Posterior contraction rates for constrained deep Gaussian processes in density estimation

François Bachoc and Agnès Lagnoux

Institut de Mathématiques de Toulouse
France

JDS Lyon

June 2022

Bayesian framework in density estimation

Bayesian framework

- Fixed unknown density function $p_0 : [-1, 1]^d \to [0, \infty)$.
- We observe random $X_1, \dots, X_n$, iid with density p_0.
- Bayesian prior P_0 on p_0. P_0 is a random density : $[-1, 1]^d \to [0, \infty)$.
  - Not the same randomness as for $X_1, \dots, X_n$.
- Bayes' rule provides a posterior distribution on p_0 :

$$\mathbb{P}\left(P_0 \in \cdot \,|\, X_1, \dots, X_n \right).$$

Frequentist analysis

- The posterior distribution provides an estimator of p_0 (e.g. posterior median).
- The posterior distribution is function of $X_1, \dots, X_n$ (data) and of the prior P_0 (parameter of the estimator).

Posterior contraction rates

Posterior contraction rates, e.g. [Ghosal and Van der Vaart, 2017]

A sequence $(\varepsilon_n)_{n \geq 1}$ is a posterior contraction rate when

$$\mathbb{P}\left(h(P_0, p_0) \geq M_n \varepsilon_n \middle| X_1, \ldots, X_n \right) \xrightarrow[n \to \infty]{p} 0,$$

for any sequence $M_n \to \infty$.

- h is the Hellinger distance.
- The convergence in probability is w.r.t. the law of $X_1, \ldots, X_n$.

Goal : obtaining contraction rate $\epsilon_n \to 0$ as fast as possible.

- For instance matching known frequentist rates in non-parametric statistics.

Gaussian processes

Gaussian processes

A stochastic process (random field) $Z : [-1, 1]^d \to \mathbb{R}$ is a Gaussian process when for any $x_1, \ldots, x_n \in [-1, 1]^d$,

$$(Z(x_1), \ldots, Z(x_n))$$

is a Gaussian vector.

Covariance function

The function $u, v \in [-1, 1]^d \mapsto \mathrm{Cov}(Z(u), Z(v))$.

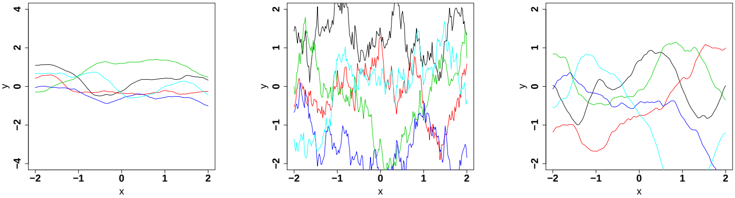


FIGURE: Various Gaussian process realizations, for various covariance functions.

RKHS and Bayesian prior

RKHS, e.g. [Berlinet and Thomas-Agnan, 2004, van der Vaart and van Zanten, 2008b]

The reproducing kernel Hilbert space (RKHS) of a Gaussian process Z is the Hilbert space $\mathbb{H}_Z$ of functions : $[-1,1]^d \to \mathbb{R}$ obtained by the "completion" of the linear combinations of the functions

$$v \in [-1,1]^d \mapsto \mathrm{Cov}(Z(u), Z(v)),$$

for $u \in [-1,1]^d$.
It provides the RKHS norm, for $f \in \mathbb{H}_Z$,

$$||f||_{\mathbb{H}_Z}.$$

Bayesian prior

For $x \in [-1,1]^d$,

$$P_0(x) = \frac{e^{Z(x)}}{\int_{[-1,1]^d} e^{Z(t)} dt},$$

where Z is a Gaussian process.

Posterior contraction rates

Concentration function, [van der Vaart and van Zanten, 2008a]

For $\varepsilon > 0$,

$$\phi_{\log p_0}(\varepsilon) = \inf_{\substack{h \in \mathbb{H}_Z \\ ||h - \log p_0||_\infty < \varepsilon}} ||h||^2_{\mathbb{H}_Z} - \log \mathbb{P}\left(||Z||_\infty < \varepsilon\right).$$

Theorem [van der Vaart and van Zanten, 2008a]

We have posterior contraction rate ϵ_n for any sequence ϵ_n such that

$$\phi_{\log p_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

Deep Gaussian processes

Deep Gaussian process, [Damianou and Lawrence, 2013]

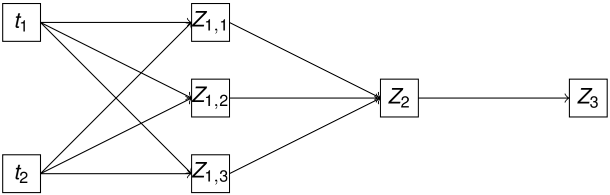
- Depth $H \in \mathbb{N}$.
- Let $d_1 = d$, d_2,...,$d_H \in \mathbb{N}$, and $d_{H+1} = 1$.
- For $h = 1, \ldots, H$, multivariate Gaussian process

$$Z_h = (Z_{h,1}, \ldots Z_{h,d_{h+1}}) : \mathbb{R}^{d_h} \to \mathbb{R}^{d_{h+1}}.$$

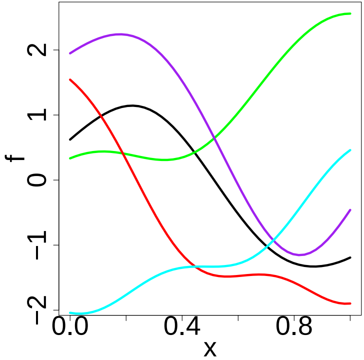
- Deep Gaussian process :

$$Z_H \circ \cdots \circ Z_1 : [-1,1]^d \to \mathbb{R}.$$

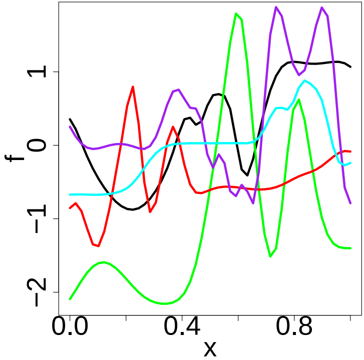
$t = (t_1, t_2)$ $\quad Z_1 = (Z_{1,1}, Z_{1,2}, Z_{1,3})$



An illustration



Gaussian process realizations.



Compositions of Gaussian process realizations.

Our work and existing work on deep Gaussian processes

- Only existing contraction rates : [Finocchio and Schmidt-Hieber, 2021].
  - Regression.
  - Sparsity and smoothness adaptation.
- Our work [Bachoc and Lagnoux, 2021].
  - Density estimation and classification.
  - No sparsity and smoothness adaptation.

Constrained prior

For the proofs, we needed to constrain the prior.

Value constraints

For $h = 1, \ldots, H-1$, $i = 1, \ldots, d_{h+1}$,

$$||Z_{h,i}||_{\infty} \leqslant 1. \tag{1}$$

Derivative constraints

For $h = 2, \ldots, H$, for $i = 1, \ldots, d_{h+1}$, and for $j = 1, \ldots, d_h$,

$$\left|\left|\frac{\partial Z_{h,i}}{\partial x_j}\right|\right|_{\infty} \leqslant K_{h,i,j}, \tag{2}$$

with constants $K_{h,i,j} > 0$.

Prior

Prior on p_0, for $x \in [-1,1]^d$,

$$P_0(x) = \frac{e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(x)}}{\int_{[-1,1]^d} e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(t)} dt},$$

where $Z_{c,1}, \ldots, Z_{c,H}$ have law of $Z_1, \ldots, Z_H$ conditioned by (1) and/or (2).

Concentration function

- Write $\log p_0$ with same composition structure as prior :

$$\log p_0 = z_{0,H} \circ \cdots \circ z_{0,1}$$

  with, for $h = 1, \ldots, H$, $z_{0,h} = (z_{0,h,1}, \ldots, z_{0,h,d_{h+1}})$.

- For $\varepsilon > 0$, let

$$\Phi_{c,z_0}(\varepsilon) = \sum_{i=1}^{d_2} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{1,i} \\ \|g - z_{0,1,i}\|_\infty < \varepsilon}} \|g\|^2_{\mathbb{H}_{1,i}} - 2 \log \mathbb{P}\left(\|Z_{1,i}\|_\infty < \varepsilon \right) \right)$$
$$+ \sum_{\substack{(h,i) \in \mathcal{I} \\ h \geqslant 2}} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{h,i} \\ \|g - z_{0,h,i}\|_\infty < \frac{\varepsilon}{2} \\ \|\partial g / \partial x_j - \partial z_{0,h,i} / \partial x_j\|_\infty < \frac{K_{\min}}{4}, \\ j = 1, \ldots, d_h}} \|g\|^2_{\mathbb{H}_{h,i}} \right.$$
$$\left. - 2 \log \mathbb{P}\left(\|Z_{h,i}\|_\infty \leqslant \frac{\varepsilon}{2} \right) - 2 \sum_{j=1}^{d_h} \log \mathbb{P}\left(\|\partial Z_{h,i} / \partial x_j\|_\infty \leqslant \frac{K_{\min}}{4} \right) \right),$$

  with

$$K_{\min} = \min_{h=2,\ldots,H} \min_{i=1,\ldots,d_{h+1}} \min_{j=1,\ldots,d_h} K_{h,i,j}.$$

Contraction rates

Theorem [Bachoc and Lagnoux, 2021]

For ε_n such that

$$\Phi_{c,z_0}(\varepsilon_n) \leq n\varepsilon_n^2,$$

we have posterior contraction rates at rate ε_n.

$\Longrightarrow$ Typically Φ_{c,z_0} (deep, constrained) has same order of magnitude as the corresponding individual concentration functions (standard Gaussian).

$\Longrightarrow$ Rate holds for all decompositions

$$\log p_0 = z_{0,H} \circ \cdots \circ z_{0,1}.$$

Example of Matérn prior

When p_0 is β-smooth (Hölder and Sobolev), and the Gaussian processes have Matérn covariance functions, if the Matérn smoothness parameters are well-chosen, we obtain the rate

$$\varepsilon_n = n^{-\beta/2\beta+d}.$$

Conclusion

Many remaining open problems :

- Adaptation to smoothness and sparsity [Finocchio and Schmidt-Hieber, 2021] in density estimation and classification.
- Theoretical analysis of computational approximations.

Thank you for your attention !

Bachoc, F. and Lagnoux, A. (2021).
Posterior contraction rates for constrained deep gaussian processes in density estimation and classication.
arXiv preprint arXiv :2112.07280.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability in Statistics.
Kluwer, Dordrecht.

Damianou, A. and Lawrence, N. (2013).
Deep Gaussian processes.
In Carvalho, C. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, AISTATS '13, pages 207–215. JMLR W&CP 31.

Finocchio, G. and Schmidt-Hieber, J. (2021).
Posterior contraction for deep Gaussian process priors.
arXiv preprint arXiv :2105.07410.

Ghosal, S. and Van der Vaart, A. (2017).
Fundamentals of nonparametric Bayesian inference, volume 44.
Cambridge University Press.

van der Vaart, A. W. and van Zanten, J. H. (2008a).
Rates of contraction of posterior distributions based on Gaussian process priors.
The Annals of Statistics, 36(3) :1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b).
Reproducing kernel Hilbert spaces of Gaussian priors.
In *Pushing the limits of contemporary statistics : contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics.