Adaptive posterior contraction results for Bayesian methods for diffusions

Jan van Waaij, Humboldt University of Berlin

Harry van Zanten
(University of Amsterdam)

Frank van der Meulen
(Delft University of Technology)

Moritz Schauer
(Leiden University)

Slides are on https://github.com/Jan-van-Waaij/wias

Bayesian vs. frequentist statistics

- Frequentist statistics: one true parameter, recover the parameter from the data.
- Bayesian statistics: no 'true' parameter. Believe expressed in probability distribution on the parameter space.
- 'Frequentist behaviour' of Bayesian methods.
- Bayesian estimators:
  - Posterior mean
  - Posterior mode
- Bayesian uncertainty quantification: credible sets

Parametric vs. nonparametric situation

- 🙂 Parametric Bayesian models: $\sqrt{n}$-consistent.
- 😐 Nonparametric models:
  - Freedman & Diaconis: inconsistent posteriors are possible.
  - There is a gap between priors with good theoretical performance and priors with good numerical performance.

Frequentist analysis of Bayesian methods

- Consistency
- Posterior contraction rates
- Coverage of credible sets
- Bernstein-von Mises 'central limit' theorems

Subject of today

- Posterior contraction rates for diffusion processes.
- Adaptation to unknown smoothness of the function.
- Empirical Bayes.

Statistical inference for diffusions

- Diffusions on the line: real-valued strong Markov processes with continuous paths,
- Under weak conditions a diffusion is described via an SDE

$$dX_t = \theta(X_t)dt + \sigma(X_t)dW_t,$$

- We assume $\sigma \equiv 1$,
- $\theta : \mathbb{R} \to \mathbb{R}$ is measurable, 1-periodic and $\int_0^1 \theta(x)^2 dx < \infty$.
- Observations $X^T = \{X_t : t \in [0, T]\}$ of

$$dX_t = \theta(X_t)dt + dW_t,$$

- Goal: estimate θ.

Key ingredients for posterior convergence

$$\mathbb{E}_{\theta_0}\, \Pi(\{\theta : \|\theta - \theta_0\|_2 \geq \varepsilon_T\} \mid X^T) \to 0$$

$$\|$$

$$\mathbb{E}_{\theta_0}\left[\frac{\int_{\{\theta\in\Theta_T:\|\theta-\theta_0\|_2\geq\varepsilon_T\}} p_\theta(X^T)d\Pi(\theta)}{\int p_\theta(X^T)d\Pi(\theta)}\right]$$

$$= \mathbb{E}_{\theta_0}\left[\frac{\int_{\{\theta\in\Theta_T:\|\theta-\theta_0\|_2\geq\varepsilon_T\}} p_\theta(X^T)d\Pi(\theta)}{\int p_\theta(X^T)d\Pi(\theta)}\right] + \Pi(\Theta_T^c \mid X^T)$$

$$\leq \frac{e^{-CT\varepsilon_T^2}}{e^{-cT\varepsilon_T^2}} + \mathbf{o}(1)$$

1. Tests,
2. Enough prior mass around true parameter.
3. Model is not too big.

Posterior convergence

When

$$\Pi(\theta : \|\theta - \theta_0\| < \varepsilon_T) \geq e^{-\xi T \varepsilon_T^2},$$

For every $K > 0$, there are measurable sets Θ_T so that

$$\Pi(\Theta_T) = 1 - e^{-KT\varepsilon_T^2}$$

and for every $a \in (0, 1)$, there is a $C > 0$

$$N\left(a\varepsilon_T, \{\theta \in \Theta_T : \|\theta - \theta_0\|_2 < \varepsilon_T\}\right) \leq e^{CT\varepsilon_T^2},$$

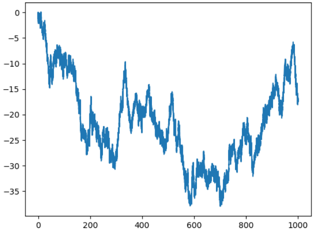
then for some $M > 0$,

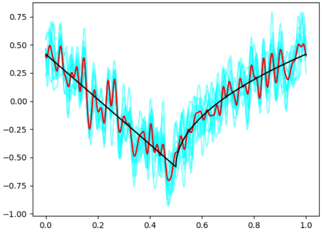
$$\mathbb{E}_{\theta_0} \Pi(\theta : \|\theta - \theta_0\|_2 \leq M\varepsilon_T \mid X^T) \to 1, \text{ as } T \to \infty.$$

Example

- SDE $dX_t = \theta_0(X_t)dt + dW_t$,
- Prior:

$$\theta = \sum_{k=1}^{100} k^{-1} Z_k \phi_k$$

- 



Posterior contraction for the Gaussian process prior

- Continuous stochastic process with fdd are multivariate Gaussian.
- Prior:

$$\theta = \sum_{k=1}^{\infty} k^{-\alpha-1/2} Z_k \phi_k.$$

- Posterior = Gaussian process + explicit formula
- True function $\theta_0 = \sum_{k=1}^{\infty} \theta_k \phi_k$ satisfies $\sum_{k=1}^{\infty} k^{2\beta} \theta_k^2 < \infty$.
- Minimax posterior convergence rate $T^{-\frac{\beta}{1+2\beta}}$ if and only if $\alpha = \beta$.
- Not adaptive!

Solution I: Hierarichal Bayes

- Let the posterior choose the right smoothness.
- Hyperpriors on the hyperparameters.

$$1 \sum_{k=1}^{\infty} k^{-\alpha-1/2} Z_k \phi_k$$

- Requirements:
  - Hyperprior should give enough mass range of optimal hyperparameters (easy).
  - Remaining mass condition (hard).

Scaling parameter

- Hyperprior on the scaling.

$$
\begin{aligned}
E &\sim \mathrm{Exp}(1), \\
S &= \frac{E^{1/2+\alpha}}{\sqrt{T}}, \\
\theta \mid S &= S \sum_{k=1}^{\infty} k^{-\alpha-1/2} Z_k \phi_k
\end{aligned}
$$

- When $0 < \beta \leq \alpha + 1/2$, and $\theta \in H^{\beta}$, then the posterior contracts with rate $T^{-\frac{\beta}{1+2\beta}}$.
- 😐 Optimal rates for the most import range, suboptimal for supersmooth functions.
- ☹ Prior on S intricate.

Prior on the baseline smoothness

- Hyperprior on α

$$\pi(\alpha) \propto e^{-T^{\frac{1}{1+2\alpha}}}, \alpha \in (0, \log T]$$

$$\theta \mid \alpha = \sum_{k=1}^{\infty} k^{-\alpha-1/2} Z_k \phi_k$$

- ☺ Adaptivity to every Sobolev smoothness!
- ☹ Are other possibilities on α possible?

A superior solution!

- $$\begin{aligned} J &\sim \text{geometric},\\ S^2 &\sim \text{inverse gamma},\\ \theta \mid J, S &\sim S \sum_{j=1}^{J} j^{-1/2-\alpha} Z_k \phi_k. \end{aligned}$$
- When $0 < \beta \leq \alpha + 1/2$ the posterior contracts with rate $T^{-\frac{\beta}{1+2\beta}}$ and when $\beta > \alpha + 1/2$, the posterior contracts with rate $\left(\frac{T}{\log T}\right)^{-\frac{\beta}{1+2\beta}}$.
- (Nearly) optimal rates for every smoothness.
- Has good numerical properties!


It's Tea Time!

Empirical Bayes

- Use prior Π_s

$$\theta = s \sum_{k=1}^{\infty} k^{-1/2-\alpha} Z_k \phi_k,$$

- estimate s from the data,
- use

$$\Pi_{\hat{s}} = \Pi_s(\cdot \mid X^T)\Big|_{s=\hat{s}}$$

for the inference.

Hierarchical vs. empirical Bayes

- ☹ Gaussian process prior is not adaptive.
- Hierarchical Bayes solution: equip hyperparameters with additional prior.
  - ☹ Prior is not Gaussian
  - ☺ Adaptivity + (near) optimal rates.
- Empirical Bayes solution: estimate hyperparameter from the data and use plug-in posterior for inference.
  - ☺ the (data-driven) prior still Gaussian.
  - 😐 But a lot unknown about the theoretical and computational performance...
  - The analysis is considerably harder.

- Prior Π_s defined by $s\sum_{k=1}^{\infty} k^{-\alpha-1/2} Z_k \phi_k$.
- "Prior behaves best when it puts a lot of prior mass around θ_0."
- That is when $s \asymp T^{\frac{\alpha-\beta}{1+2\beta}}$ for $0 < \beta \leq \alpha + 1/2$.
- Optimise $\Pi_s\left(\|\theta - \theta_0\|_2 \leq \varepsilon_T\right)$ over

$$\Lambda = \left\{ kT^{-\frac{1}{4+4\alpha}} : k \in \mathbb{N}, kT^{-\frac{1}{4+4\alpha}} \leq T^{\alpha} \right\}.$$

- Marginal maximum likelihood estimator (MMLE)

$$\hat{s} = \mathrm{argmax}_{s\in\Lambda} \int p_\theta(X^T) d\Pi_s(\theta),$$

$$p_\theta(X^T) = \exp\left\{ \int_0^T \theta(X_t) dX_t - \frac{1}{2}\int_0^T \theta(X_t)^2 dt \right\}$$

Theorem

When θ_0 is β-Sobolev smooth, $0 < \beta \leq \alpha + 1/2$, then for some $M > 0$,

$$\Pi_{\hat{s}} \left(\theta : \|\theta - \theta_0\|_2 \leq M T^{-\frac{\beta}{1+2\beta}} \;\middle|\; X^T \right) \to 1$$

in $\mathbb{P}_{\theta_0}$-probability as $T \to \infty$.

Outline of the proof

Show $\Pi_{\hat{s}}\left(\|\theta-\theta_0\| > MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right) \to 0$.

Ingredients of the proof:

1. Determine $\Lambda_0 \subseteq \Lambda$ where $\Pi_s(\cdot \mid X^T)$ enjoys good rates.
2. $\mathbb{P}_{\theta_0}(\hat{s} \in \Lambda_0) \to 1$, as $T \to \infty$,
3.

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left(\Pi_{\hat{s}}\left(\|\theta-\theta_0\|_2 \geq MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right)\right) \\
\leq &\mathbb{E}_{\theta_0}\left(\Pi_{\hat{s}}\left(\|\theta-\theta_0\|_2 \geq MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right)\mathbb{I}_{\{\hat{s}\in\Lambda_0\}}\right) \\
&+\mathbb{E}_{\theta_0}\left(\Pi_{\hat{s}}\left(\|\theta-\theta_0\|_2 \geq MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right)\mathbb{I}_{\{\hat{s}\notin\Lambda_0\}}\right) \\
\leq &\mathbb{E}_{\theta_0}\left(\max_{s\in\Lambda_0}\Pi_{s}\left(\|\theta-\theta_0\|_2 \geq MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right)\right) \\
&+\mathbb{P}_{\theta_0}(\hat{s}\notin\Lambda_0) \to 0.
\end{aligned}$$

Determining Λ_0

Let $K > 0$ be constant. There is a unique $\varepsilon_s > 0$ so that

$$\Pi_s(\|\theta - \theta_0\|_2 < K\varepsilon_s) = e^{-T\varepsilon_s^2}.$$

Let

$$\varepsilon_0 = \min_{s \in \Lambda} \varepsilon_s.$$

Let $L > 1$ be a constant and

$$\Lambda_0 = \left\{ s \in \Lambda : \varepsilon_s \leq L\varepsilon_0 \right\}.$$

Lemma

For $L > 1$ big enough, with $\mathbb{P}_{\theta_0}$-probability converging to one $\hat{s} \in \Lambda_0$.

Step 1 Take p_θ / p_{θ_0} instead.

$$\operatorname{argmax}_{s\in\Lambda} \int p_\theta(X^T) d\Pi_s(\theta)$$
$$= \operatorname{argmax}_{s\in\Lambda} \int p_\theta(X^T)/p_{\theta_0}(X^T) d\Pi_s(\theta).$$

Step 2 Let $s_0 \in \Lambda, \varepsilon_{s_0} = \varepsilon_0$. There are constants $0 < A < B$ so that with $\mathbb{P}_{\theta_0}$-probability converging to one,

$$\int p_\theta(X^T)/p_{\theta_0}(X^T) d\Pi_{s_0}(\theta) \geq e^{-AT\varepsilon_0^2}$$
$$> e^{-BT\varepsilon_0^2} \geq \max_{s\in\Lambda\setminus\Lambda_0} \int p_\theta(X^T)/p_{\theta_0}(X^T) d\Pi_s(\theta)$$

From

$$
\begin{aligned}
&\int p_\theta(X^T)/p_{\theta_0}(X^T)\,d\Pi_{\hat{s}}(\theta) \\
\geq &\int p_\theta(X^T)/p_{\theta_0}(X^T)\,d\Pi_{s_0}(\theta) \geq e^{-AT\varepsilon_0^2} \\
> &e^{-BT\varepsilon_0^2} \geq \max_{s\in\Lambda\setminus\Lambda_0} \int p_\theta(X^T)/p_{\theta_0}(X^T)\,d\Pi_s(\theta)
\end{aligned}
$$

follows that $\hat{s} \in \Lambda_0$ (on this event).

Goal: show that

$$\mathbb{P}_{\theta_0}\left(\max_{s\in\Lambda\backslash\Lambda_0}\int p_\theta(X^T)/p_{\theta_0}(X^T)d\Pi_s(\theta)\geq e^{-BT\varepsilon_0^2}\right)\to 0.$$

$$\begin{aligned}&\mathbb{P}_{\theta_0}\left(\max_{s\in\Lambda\backslash\Lambda_0}\int p_\theta(X^T)/p_{\theta_0}(X^T)d\Pi_s(\theta)\geq e^{-BT\varepsilon_0^2}\right)\\ \leq\, & T^{\alpha+\frac{1}{4+4\alpha}}\max_{s\in\Lambda\backslash\Lambda_0}\mathbb{P}_{\theta_0}\left(\int p_\theta(X^T)/p_{\theta_0}(X^T)d\Pi_s(\theta)\geq e^{-BT\varepsilon_0^2}\right).\end{aligned}$$

Let $s \in \Lambda \backslash \Lambda_0$. Consider

$$
\begin{aligned}
&\mathbb{P}_{\theta_0}\left(\int p_\theta(X^T)/p_{\theta_0}(X^T) d\Pi_s(\theta) \geq e^{-BT\varepsilon_0^2}\right) \\
=& \mathbb{E}_{\theta_0}\left[\mathbb{I}\left\{\int p_\theta(X^T)/p_{\theta_0}(X^T) d\Pi_s(\theta) \geq e^{-BT\varepsilon_0^2}\right\}(\varphi + 1 - \varphi)\right] \\
\leq& \mathbb{E}_{\theta_0}\varphi + e^{BT\varepsilon_0^2}\int \mathbb{E}_\theta[1-\varphi] d\Pi_s(\theta).
\end{aligned}
$$

Use $\varepsilon_s \geq L\varepsilon_0$ for $s \in \Lambda \backslash \Lambda_0$.

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}\varphi \\
\leq& e^{-C_1 T\varepsilon_s^2} \\
\leq& e^{-C_1 L^2 T\varepsilon_0^2}.
\end{aligned}
$$

$$
\begin{aligned}
&\int \mathbb{E}_\theta[1-\varphi] d\Pi_s(\theta) \\
\leq& \int_{\|\theta-\theta_0\| \leq K\varepsilon_s} d\Pi_s(\theta) \\
&+ \int_{\|\theta-\theta_0\| > K\varepsilon_s} \mathbb{E}_\theta[1-\varphi] d\Pi_s(\theta) \\
\leq& e^{-T\varepsilon_s^2} + e^{-C_2 T\varepsilon_s^2} \\
\leq& e^{-L^2 T\varepsilon_0^2} + e^{-C_2 L^2 T\varepsilon_0^2}
\end{aligned}
$$

Future work

- The asymptotic behaviour of credible sets.
- “Empirical Bayes” as tool to show rates for hierarchical Bayes priors.
- Simulation studies.

Thank you!