Bayesian posterior contraction rates via classical regularization techniques

Sergios Agapiou

Department of Statistics, University of Warwick

Applied Inverse Problems Conference
May 25-29, 2015, Helsinki

Enabling Quantification of EQUIP Uncertainty for Inverse Problems

http://www.sergiosagapiou.com

S. Agapiou, P. Mathé, *Preconditioning the prior to overcome saturation in Bayesian inverse problems*, submitted, arXiv:1409.6496.

S. Agapiou, S. Larsson, A. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stochastic Processes and their Applications, 2013.

S. Agapiou, A. Stuart, Y-X Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, Journal of Inverse and Ill-posed Problems, 2014.

S. Agapiou, J. Bardsley, O. Papaspiliopoulos, A. Stuart, *Analysis of the Gibbs Sampler for Hierarchical Inverse Problems*, SIAM J. UQ, 2014.

Outline

Outline

Bayesian linear inverse problems

$$y = Ku + \delta\xi$$

- $u \in X$ unknown, $y \in Y$ observation, X, Y separable Hilbert.
- $K : X \to Y$ linear bounded forward operator.
- $\xi \sim N(0, I)$ noise, $\delta > 0$ known noise level.
- Likelihood $y|u \sim N(Ku, \delta^2 I)$.
- Prior $u \sim N(m_0, \frac{\delta^2}{\alpha} C_0)$, C_0 trace-class, $\alpha > 0$ scaling parameter.
- Posterior $u|y \sim \mu_\alpha^{y,\delta} = N(u_\alpha^\delta, C_\alpha^\delta)$.

Bayesian linear inverse problems

- Formally

$$\mu_\alpha^{y,\delta}(du) \propto \exp\left(-\frac{1}{2\delta^2}\|y-Ku\|^2 - \frac{\alpha}{2\delta^2}\|C_0^{-\frac{1}{2}}(u-m_0)\|^2\right) du.$$

(α regularization parameter)

- Completing the square

$$(C_\alpha^\delta)^{-1} = \frac{\alpha}{\delta^2}C_0^{-1} + \frac{1}{\delta^2}K^*K$$

$$(C_\alpha^\delta)^{-1}u_\alpha^\delta = \frac{1}{\delta^2}K^*y + \frac{\alpha}{\delta^2}C_0^{-1}m_0$$

- Let $B = KC_0^{\frac{1}{2}}$ compact

$$(C_\alpha^\delta)^{-1} = \frac{1}{\delta^2}C_0^{-\frac{1}{2}}(\alpha I + B^*B)\,C_0^{-\frac{1}{2}}$$

$$u_\alpha^\delta = C_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}B^*(y-Km_0) + m_0.$$

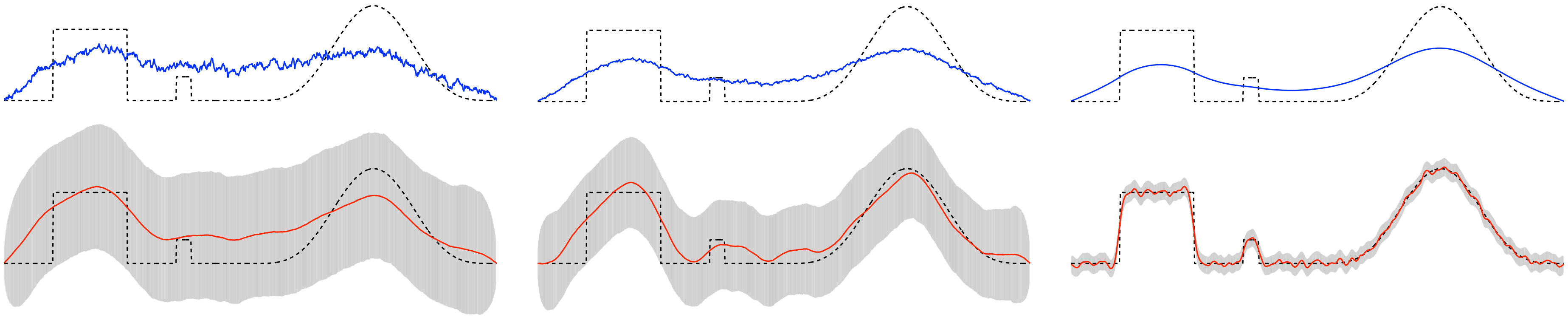
Outline

Frequentist asymptotic performance in small noise limit

Consider sequence of data generated from fixed underlying truth $u^\dagger$, where $\delta \to 0$

$$y^\delta = Ku^\dagger + \delta\xi.$$

- Posterior $\mu_\alpha^{y^\delta,\delta} := \mu_\alpha^{y=y^\delta,\delta}$.
- As $\delta \to 0$, can we choose $\alpha = \alpha(\delta) \to 0$ s.t. "$\mu_\alpha^{y=y^\delta,\delta} \to \delta_{u^\dagger}$"?



- First studies of convergence of posterior in BIP, in Ky-Fan metric (finite-dim).

A. Hofinger, H. Pikkarainen, *Convergence rate for the Bayesian approach to linear inverse problems*, Inverse Problems, 2007.

A. Neubauer, H. Pikkarainen, *Convergence results for the Bayesian inversion theory*, Journal of Inverse and Ill-posed Problems, 2008.

Squared Posterior Contraction

Let $\mathcal{P}^\delta = N(Ku^\dagger, \delta^2 I)$ distribution generating y^δ. Define

$$\begin{aligned}
\mathrm{SPC} \; &:= \mathbb{E}^{\mathcal{P}^\delta} \mathbb{E}^{\mu_\alpha^{y^\delta,\delta}} \|u^\dagger - u\|^2 \\
&= \mathbb{E}^{\mathcal{P}^\delta} \left(\|u^\dagger - u_\alpha^\delta\|^2 + \mathrm{Tr}\,(C_\alpha^\delta) \right) \\
&= \underbrace{\|u^\dagger - \mathbb{E}^{\mathcal{P}^\delta} u_\alpha^\delta\|^2}_{\text{squared bias}} + \underbrace{\mathbb{E}^{\mathcal{P}^\delta}\|u_\alpha^\delta - \mathbb{E}^{\mathcal{P}^\delta} u_\alpha^\delta\|^2}_{\text{est. variance}} + \underbrace{\mathrm{Tr}\,(C_\alpha^\delta).}_{\text{pos. spread}} \\
&:= b_{u^\dagger}^2(\alpha) + V_\alpha^\delta + \mathrm{Tr}\,(C_\alpha^\delta).
\end{aligned}$$

Minimax framework

- Assume $u^\dagger \in X^\gamma$, γ regularity parameter.
- Minimax rate: benchmark rate for given forward operator K and smoothness of truth X^γ.
- Minimax rates for standard forward operators and smoothness classes available in
  - L. Cavalier, *Nonparametric statistical inverse problems*, Inverse Problems, 2008.
- Optimality: can we choose $\alpha = \alpha(\delta; \gamma) \to 0$ such that $\mathrm{SPC} \to 0$ at minimax rate?
- First study of SPC rates
  - B. Knapik, A. van der Vaart, H. van Zanten, *Bayesian inverse problems with Gaussian priors*, Annals of Statistics, 2011.

Example - diagonal setting

$X = L^2(0,1)$, let $\mathcal{A} = -\boldsymbol{\Delta}$, for $\boldsymbol{\Delta}$ the Dirichlet-Laplacian.

- Moderately ill-posed forward operator

$$K^*K = \mathcal{A}^{-\ell}, \ell > 0.$$

- Severely ill-posed forward operator

$$K^*K = \exp(-\mathcal{A}^{\frac{b}{2}}), b > 0.$$

- Sobolev-type smoothness

$$X^\gamma = S^\gamma := \{u : u = \mathcal{A}^{-\frac{\gamma}{2}} w, \|w\| \leq 1\}.$$

- Analytic-type smoothness

$$X^\gamma = A^\gamma := \{u : u = \exp(-\gamma\mathcal{A}^{\frac{1}{2}}) w, \|w\| \leq 1\}.$$

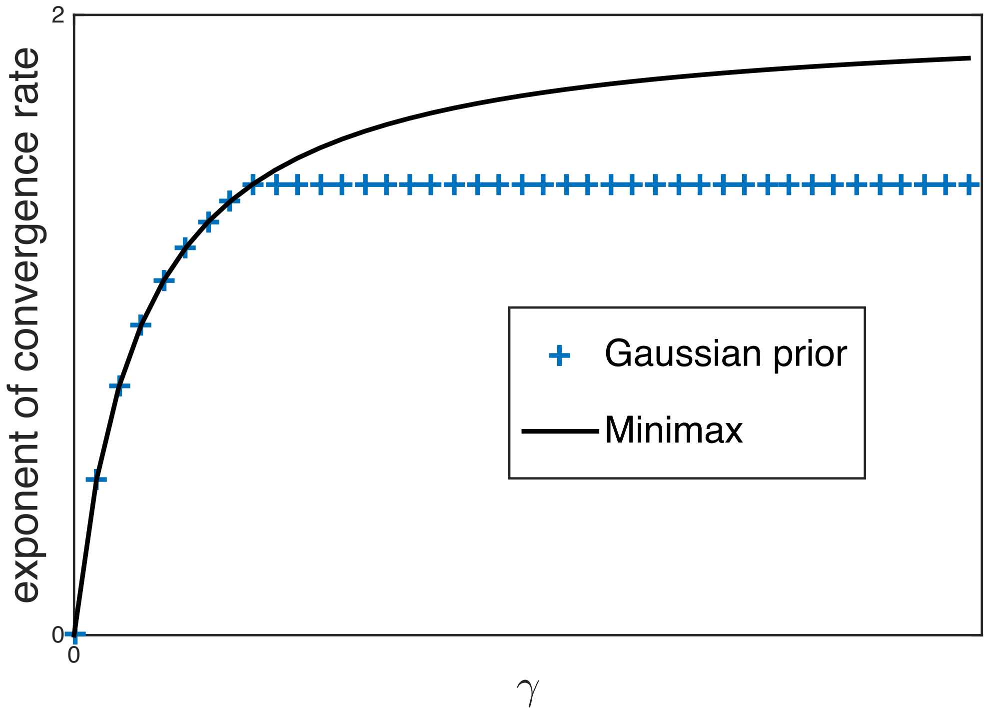
- Prior covariance with $C_0 = \mathcal{A}^{-\frac{1}{2}-p}, p > 0$.
- C_0, K^*K commute.

Moderately ill-posed operators under Sobolev-type truth regularity

- $K^*K = \mathcal{A}^{-\ell}, \ u^\dagger \in S^\gamma, \ C_0 = \mathcal{A}^{-\frac{1}{2}-p}, \ m_0 = 0.$
- KVZ11 studied diagonal setting, ALS13 extended to non-diagonal setting.
- Fix $p, \ell > 0$. Then for optimal choice $\alpha = \alpha(\delta;\gamma)$

$$\mathrm{SPC} \asymp \delta^{\mathrm{c}(\gamma;\mathrm{a},\ell)}.$$

- Saturation when truth too smooth.



Severely ill-posed operators under Sobolev-type truth regularity

- $K^*K = \exp(-\mathcal{A}^{\frac{b}{2}}),\ u^\dagger \in S^\gamma,\ C_0 = \mathcal{A}^{-\frac{1}{2}-p},\ m_0 = 0.$
- Studied in ASZ14 and

  B. Knapik, A. van der Vaart, H. van Zanten, *Bayesian recovery of the initial condition for the heat equation*, Communications in Statistics - Theory and Methods, 2013.

- Fix $p, b > 0$. Then for optimal choice $\alpha = \alpha(\delta; \gamma)$
$$\mathrm{SPC} \asymp \log^{-\mathrm{c}(\gamma;\mathrm{b})}(1/\delta).$$
- No saturation phenomenon.

Outline

Contribution

- Focus on commuting/diagonal setting.
- Previous studies rely on explicit calculations.
- We use abstract regularization theory techniques from

B. Hofmann, P. Mathé, *Analysis of profile functions for general linear regularization methods*, SIAM Journal of Numerical Analysis, 2007.

P. Mathé, *Saturation of regulaziation methods for linear ill-posed problems in Hilbert spaces*, SIAM Journal of Numerical Analysis, 2004.

- Formulate abstract theory. Existing and new (diagonal) results obtained as special cases.
- Propose data dependent choice of prior mean m_0 resulting in delaying/removing saturation.

Estimation variance and posterior spread

$$\mathrm{SPC} = b_{u^\dagger}^2(\alpha) + V_\alpha^\delta + \mathrm{Tr}\,(C_\alpha^\delta).$$

- As shown in

K. Lin, S. Lu, P. Mathé, *Oracle-type posterior contraction rates in Bayesian inverse problems*, 2014.

$$V_\alpha^\delta \le c\,\mathrm{Tr}\,(C_\alpha^\delta),$$

$c > 0$ independent of δ, α.

- Suffices to estimate the posterior spread (straightforward) and the bias.
- We focus on the bias, the source of saturation (since only term depending on $u^\dagger$).

Regularization theory - regularization filters

- Loosely speaking $g_\alpha : (0,\infty) \to \mathbb{R}$, $\alpha > 0$, is a regularization filter if
  - $g_\alpha(t)$ bounded $\forall \alpha > 0$;
  - $g_\alpha(t) \to \frac{1}{t}$ as $\alpha \to 0$.
- Associated residual function $r_\alpha(t) = 1 - t g_\alpha(t)$.
- Tikhonov filter

$$g_\alpha(t) = \frac{1}{\alpha + t}, \qquad r_\alpha(t) = \frac{\alpha}{\alpha + t}.$$

- Spectral cut-off

$$g_\alpha(t) = \begin{cases} \frac{1}{t}, & t \geq \alpha \\ 0, & t < \alpha \end{cases}, \qquad r_\alpha(t) = \begin{cases} 0, & t \geq \alpha \\ 1, & t < \alpha. \end{cases}$$

Regularization theory - index functions and source sets

- $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ index function if continuous, nondecreasing, $\varphi(0) = 0$.
- eg $\varphi(t) = t^s$, $s > 0$ index function.
- For φ, ψ index functions, we write $\varphi \prec \psi$ if ψ decays to zero faster than φ.
- eg $\varphi(t) = t^s$, $\psi(t) = t^r$, $\varphi \prec \psi$ if $s < r$.
- Source set: assume $u^\dagger \in A_{\varphi_\gamma} = \{u : u = \varphi_\gamma(B^*B)w, \|w\| \leq 1\}$, for index function φ_γ.

Regularization theory - qualification

- Index function φ is qualification for regularization g_α if

$$r_\alpha(t)\varphi(t) \le c\varphi(\alpha), \ \forall \alpha, t.$$

- If φ qualification for g_α and $\psi \prec \varphi$, then ψ qualification for g_α.
- Quantifies ability of regularization to take smoothness into account.
- Bias of classically regularized inversion of B

$$\|r_\alpha(B^*B)u^\dagger\| = \|r_\alpha(B^*B)\varphi_\gamma(B^*B)w\| \le \|r_\alpha(B^*B)\varphi_\gamma(B^*B)\| \le c\varphi_\gamma(\alpha),$$

  provided $\varphi_\gamma \prec \varphi$.
- Want qualification to decay to zero as quickly as possible.

Regularization theory - qualification

- eg Tikhonov
$$r_\alpha(t)t = \frac{\alpha}{t+\alpha} t \leq \alpha$$
hence $\varphi(t) = t$ (maximal) qualification.

- eg Spectral cut-off
$$r_\alpha(t)\varphi(t) = \begin{cases} 0 \cdot \varphi(t) \leq \varphi(\alpha), & t \geq \alpha \\ 1 \cdot \varphi(t) \leq \varphi(\alpha), & t < \alpha \end{cases},$$
has arbitrary qualification.

Main result about bias

Investigate effect on bias of choosing

$$m_0 = m_\alpha^\delta = C_0^{\frac{1}{2}} g_\alpha(B^*B)B^*y^\delta.$$

Proposition (A., Mathé 2014)

Assume $u^\dagger \in A_{\varphi_\gamma} = \{u : u = \varphi_\gamma(B^*B)w, \|w\| \leq 1\}$.

i) *Low smoothness* $\varphi_\gamma(t) \prec t$: independently of m_0

$$b_{u^\dagger}(\alpha) \leq c\varphi_\gamma(\alpha).$$

ii) *High smoothness* $t \prec \varphi_\gamma(t)$, *no preconditioning* $m_0 = 0$:

$$b_{u^\dagger}(\alpha) \asymp \alpha.$$

iii) *High smoothness* $t \prec \varphi_\gamma(t)$, *with preconditioning* $m_0 = m_\alpha^\delta$: if $\frac{\varphi_\gamma(t)}{t}$ qualification for g_α

$$b_{u^\dagger}(\alpha) \leq c\varphi_\gamma(\alpha).$$

Sketch of proof - fixed mean

- Assume $m_0 = 0$.

$$b_{u^\dagger}(\alpha) = \|u^\dagger - \mathbb{E}^{\mathcal{P}^\delta} u_\alpha^\delta\| = \alpha \|(\alpha I + B^* B)^{-1} u^\dagger\|.$$

- Norm term on rhs, at best $\mathcal{O}(1)$ as $\alpha \to 0$. Happens if $u^\dagger \in \mathcal{D}((B^*B)^{-1})$, i.e. $\varphi_\gamma(t) = t$.

- For low smoothness $u^\dagger \in A_{\varphi_\gamma}$ with $\varphi_\gamma(t) \prec t$

$$\begin{aligned} b_{u^\dagger}(\alpha) &= \|\alpha(\alpha I + B^* B)^{-1} \varphi_\gamma(B^* B) w\| \\ &\leq \|\alpha(\alpha I + B^* B)^{-1} \varphi_\gamma(B^* B)\| \\ &\leq \varphi_\gamma(\alpha), \end{aligned}$$

since $\frac{\alpha}{\alpha+t}$ residual of Tikhonov which has maximal qualification t.

Sketch of proof - preconditioned mean

- For $m_0 = m_\alpha^\delta = C_0^{\frac{1}{2}} g_\alpha(B^*B)B^*y^\delta$,

$$b_{u^\dagger}(\alpha) = \|u^\dagger - \mathbb{E}^{\mathcal{P}^\delta} u_\alpha^\delta\| = \|\alpha(\alpha I + B^*B)^{-1} r_\alpha(B^*B)u^\dagger\|$$

- eg if g_α Tikhonov filter

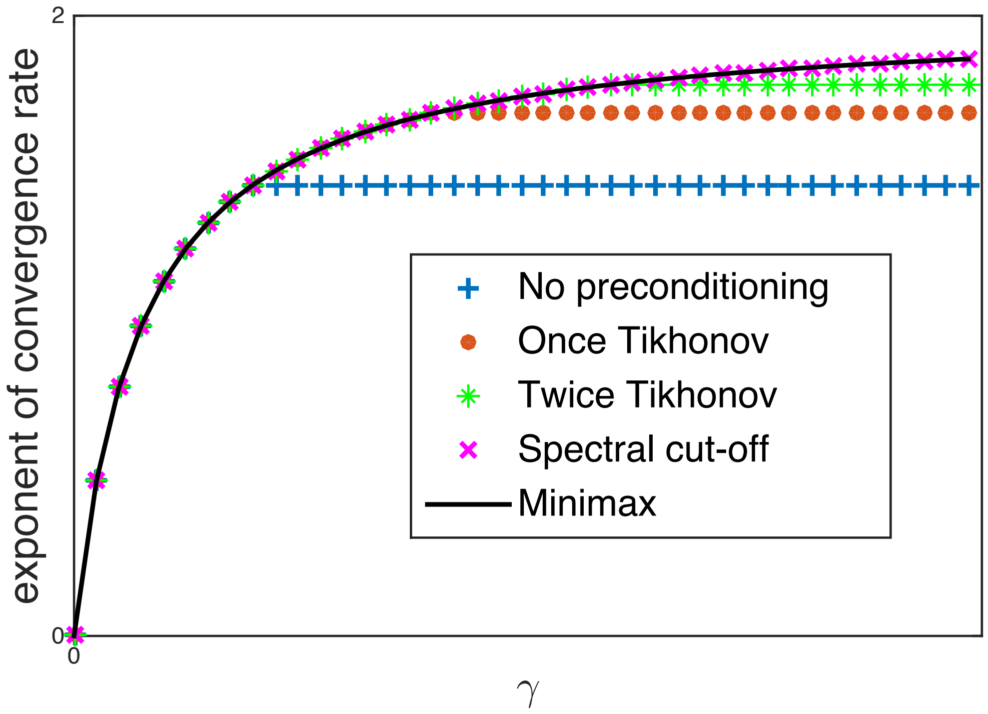
$$b_{u^\dagger}(\alpha) = \alpha^2 \|(\alpha I + B^*B)^{-2} u^\dagger\|.$$

- Norm term on rhs, at best $\mathcal{O}(1)$ as $\alpha \to 0$. Happens if $u^\dagger \in \mathcal{D}((B^*B)^{-2})$, i.e. $\varphi_\gamma(t) = t^2$.
- For general g_α with $\varphi_\gamma(t)/t$ qualification,

$$\begin{aligned} b_{u^\dagger}(\alpha) &= \|\alpha(\alpha I + B^*B)^{-1}(B^*B)(B^*B)^{-1} r_\alpha(B^*B)\varphi_\gamma(B^*B)w\| \\ &\leq \|\alpha(\alpha I + B^*B)^{-1}B^*B\| \|r_\alpha(B^*B)\varphi_\gamma(B^*B)(B^*B)^{-1}\| \\ &\leq \alpha \frac{\varphi_\gamma(\alpha)}{\alpha} = \varphi_\gamma(\alpha). \end{aligned}$$

Moderately ill-posed operators under Sobolev-type regularity

- $K^*K = \mathcal{A}^{-\ell}, \; u^\dagger \in S^\gamma, \; C_0 = \mathcal{A}^{-\frac{1}{2}-p}, \; m_0 = 0.$
- Applying proposition and combining with existing estimates for posterior spread



- Delayed/removed saturation!

Outline

Conclusions

- Proposition easily applied for eg severely ill-posed operators, analytic-type truth regularity.
- Summary and benefits of preconditioning method in general setting:
  - the user chooses a (centered) Gaussian prior of arbitrary smoothness;
  - after observing data y, a prior center $m_0 = m_\alpha^\delta$, is chosen by some deterministic regularization;
  - if preprocessing regularization has enough qualification, posterior contracts 'optimally' regardless of solution smoothness. If not, contraction rate at least as good as rate for centered prior;
  - preprocessing step has no effect on optimal regularization parameter choice; *any* choice $\alpha = \alpha(\delta; y)$ which yields 'optimal' contraction without preprocessing retains this property, and eventually extends optimality to higher solution smoothness.

Conclusions

- Proposition easily applied for eg severely ill-posed operators, analytic-type truth regularity.
- Summary and benefits of preconditioning method in general setting:
  - the user chooses a (centered) Gaussian prior of arbitrary smoothness;
  - after observing data y, a prior center $m_0 = m_\alpha^\delta$, is chosen by some deterministic regularization;
  - if preprocessing regularization has enough qualification, posterior contracts 'optimally' regardless of solution smoothness. If not, contraction rate at least as good as rate for centered prior;
  - preprocessing step has no effect on optimal regularization parameter choice; *any* choice $\alpha = \alpha(\delta; y)$ which yields 'optimal' contraction without preprocessing retains this property, and eventually extends optimality to higher solution smoothness.

Thank you for your attention!

http://www.sergiosagapiou.com

S. Agapiou, P. Mathé, *Preconditioning the prior to overcome saturation in Bayesian inverse problems*, submitted, arXiv:1409.6496.

S. Agapiou, S. Larsson, A. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stochastic Processes and their Applications, 2013.

S. Agapiou, A. Stuart, Y-X Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, Journal of Inverse and Ill-posed Problems, 2014.

S. Agapiou, J. Bardsley, O. Papaspiliopoulos, A. Stuart, *Analysis of the Gibbs Sampler for Hierarchical Inverse Problems*, SIAM Journal on Uncertainty Quantification, 2014.