Vitesses de contraction du posterior pour les processus gaussiens profonds contraints en classification et estimation de densité

François Bachoc [1] & Agnès Lagnoux [2]

[1] *Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.* francois.bachoc@math.univ-toulouse.fr

[2] *Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.* lagnoux@univ-tlse2.fr

Résumé. Nous fournissons des vitesses de contraction du posterior pour les processus gaussiens contraints profonds en estimation non paramétrique de densité et en classification. Les contraintes sont des bornes sur les valeurs et dérivées des processus gaussiens dans les couches de la structure de composition. Les vitesses de contraction sont d'abord données dans un cadre général, sous la forme d'une nouvelle fonction de concentration que l'on introduit et qui prend les contraintes en compte. Ensuite, le cadre général est appliqué au mouvement Brownien intégré, au processus de Riemann Liouville et au processus de Matérn, avec des classes de fonctions standard. Dans chacun des exemples, on retrouve des vitesses minimax classiques.

Mots-clés. Prior gaussien, prior gaussien profond, inférence bayésienne, estimation non paramétrique de densité, classification, contraction du posterior, classes de fonctions régulières, fonctions de covariance de Matérn.

Abstract. We provide posterior contraction rates for constrained deep Gaussian processes in non-parametric density estimation and classification. The constraints are in the form of bounds on the values and on the derivatives of the Gaussian processes in the layers of the composition structure. The contraction rates are first given in a general framework, in terms of a new concentration function that we introduce and that takes the constraints into account. Then, the general framework is applied to integrated Brownian motions, Riemann-Liouville processes, and Matérn processes and to standard smoothness classes of functions. In each of these examples, we can recover known minimax rates.

Keywords. Gaussian priors, deep Gaussian priors, Bayesian inference, nonparametric density estimation, classification, posterior contraction, smoothness classes, Matérn covariance functions.

1 Introduction

1.1 Vitesse de concentration du posterior

On considère une fonction densité de probabilité fixée, continue et inconnue $p_0 : [-1,1]^d \to [0,\infty)$ et l'on observe des vecteurs aléatoires $X_1, \ldots, X_n$ i.i.d. de densité p_0. Dans cette communication, on se limite au cas de l'estimation d'une densité, mais le manuscrit long Bachoc et Lagnoux (2021), que résume cette communication, considère aussi un problème de classification. On considère un prior Bayésien sur la densité p_0 sous la forme d'une densité de probabilité aléatoire P_0. La règle de Bayes fournit alors un posterior qui prend la forme d'une distribution de probabilité sur l'espace $\mathcal{D}$ des densités de probabilité continues sur $[-1,1]^d$. On exprime ce posterior sous la forme

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

et on le voit comme une distribution dépendant des observations $X_1, \ldots, X_n$. On note que la loi de ce posterior aléatoire dépend de la densité inconnue p_0. On s'intéresse alors à montrer une vitesse de contraction du posterior aléatoire vers la vraie densité p_0. C'est-à-dire que l'on cherche une suite $(\varepsilon_n)_{n\in\mathbb{N}}$, tendant vers zéro aussi vite que possible, telle que

$$\mathbb{P}\left(h\left(P_0, p_0\right) \geq M_n \varepsilon_n \middle| X_1, \ldots, X_n\right) \xrightarrow[n\to\infty]{p} 0, \tag{1}$$

pour toute suite $(M_n)_{n\in\mathbb{N}}$ qui tend vers l'infini. Ici h est la distance de Hellinger. Dans cette convergence, la probabilité porte sur les observations $X_1, \ldots, X_n$ et dépend donc de p_0. Un référence générale sur les vitesses de contraction du posterior est Ghosal, Ghosh et van der Vaart (2000).

1.2 Processus Gaussiens

On considère ici que la densité aléatoire P_0 du prior Bayésien s'écrit, pour $x \in [-1,1]^d$,

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

où $Z : [-1,1]^d \to \mathbb{R}$ est un processus gaussien continu. Notons $w_0 = \log(p_0)$. Notons $\mathbb{H}_Z$ l'espace de Hilbert à noyau reproduisant (RKHS) de Z, avec la norme associée $||\cdot||_{\mathbb{H}_Z}$. Dans van der Vaart et van Zanten (2008), il est introduit la fonction $\phi_{w_0} : (0,\infty) \to [0,\infty)$ donnée par, pour $\varepsilon > 0$,

$$\phi_{w_0}(\varepsilon) = \inf_{\substack{h\in\mathbb{H}_Z \\ ||h-w_0||_\infty < \varepsilon}} ||h||^2_{\mathbb{H}_Z} + \mathbb{P}\left(||Z||_\infty < \varepsilon\right). \tag{2}$$

Cette fonction est interprétée comme mesurant la concentration du processus gaussien Z autour de la fonction w_0 et est donc appelée fonction de concentration. Dans van der

Vaart et van Zanten (2008), il est montré qu'il y aura contraction du posterior à vitesse ε_n, au sens de (1), for toute suite ε_n satisfaisant

$$\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

Ainsi on voit que la fonction de concentration ϕ_{w_0} mesure la vitesse de contraction : plus cette fonction tend vite vers l'infini en zéro, plus la vitesse de contraction sera lente. De plus, on voit que ϕ_{w_0} est plus petite lorsque w_0 appartient au RKHS $\mathbb{H}_Z$, ce qui montre que l'on a une vitesse plus rapide lorsque la densité inconnue est en accord avec le prior (en terme de régularité). Le second terme dans (2) est plus petit lorsque le processus Z est régulier, ce qui laisse voir que la vitesse de contraction sera plus rapide pour une fonction régulière et un prior régulier.

1.3 Processus gaussiens profonds

Les processus gaussiens profonds sont fournis par des compositions de processus gaussiens et on étés proposés initialement par Damianou et Lawrence (2013). Ici, on va considérer des processus profonds de la forme suivante. On prend $H \in \mathbb{N}^* \setminus \{1\}$ et $d_1 = d$, d_2,...,$d_H \in \mathbb{N}^*$, et $d_{H+1} = 1$. Pour $h = 1, \ldots, H$, on considère un processus gaussien centré multivarié $Z_h = (Z_{h,1}, \ldots Z_{h,d_{h+1}}) : \mathbb{R}^{d_h} \to \mathbb{R}^{d_{h+1}}$. On suppose que $Z_1, \ldots, Z_H$ sont indépendants, à composantes indépendantes et continues. On considère alors un processus gaussien profond donné par

$$Z_H \circ \cdots \circ Z_1,$$

qui fournit un prior (non Gaussien) pour les fonctions de $[-1, 1]^d$ dans $\mathbb{R}$. Les processus profonds peuvent être vus comme des réseaux de neurones profonds dans lesquels les fonctions d'activation ont un prior Bayésien donné par des processus gaussiens, voir la Figure 1. Les processus gaussiens profonds ont peu de garanties théoriques à notre connaissance, à l'exception de Bachoc et Lagnoux (2021) et de Finocchio et Schmidt-Hieber (2021).

2 Vitesses de contraction générales

2.1 Prior avec contraintes

Pour les preuves que nous fournissons dans Bachoc et Lagnoux (2021), nous devons conditionner les processus $Z_1, \ldots, Z_H$ par des contraintes sur les valeurs, pour $h = 1, \ldots, H-1$, $i = 1, \ldots, d_{h+1}$,

$$||Z_{h,i}||_\infty \leqslant 1, \tag{3}$$

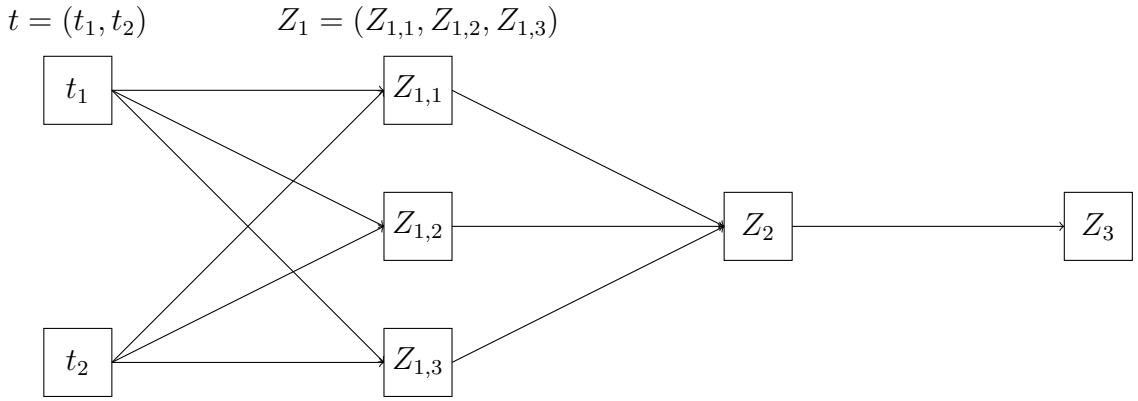

Figure 1: Exemple d'un processus gaussien profond de $[-1,1]^2$ dans $\mathbb{R}$. Ici $H = 3$, $d_1 = d = 2$, $d_2 = 3$, $d_3 = 1$, et $d_{H+1} = d_4 = 1$.

et des contraintes sur les dérivées, pour $h = 2, \ldots, H$, pour $i = 1, \ldots, d_{h+1}$, et pour $j = 1, \ldots, d_h$,

$$\left|\left|\frac{\partial Z_{h,i}}{\partial x_j}\right|\right|_\infty \leqslant K_{h,i,j}, \tag{4}$$

avec des constantes $K_{h,i,j} > 0$. Dans (3) et (4), les normes infinies sont prises sur le domaine $[-1,1]^d$. On utilisera alors le prior gaussien profond $Z_{c,H} \circ \cdots \circ Z_{c,1}$, où $Z_{c,1}, \ldots, Z_{c,H}$ sont des processus qui suivent la loi de $Z_1, \ldots, Z_H$ conditionnée par (3) et (4). Le prior utilisé sur p_0 est finalement donné par, pour $x \in [-1,1]^d$,

$$P_0(x) = \frac{e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(x)}}{\int_{[-1,1]^d} e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(t)} dt}. \tag{5}$$

On note que les contraintes (3) (resp. (4)) sont satisfaites avec probabilité non nulle pour tout les processus gaussiens à trajectoires continues (resp. continuement différentiables), ce qui autorise la plupart des fonctions de covariance continues (resp. dérivables) usuelles.

2.2 Fonction de concentration adaptée à la composition et aux contraintes

Pour $h = 1, \ldots, H$, $i = 1, \ldots, d_{h+1}$, on note $\mathbb{H}_{h,i}$ le RKHS de $Z_{h,i}$ avec la norme $||\cdot||_{\mathbb{H}_{h,i}}$. On écrit également la log densité $w_0 = \log(p_0)$ avec la même structure de composition que $\log(P_0)$: $w_0 = z_{0,H} \circ \cdots \circ z_{0,1}$ avec, pour $h = 1, \ldots, H$, $z_{0,h} = (z_{0,h,1}, \ldots, z_{0,h,d_{h+1}})$. On suppose que les fonctions $z_{0,1}, \ldots, z_{0,H}$ satisfont les mêmes contraintes que $Z_{c,1}, \ldots, Z_{c,H}$ ci-dessus et on remarque que Bachoc et Lagnoux (2021) argumentent qu'il s'agit en fait d'une hypothèse peu restrictive. Nous introduisons alors la fonction de concentration

$\Phi_{c,z_0} : (0,\infty) \to [0,\infty)$ suivante, prenant en compte la composition et les contraintes. On écrit $\mathcal{I} = \{(h,i);\, h \in \{1,\dots,H\}, i \in \{1,\dots,d_{h+1}\}\}$. Pour $\varepsilon > 0$, on définit

$$\begin{aligned}\Phi_{c,z_0}(\varepsilon) = &\sum_{i=1}^{d_2} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{1,i} \\ \|g - z_{0,1,i}\|_\infty < \varepsilon}} \|g\|^2_{\mathbb{H}_{1,i}} - 2\log \mathbb{P}\left(\|Z_{1,i}\|_\infty < \varepsilon\right) \right) \\ &+ \sum_{\substack{(h,i)\in\mathcal{I} \\ h \geqslant 2}} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{h,i} \\ \|g - z_{0,h,i}\|_\infty < \frac{\varepsilon}{2} \\ \left\|\partial g/\partial x_j - \partial z_{0,h,i}/\partial x_j\right\|_\infty < \frac{K_{\min}}{4}, \\ j=1,\dots,d_h}} \|g\|^2_{\mathbb{H}_{h,i}} \right. \\ &\left. - 2\log \mathbb{P}\Big(\|Z_{h,i}\|_\infty \leqslant \frac{\varepsilon}{2}\Big) - 2\sum_{j=1}^{d_h} \log \mathbb{P}\Big(\|\partial Z_{h,i}/\partial x_j\|_\infty \leqslant \frac{K_{\min}}{4}\Big) \right),\end{aligned} \tag{6}$$

avec

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

2.3 Vitesses de contraction

Nous montrons alors le résultat de contraction général suivant.

Théorème 1 *Considérons une suite $(\varepsilon_n)_{n\in\mathbb{N}}$ satisfaisant $\Phi_{c,z_0}(\varepsilon_n) \leq n\varepsilon_n^2$. Alors, le prior par processus Gaussien profond contraint* (5) *présente une contraction du posterior à vitesse ε_n comme dans* (1).

Un avantage de ce résultat est que les termes de la fonction de concentration (6), correspondants à chaque fonction $z_{0,h,i}$, ont généralement le même ordre de grandeur que dans (2) dans le cas non contraint. De plus, on est libre de choisir la décomposition de $\log(p_0)$ qui fournit la plus petite fonction Φ_{c,z_0}.

3 Exemple du prior de Matérn

On considère que la densité p_0 appartient à une classe de régularité $\mathcal{F}^\beta([-1,1]^d,\mathbb{R}) \cap \mathcal{H}^\beta([-1,1]^d,\mathbb{R})$, ou $\beta > 0$ est le paramètre de régularité, ce qui correspond à avoir $\lfloor\beta\rfloor$ dérivées qui sont $\beta - \lfloor\beta\rfloor$ Hölder, avec une autre condition similaire portant sur la transformée de Fourier.

On prend alors pour chaque processus gaussien $Z_{h,i}$ une fonction de covariance de Matérn de paramètre de régularité $\alpha_{h,i} \geq \beta$. On a alors la vitesse de contraction du posterior suivante.

Théorème 2 *On a une contraction du posterior comme dans* (1) *à vitesse*

$$n^{-\beta/(2\alpha_{1,\min}+d)},$$

où $\alpha_{1,\min} = \min(\alpha_{1,1}, \ldots, \alpha_{1,d_2})$.

Lorsque $\alpha_{1,\min} = \beta$ on obtient la vitesse minimax $n^{-\beta/\beta+d}$, ce qui corrrespond à un prior bien adapté.

4 Idées des preuves

L'idée de la preuve du Théorème 1 est d'utiliser les résultats de van der Vaart et van Zanten (2008) qui montrent que la vitesse contraction est régie par ϕ_{w_0} dans (2). Ces résultats peuvent en fait s'appliquer à n'importe quel processus gaussien défini sur un compact. On voit alors les composants du processus profond $(Z_{h,i}; (h,i) \in \mathcal{I})$ comme un unique processus gaussien W indexé par le compact $\mathcal{X} = [-1,1]^{d_{\max}} \times \mathcal{I}$, avec $d_{\max} = \max(d_1, \ldots, d_H)$. On construit son RKHS $\mathbb{H}$ et on introduit sa fonction de concentration ϕ_{c,w_0}, comme dans (2) autour d'une fonction $w_0 : \mathcal{X} \to \mathbb{R}$. On montre alors que pour $\varepsilon > 0$

$$\phi_{c,w_0}(\varepsilon) \leq \Phi_{c,w_0}(\varepsilon),$$

où Φ_{c,w_0} est comme dans (6). On peut alors conclure en utilisant les mêmes techniques que dans van der Vaart et van Zanten (2008).

La preuve du Théorème 2 consiste à contrôler les termes dans (6), avec des méthodes similaires à van der Vaart et van Zanten (2008).

Bibliographie

Bachoc, F. and Lagnoux, A. (2021). Posterior contraction rates for constrained deep Gaussian processes in density estimation and classication, *https://arxiv.org/abs/2112.07280*.
Damianou, A. and Lawrence, N. (2013). Deep Gaussian Processes, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, JMLR W&CP 31, pp. 207–215.
Finocchio, G. and Schmidt-Hieber, J. (2021). Posterior contraction for deep Gaussian process priors, *https://arxiv.org/abs/2105.07410*.
Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions, *Annals of Statistics*, 28, pp. 500–531.
van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors, *Annals of Statistics*, 36, pp. 1435–1463.