Posterior Contraction Rates for Bayesian Inverse Problems

Sergios Agapiou

Mathematics Institute
University of Warwick, UK

Canberra Symposium on Regularization,
19-23 November 2012, Australian National University in Canberra

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


S. Agapiou, S. Larsson and A. M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, http://arxiv.org/abs/1203.5753


Y. Pokern, A. M. Stuart and J. H. Van Zanten, *Posterior consistency via precision operators for nonparametric drift estimation in SDEs*, http://arxiv.org/abs/1202.0976


S. Agapiou, A. M. Stuart and Y. X. Zhang, *An abstract framework for the study of posterior contraction in Bayesian inverse problems*, in preparation.


S. Agapiou, A. M. Stuart and Y. X. Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, http://arxiv.org/abs/1210.1563

Outline

Outline

Bayesian Inverse Problems

- $(X, \langle \cdot, \cdot \rangle, \|\cdot\|)$ separable Hilbert space.
- Probabilistic approach to problem of recovering u from noisy, indirect observations, y.
- *Likelihood*: distribution of $y|u$.
- *Prior*: $u \sim \mu_0$, encoding prior beliefs. Here $\mu_0 = \mathcal{N}(0, \tau^2 \mathcal{C}_0)$.
- *Posterior*: $u|y \sim \mu^y$, object of interest.
- Link: *Bayes'* theorem

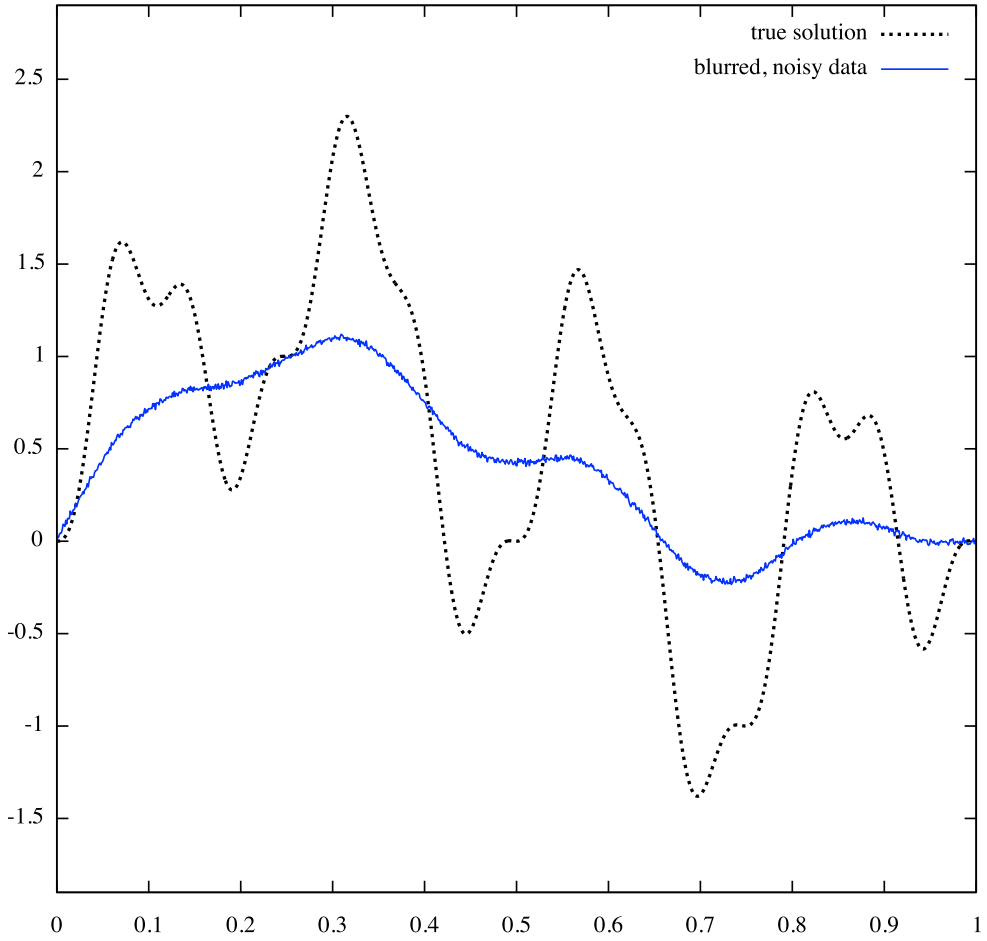
$$\mathbb{P}(u|y) \propto \mathbb{P}(y|u)\mathbb{P}(u)$$

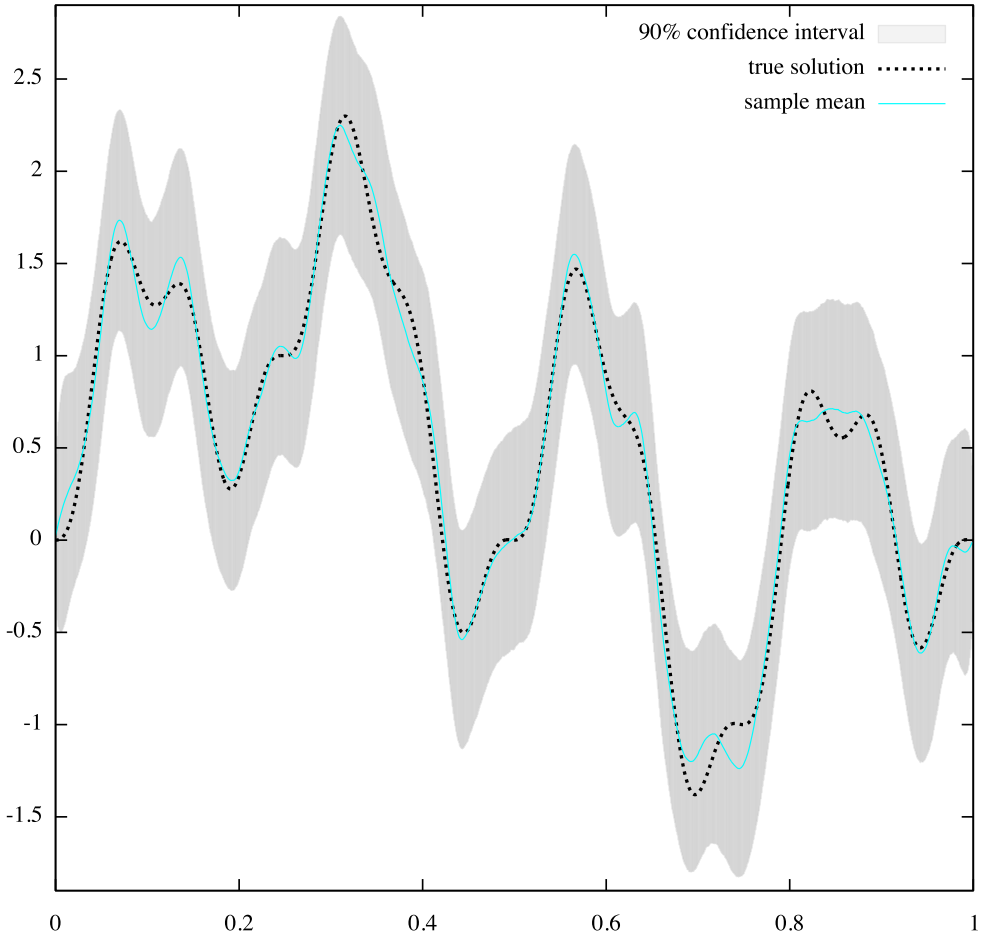
Frequentist Posterior Consistency

- Assume posterior is Gaussian $\mu^y = \mathcal{N}(m, \mathcal{C})$.
- Consider data $y^\dagger$ produced from an underlying truth $u^\dagger$.
- Posterior Consistency: as more data become available, can we recover the truth?
- $y^\dagger = y^\dagger(\epsilon)$ and $\mu^y = \mu^y(\epsilon)$, $\epsilon \to 0$ models improvement in the data (and model).

AIM: Show $\mu^{y^\dagger}(\epsilon) \to \delta_{u^\dagger}$, as $\epsilon \to 0$ for appropriate choice $\tau = \tau(\epsilon)$.

Frequentist Posterior Consistency





Gaussian Measures in Separable Hilbert Spaces

- $\mathcal{N}(m, \mathcal{C})$ in X

  mean $m \in X$;

  covariance $\mathcal{C}$ selfadjoint trace class linear operator in X, eigenpairs $\{\phi_k, \lambda_k\}_{k\in\mathbb{N}}$.

- *Karhunen-Loeve* expansion: $u \sim \mathcal{N}(m, \mathcal{C})$

$$u = m + \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k\phi_k,$$

$\{\xi_k\}_{k\in\mathbb{N}}$ i.i.d. $\mathcal{N}(0, 1)$ in $\mathbb{R}$.

- Decay of λ_k determines regularity of u.

Outline

Linear Inverse Problem

- $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset X \to X$ selfadjoint positive definite linear operator with bounded inverse.
- Find u from noisy observation of $\mathcal{A}^{-1}u$, y.
- Model:

$$y = \mathcal{A}^{-1}u + \frac{1}{\sqrt{n}}\xi,$$

- Interested in *small noise limit*, $n \to \infty$.
- Naive approach $u \approx \mathcal{A}y$, ξ is rough, need to regularize.

Linear Inverse Problem - Bayesian Approach

Assume $\xi \sim \mathcal{N}(0, \mathcal{C}_1)$, $\mathcal{C}_1 : X \to X$ selfadjoint positive definite.

- *Likelihood*: $y|u \sim \mathcal{N}(\mathcal{A}^{-1}u, \frac{1}{n}\mathcal{C}_1)$.
- *Prior*: $u \sim \mathcal{N}(0, \tau^2\mathcal{C}_0)$, $\mathcal{C}_0 : X \to X$ selfadjoint positive definite trace class.
- *Posterior* distribution on $u|y$, μ^y.

Linear Inverse Problem - Assumptions

- *Hilbert Scale* $(X^s)_{s\in\mathbb{R}}$, for $X^s = \mathcal{D}(\mathcal{C}_0^{-\frac{s}{2}})$ with $\langle u, v\rangle_s = \langle \mathcal{C}_0^{-\frac{s}{2}} u, \mathcal{C}_0^{-\frac{s}{2}} v\rangle$.
- Assumptions

$$\exists s_0 \in [0,1) \ \text{ s.t. } \ \mathrm{Tr}\left(\mathcal{C}_0^{\mathrm{s}}\right) < \infty \ \ \forall \mathrm{s} > \mathrm{s}_0;$$

$$\mathcal{C}_1 \simeq \mathcal{C}_0^{\beta}, \quad \beta \geq 0;$$

$$\mathcal{A}^{-1} \simeq \mathcal{C}_0^{\ell}, \quad \ell > 0.$$

- $\Delta := 1 + 2\ell - \beta > 2s_0$. For simplicity $2\ell - \beta \geq 0$.

Linear Inverse Problem - Posterior Identification

Theorem (A., Larsson, Stuart 12)

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z(y)} \exp\left(-\Phi(u;y)\right)$$

where

$$\Phi(u;y) = \frac{n}{2}\left\|\mathcal{C}_1^{-\frac{1}{2}}(y - \mathcal{A}^{-1}u)\right\|^2 - \frac{n}{2}\underbrace{\left\|\mathcal{C}_1^{-\frac{1}{2}}y\right\|^2}.$$

In particular, $\mu^y = \mathcal{N}(m, \frac{1}{n}\mathcal{B}_\lambda^{-1})$

$$\mathcal{B}_\lambda = \mathcal{A}^{-1}\mathcal{C}_1^{-1}\mathcal{A}^{-1} + \lambda\mathcal{C}_0^{-1}$$

$$\mathcal{B}_\lambda m = \mathcal{A}^{-1}\mathcal{C}_1^{-1}y,$$

$\lambda = \frac{1}{n\tau^2}$ regularization parameter.

Linear Inverse Problem - Posterior Consistency

- Consider

$$y^\dagger = \mathcal{A}^{-1}u^\dagger + \frac{1}{\sqrt{n}}\xi,$$

$u^\dagger \in X^\gamma$ fixed true solution, γ a-priori known.

- Posterior $\mu^{y=y^\dagger} = \mathcal{N}(m^\dagger, \frac{1}{n}\mathcal{B}_\lambda^{-1})$.

- AIM: Find optimal rate $=$ rate(γ), such that, as $n \to \infty$,

$$\mathbb{E}\left\|m^\dagger - u^\dagger\right\|^2 + \mathrm{Tr}\,(\frac{1}{\mathrm{n}}\mathcal{B}_\lambda^{-1}) = \mathcal{O}(\mathrm{n}^{-\mathrm{rate}}).$$

Linear Inverse Problem - Error Equation

- Mean

$$\mathcal{B}_\lambda m^\dagger = \underbrace{\mathcal{A}^{-1}\mathcal{C}_1^{-1}\mathcal{A}^{-1}u^\dagger} + \frac{1}{\sqrt{n}}\mathcal{A}^{-1}\mathcal{C}_1^{-1}\xi.$$

- Truth

$$\mathcal{B}_\lambda u^\dagger = \underbrace{\mathcal{A}^{-1}\mathcal{C}_1^{-1}\mathcal{A}^{-1}u^\dagger} + \lambda\mathcal{C}_0^{-1}u^\dagger.$$

- Error $e = m^\dagger - u^\dagger$

$$e = \mathcal{B}_\lambda^{-1}\left(\frac{1}{\sqrt{n}}\mathcal{A}^{-1}\mathcal{C}_1^{-1}\xi - \lambda\mathcal{C}_0^{-1}u^\dagger\right).$$

Linear Inverse Problem - Error Analysis

Lemma (A., Larsson, Stuart 12)

For $\eta = (1-\theta)(\beta - 2\ell) + \theta,\ \theta \in [0,1]$

$$\left\| \mathcal{B}_\lambda^{-1} \right\|_{\mathcal{L}(X^{-\eta}, X)} = \mathcal{O}(\lambda^{-c}),$$

$c = c(\theta, \beta - 2\ell) \in (0,1)$, increasing in θ.

- For $\eta_i = (1-\theta_i)(\beta - 2\ell) + \theta_i$, where $\theta_i \in [0,1]$ sufficiently large,

$$\mathbb{E} \left\| \frac{1}{\sqrt{n}} \mathcal{A}^{-1} \mathcal{C}_1^{-1} \xi \right\|_{-\eta_1} = \mathcal{O}(n^{-\frac{1}{2}})$$

and

$$\left\| \lambda \mathcal{C}_0^{-1} u^\dagger \right\|_{-\eta_2} = \mathcal{O}(\lambda).$$

$$e = \mathcal{O}(\lambda^{-c_1} n^{-\frac{1}{2}}) + \mathcal{O}(\lambda^{1-c_2}).$$

Linear Inverse Problem - Convergence

- Choose $\lambda = \lambda(n)$ (hence $\tau = \tau(n)$) optimally.
- $\mathrm{Tr}\left(\frac{1}{n}\mathcal{B}_\lambda^{-1}\right)$ bounded by noise term.

Theorem (A., Larsson, Stuart 12)

Assume $u^\dagger \in X^\gamma$, $\gamma \geq 1$. For appropriate choice of $\lambda = \lambda(n) \to 0$ we have as $n \to \infty$

$$\text{rate} = \begin{cases} \frac{\gamma}{2(\Delta+\gamma-1+s_0+\delta)}, & \text{if } \gamma \in [1, \Delta+1] \\ \frac{\Delta+1}{2(2\Delta+s_0+\delta)}, & \text{if } \gamma > \Delta+1, \end{cases}$$

$\delta > 0$ arbitrarily small.

Linear Inverse Problem - Optimality, Diagonal Case

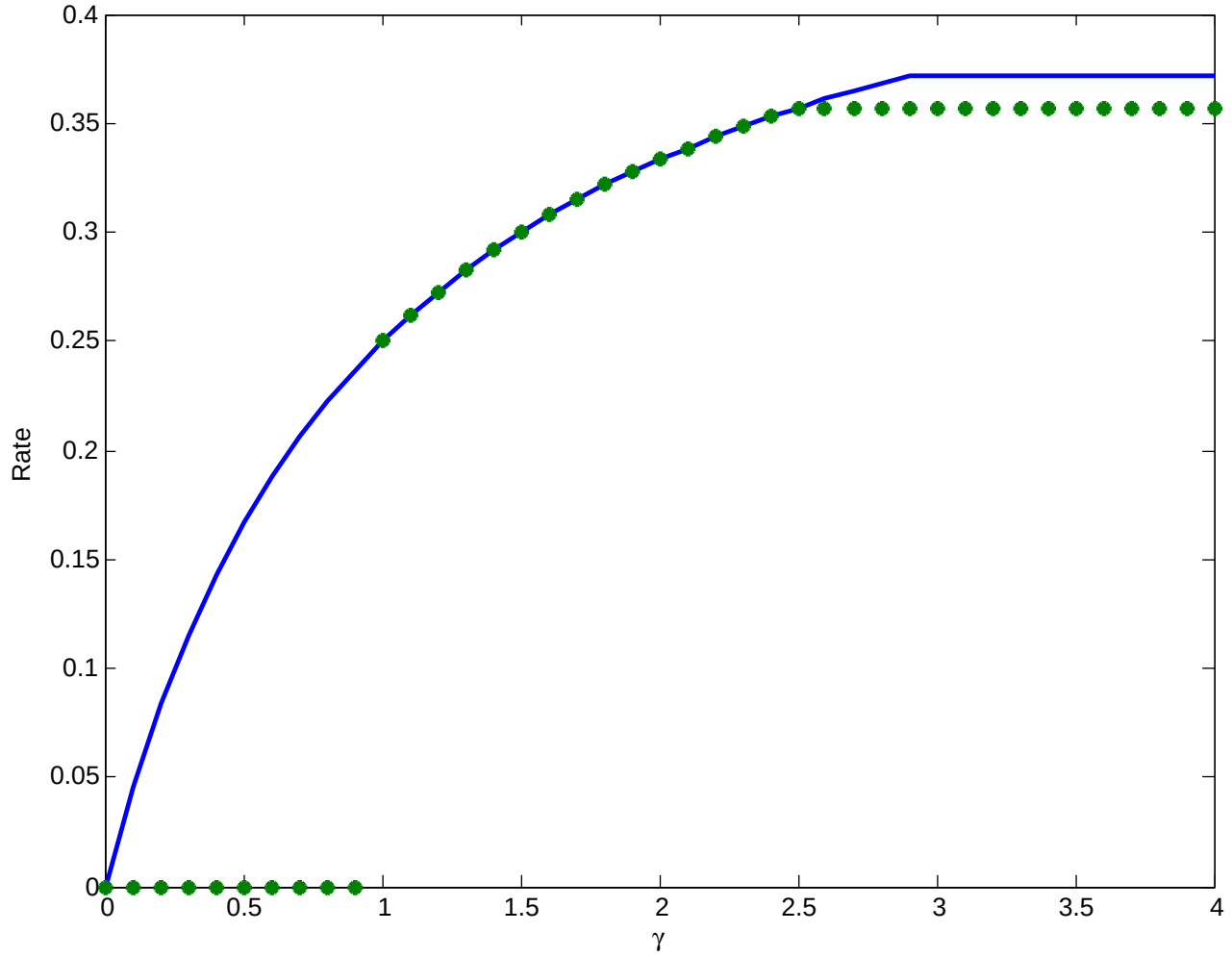
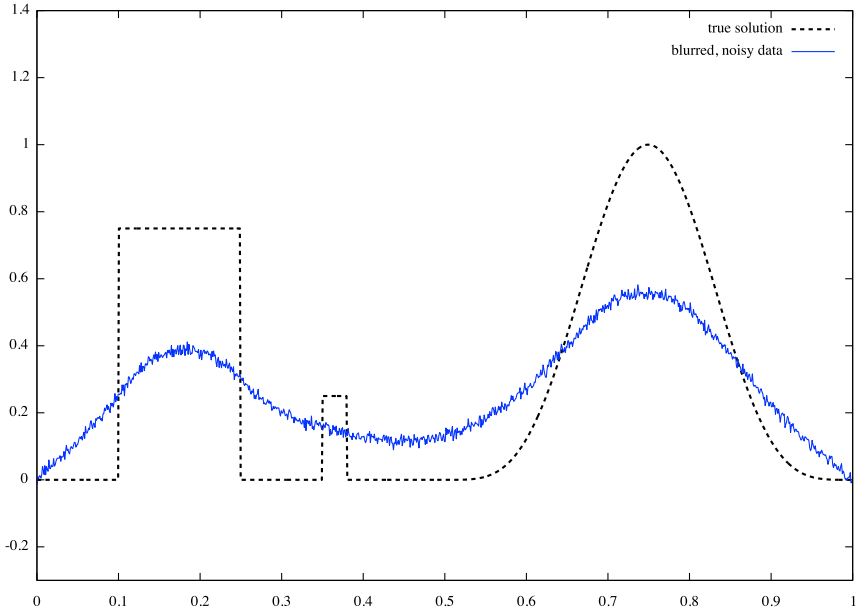
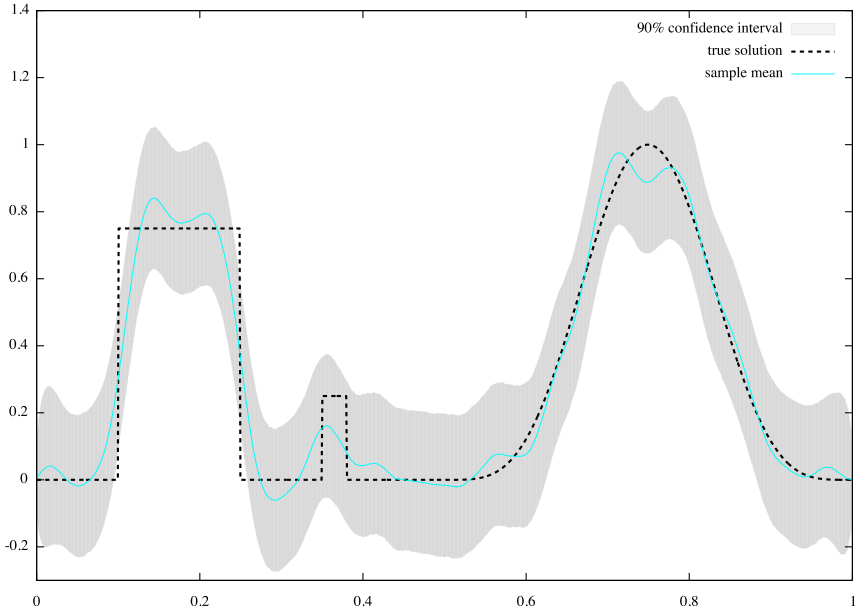
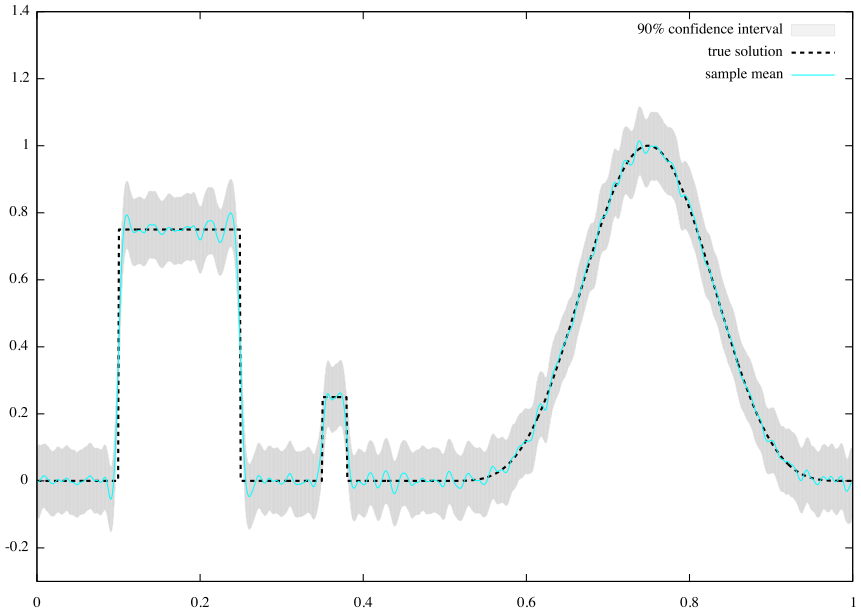


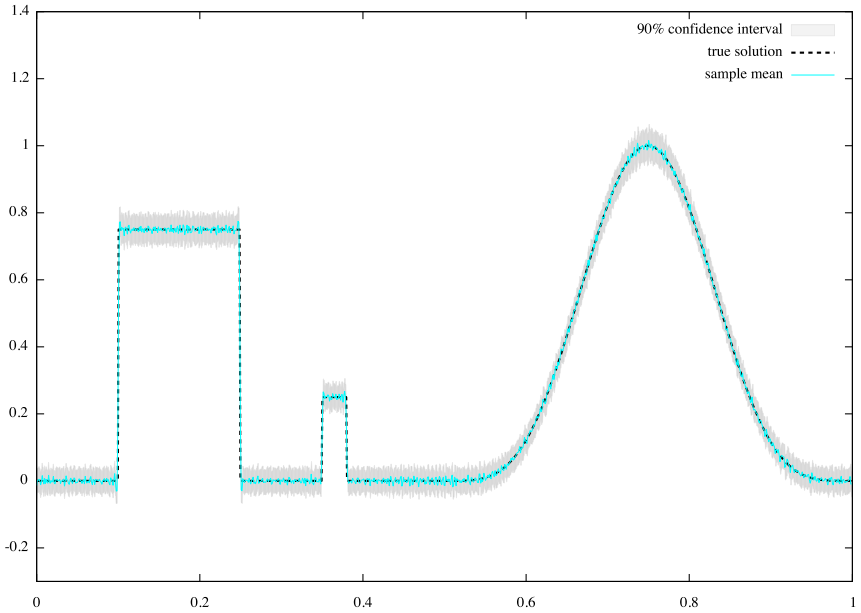
Figure: blue=diagonal analysis and green=our more general analysis for $\mathcal{C}_0^{-1} = -\frac{d^2}{dx^2}$, $\beta = \ell = \frac{1}{2}$, so that $s_0 = \frac{1}{2}$, $\Delta = \frac{3}{2}$ and rate $= \frac{\gamma}{2+2\gamma+\delta}$, $\gamma \in [1, 2.5]$.

Linear Inverse Problem - Posterior Consistency









Outline

Drift Estimation in SDEs

- Consider the SDE

$$dY_t = u(Y_t)dt + dW_t, \quad Y_0 = 0,$$

where $u \in C^1(\mathbb{T}), \quad \mathbb{T} = [0, 1)$ periodic.

- Find drift u from observations $Y = \{Y_t\}_{t\in[0,T]}$.
- Interested in the *long observation time limit*, $T \to \infty$.

Drift Estimation in SDEs - Intuition

- For $t \to \infty$, Y_t reaches equilibrium in form of invariant measure with density ρ

$$u(x) = \frac{\rho'(x)}{2\rho(x)}$$

- *Local time* $L_T(x; Y)$ measures time that Y spends around x.

Lemma (Van Zanten 12)

For every $\alpha < 1/2$,

$$\left\| \frac{1}{T} L_T(\cdot; Y) - \rho \right\|_{H^\alpha} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{T}}\right)$$

as $T \to \infty$.

- i.e. $u \approx \frac{L'_T(.;Y)}{2L_T(.;Y)}$ but in weak spaces, regularize.

Drift Estimation in SDEs - Bayesian Approach

- *Likelihood*: law of $Y|u$ (*Girsanov*)
- *Prior*: $u \sim \mathcal{N}(0, \tau^2 \mathcal{C}_0)$,

$$\mathcal{C}_0^{-1} = \left(-\frac{d^2}{dx^2}\right)^p + I,$$

  $p \in 2, 3, \ldots$; $p \geq 2$ secures $u \in C^1(\mathbb{T})$ a.s. (Gaussianity).
- Joint distribution of Y and u in general non-Gaussian.
- *Posterior:* $u|Y \sim \mu^Y$ Gaussian.

Drift Estimation in SDEs - Posterior Identification

Theorem (Pokern, Stuart, Van Zanten 12)

The posterior is Gaussian, $\mu^y = \mathcal{N}(m, \frac{1}{T}\mathcal{B}_\lambda^{-1})$

$$\mathcal{B}_\lambda = \mathcal{B}_\lambda(Y, T) = \frac{L_T(\cdot; Y)}{T} I + \lambda \mathcal{C}_0^{-1}$$

$$\mathcal{B}_\lambda m = \frac{1}{2}\frac{L_T(\cdot; Y)'}{T} + \frac{\chi_T(\cdot; Y)}{T}$$

$\lambda = \frac{1}{\tau^2 T}$ regularization parameter.

Drift Estimation in SDEs - Posterior Consistency

- Consider $Y^\dagger = \{Y_t^\dagger\}_{t\in[0,T]}$ where

$$dY_t^\dagger = u^\dagger(Y_t^\dagger)dt + dW_t$$

$u^\dagger \in H^\gamma$ fixed true drift, γ a-priori known.

- Posterior $\mu^{Y=Y^\dagger} = \mathcal{N}(m^\dagger, \frac{1}{T}\mathcal{B}_\lambda^{-1})$, where $\mathcal{B}_\lambda = \mathcal{B}_\lambda(Y^\dagger)$.

- AIM: Find optimal rate $=$ rate(γ), such that, as $T \to \infty$,

$$\left\|m^\dagger - u^\dagger\right\|^2 + \mathrm{Tr}\left(\frac{1}{\mathrm{T}}\mathcal{B}_\lambda^{-1}\right) = \mathcal{O}_{\mathbb{P}}(\mathrm{T}^{-\mathrm{rate}}).$$

Drift Estimation in SDEs - Error Equation

- Error $e = m^\dagger - u^\dagger$

$$e = \mathcal{B}_\lambda^{-1}\left(\frac{1}{2}\left(\frac{L_T(\cdot;Y^\dagger)}{T} - \rho^\dagger\right)' - \lambda\mathcal{C}_0^{-1}u^\dagger - \left(\frac{L_T(\cdot;Y^\dagger)}{T} - \rho^\dagger\right)u^\dagger + \frac{\chi_T(\cdot;Y^\dagger)}{T}\right).$$

- As $\lambda \to 0$, $\left\|\mathcal{B}_\lambda^{-1}\right\|_{\mathcal{L}(H^{-\theta},H)} = \mathcal{O}_{\mathbb{P}}(\lambda^{-\frac{\theta}{2}}), \quad \theta \in [0,1].$

- As $\lambda \to 0$ and $T \to \infty$, terms in parenthesis go to zero in different sufficiently weak spaces.

- Choose $\lambda = \lambda(T)$ (hence $\tau = \tau(T)$) optimally.

Drift Estimation in SDEs - Convergence

Theorem (Pokern, Stuart, Van Zanten 12; slightly improved by A., Stuart, Zhang)

Assume $u^\dagger \in H^\gamma$, $\gamma \geq p$. For appropriate choice $\lambda = \lambda(T) \to 0$ we have as $T \to \infty$

$$\text{rate} = \begin{cases} \frac{\gamma}{1+2\gamma+\delta}, & \text{if } \gamma \in [p, 2p] \\ \frac{2p}{1+4p+\delta}, & \text{if } \gamma > 2p, \end{cases}$$

$\delta > 0$ arbitrarily small.

Drift Estimation in SDEs - Convergence

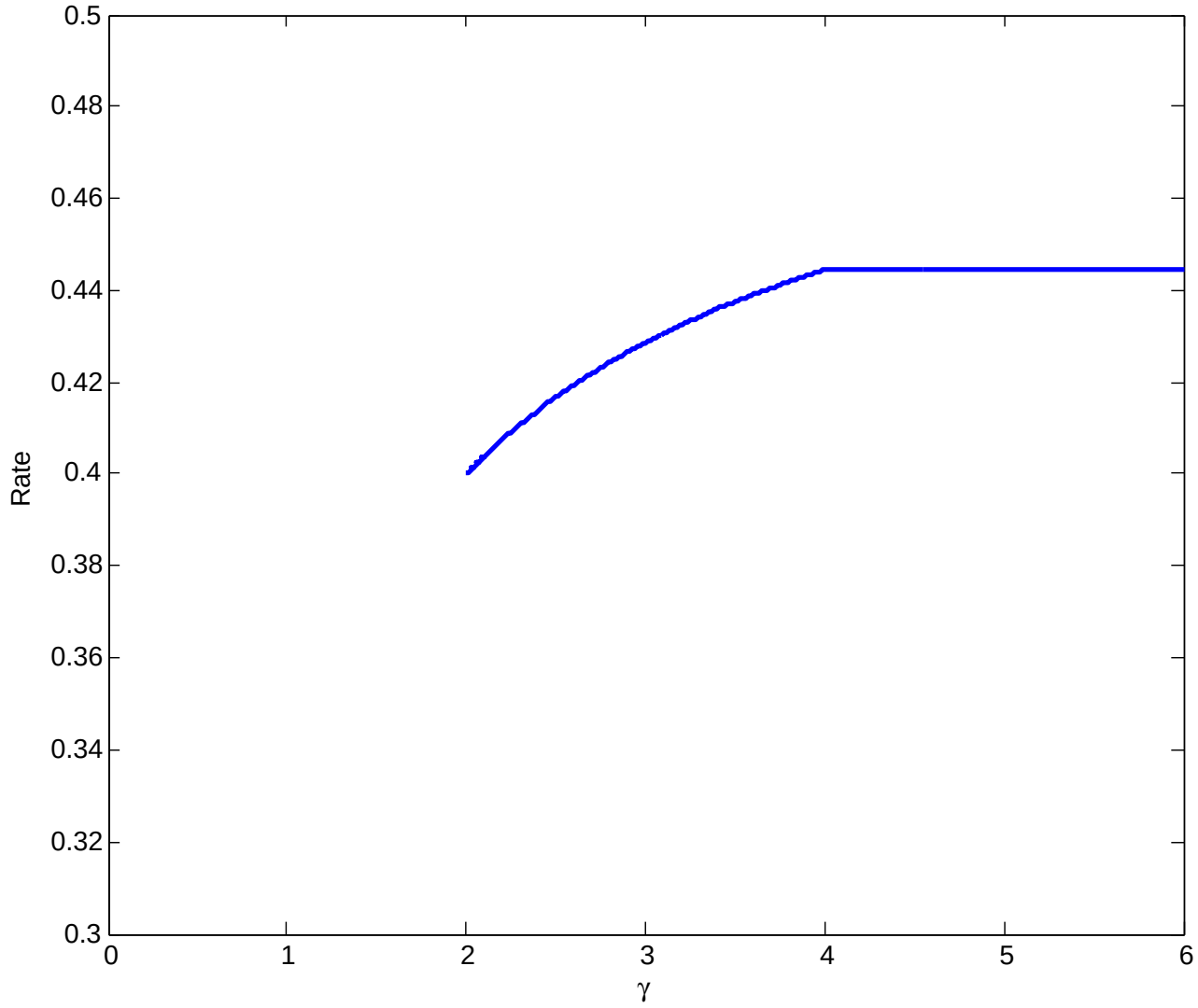


Figure: contraction for $\mathcal{C}_0^{-1} = (-\frac{d^2}{dx^2})^2 + I$ so that rate $= \frac{\gamma}{1+2\gamma+\delta}$, $\gamma \in [2, 4]$.

Outline

Abstract Theory

- *Prior*: $\mu_0 = \mathcal{N}(0, \tau^2 \mathcal{C}_0)$, $\mathcal{C}_0$ selfadjoint positive definite trace class in X.
- X^s Hilbert scale induced by $\mathcal{C}_0^{-\frac{1}{2}}$.
- *Posterior*: assume $\mu^y = \mathcal{N}(m, \epsilon \mathcal{B}_\lambda^{-1})$,

$$\mathcal{B}_\lambda = \mathcal{B}_\lambda(y, \epsilon) = Q(y; \epsilon) + \lambda \mathcal{C}_0^{-1}$$

$$\mathcal{B}_\lambda m = r(y; \epsilon),$$

$\lambda = \frac{\epsilon^2}{\tau^2}$.

- Interested in $\epsilon \to 0$ modelling improvement in data.

Abstract Theory - Meta Theorem

- Consider data $y^\dagger = y^\dagger(\epsilon)$ produced from underlying truth $u^\dagger \in X^\gamma$.
- Assume $r(y^\dagger; \epsilon) \approx \hat{r}$ and $Q(y^\dagger; \epsilon) \approx \hat{Q}$, where $u^\dagger = \hat{Q}^{-1}\hat{r}$.
- $\mu^{y=y^\dagger} = \mathcal{N}(m^\dagger, \epsilon\mathcal{B}_\lambda^{-1})$, where $\mathcal{B}_\lambda = \mathcal{B}_\lambda(y^\dagger)$.

Theorem

Find and optimize rate=rate(γ) such that as $\epsilon \to 0$,

$$\left\|m^\dagger - u^\dagger\right\|^2 + \mathrm{Tr}\left(\epsilon\mathcal{B}_\lambda^{-1}\right) = \mathcal{O}(\epsilon^{\mathrm{rate}}),$$

in some probabilistic sense since $y^\dagger$ random.

Abstract Theory - Method

- Error $e = m^\dagger - u^\dagger$ satisfies (Lax-Milgram in X^1)

$$e = \mathcal{B}_\lambda^{-1}\left((\hat{r} - r(y^\dagger;\epsilon)) + (Q(y^\dagger;\epsilon) - \hat{Q})u^\dagger + \lambda \mathcal{C}_0^{-1}u^\dagger\right).$$

- Rate of convergence determined by γ and
  - $\left\|\hat{r} - r(y^\dagger;\epsilon)\right\|_{-\alpha} = \mathcal{O}(\epsilon^a)$ as $\epsilon \to 0$;
  - $\left\|\hat{Q} - Q(y^\dagger;\epsilon)\right\|_{op} = \mathcal{O}(\epsilon^b)$ as $\epsilon \to 0$;
  - $\left\|\mathcal{B}_\lambda^{-1}\right\|_{op} = \mathcal{O}(\lambda^{-c})$ as $\lambda \to 0$.
- $\mathrm{Tr}\left(\epsilon\mathcal{B}_\lambda^{-1}\right)$ often dominated by terms arising in $\left\|m^\dagger - u^\dagger\right\|^2$.
- Rate optimized by choosing $\lambda = \lambda(\epsilon)$ (inducing choice $\tau = \tau(\epsilon)$).

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


S. Agapiou, S. Larsson and A. M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, http://arxiv.org/abs/1203.5753


Y. Pokern, A. M. Stuart and J. H. Van Zanten, *Posterior consistency via precision operators for nonparametric drift estimation in SDEs*, http://arxiv.org/abs/1202.0976


S. Agapiou, A. M. Stuart and Y. X. Zhang, *An abstract framework for the study of posterior contraction in Bayesian inverse problems*, in preparation.


S. Agapiou, A. M. Stuart and Y. X. Zhang, *Bayesian posterior contraction rates for linear severely ill-posed inverse problems*, http://arxiv.org/abs/1210.1563