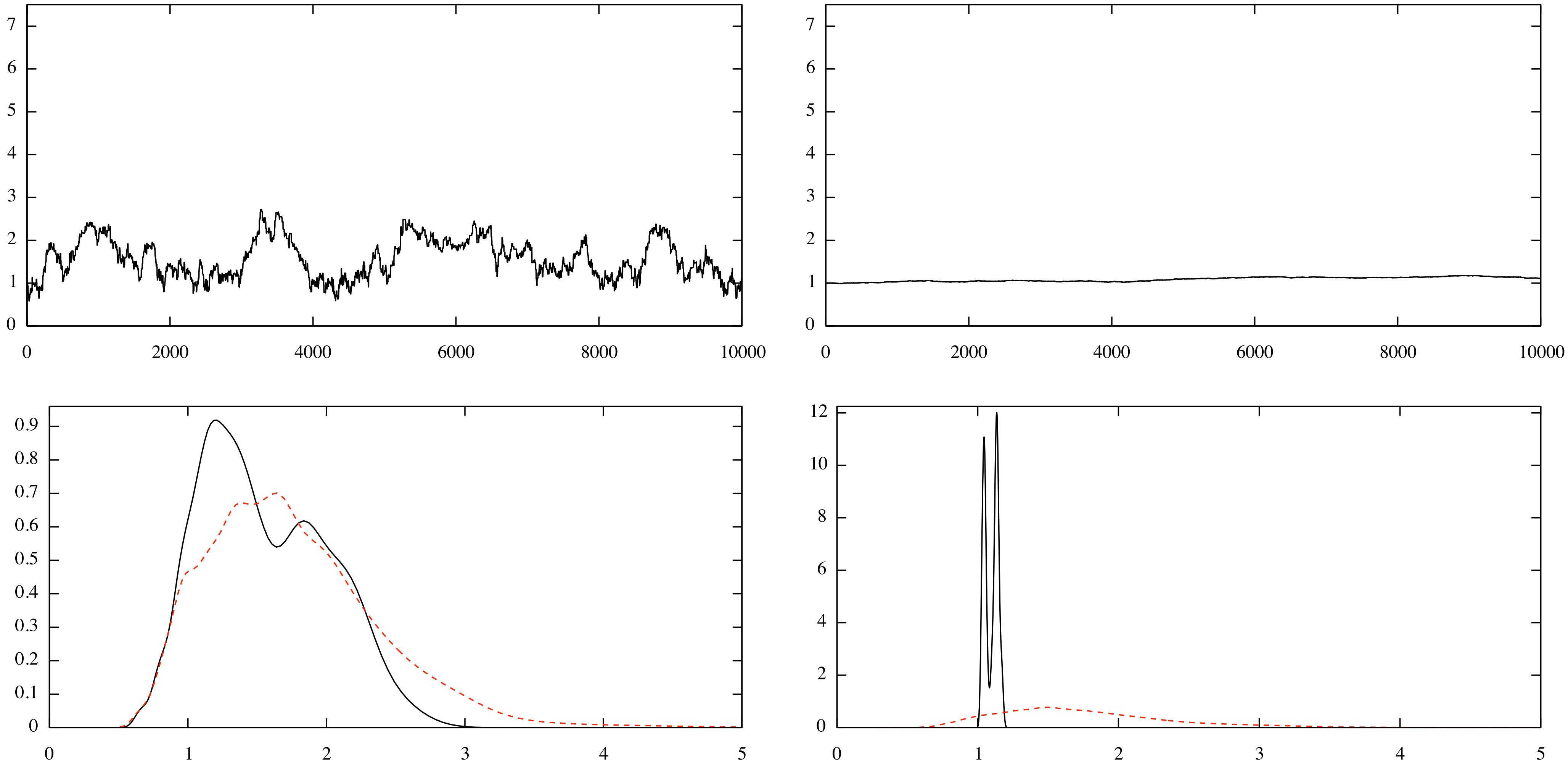
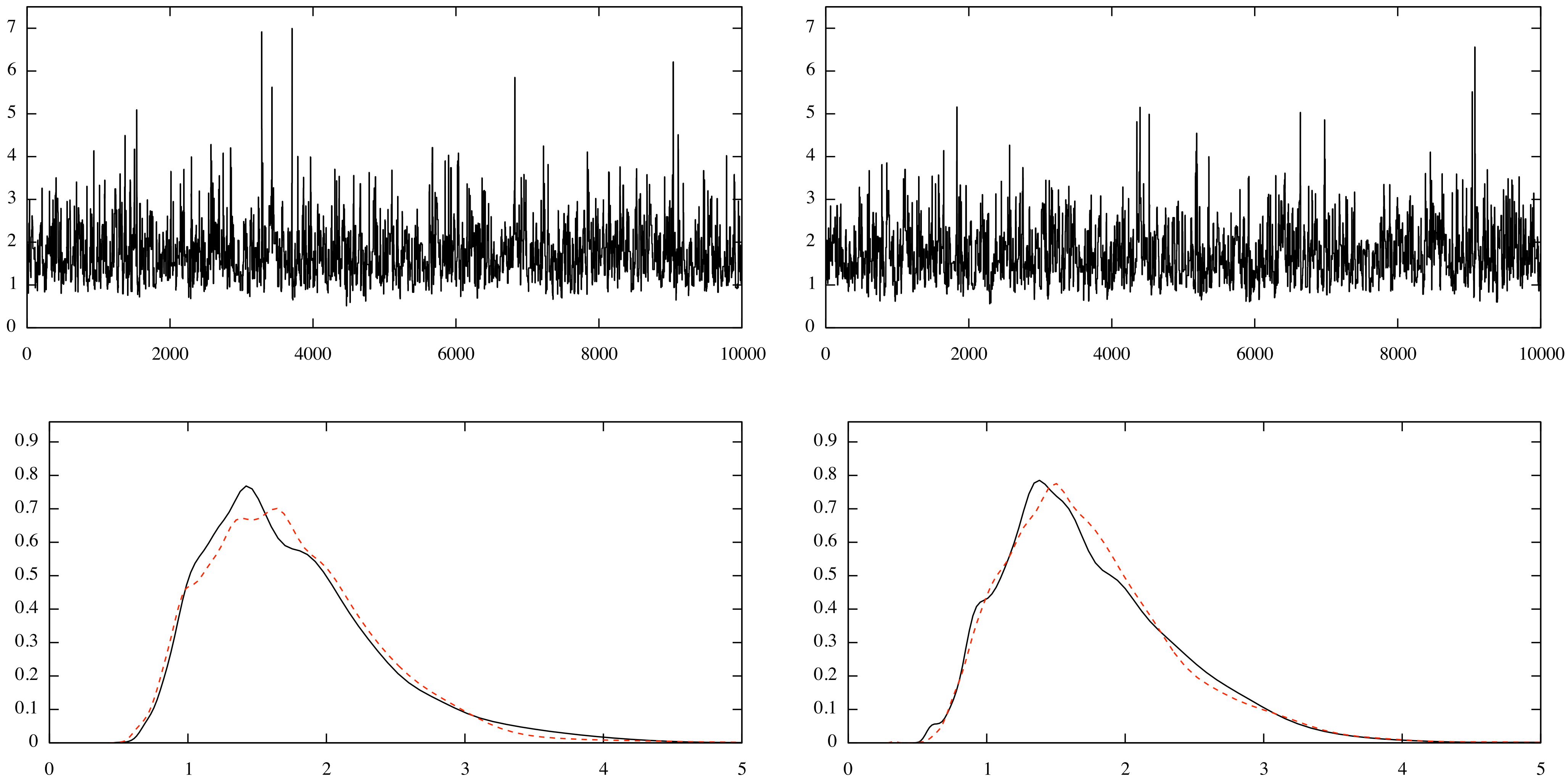


Figure: $K = I$, $\mathcal{C}_\theta = (-\Delta)^{-\theta-\frac{1}{2}}$, $p(\theta) = e^{-\theta}$, $\lambda = 200$, $N = 512$, $\theta_{true} = 1.75$

Centered Algorithm, $N = 32, 8192$



Non-Centered Algorithm, $N = 32, 8192$



Non-Centered Algorithm Regularity, Small Noise $\lambda = 200^2$