Gaussian Processes for nonparametrics

Ismaël Castillo & Elie Odin

Toulouse, July 2025

Special thanks to:

The ANR GAP project

[PI François Bachoc]

&

Botond Szabo, Yichen Zhu

1. Introduction
   - GPs: examples
   - Bayesian statistics
   - Nonparametric framework
   - Bayesian asymptotics
2. Statistical properties of GPs I
   - Contraction rates for GPs
   - Adaptation to smoothness
   - Variable selection
   - UQ and other topics
3. Statistical properties of GPs II
   - GPs and geometry: the intrinsic approach
   - GPs and geometry: the extrinsic approach
   - Deep GPs
4. Scalable GPs: approximations and surrogates
   - Variational Bayes
   - Vecchia GPs
   - Future directions

1. Introduction

1. Introduction
    - GPs: examples
    - Bayesian statistics
    - Nonparametric framework
    - Bayesian asymptotics
2. Statistical properties of GPs I
3. Statistical properties of GPs II
4. Scalable GPs: approximations and surrogates

Recap Define a Gaussian function random function [cf Francois's lectures!]

1) GP as Series expansion

- (φ_k) orthonormal basis
- (ζ_k) iid $\mathcal{N}(0, 1)$
- $\sigma_k > 0$

$$W(x) = \sum_{k \geq 1} \sigma_k \zeta_k \varphi_k(x)$$

Recap Define a Gaussian function random function [cf Francois's lectures!]

1) GP as Series expansion

- (φ_k) orthonormal basis
- (ζ_k) iid $\mathcal{N}(0,1)$
- $\sigma_k > 0$

$$W(x) = \sum_{k\geq 1} \sigma_k \zeta_k \varphi_k(x)$$

Example 1 $W_1(x) = \zeta_1 \cos(2\pi x)$

Example 2 For $\alpha > 0$ a 'regularity' parameter

$$W_2(x) = \sum_{k\geq 1} k^{-1/2-\alpha} \zeta_k \varphi_k(x)$$

2) GP as stochastic process process $W(\cdot)$ with
$(W(t_1), \ldots, W(t_p))$ multivariate Gaussian for all $t_1 < \ldots < t_p$

For $K(\cdot, \cdot)$ positive definite kernel, there exists a GP W with

$$EW_t = 0, \qquad EW_s W_t = K(s, t)$$

2) GP as stochastic process process $W(\cdot)$ with $(W(t_1), \ldots, W(t_p))$ multivariate Gaussian for all $t_1 < \ldots < t_p$

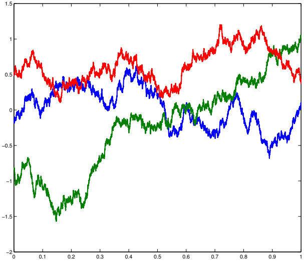
For $K(\cdot,\cdot)$ positive definite kernel, there exists a GP W with

$$EW_t = 0, \qquad EW_sW_t = K(s,t)$$

Example 1 Brownian motion (B_t)

$$K(s,t) = \min(s,t) = s \wedge t$$

Example 2 Brownian motion released at 0

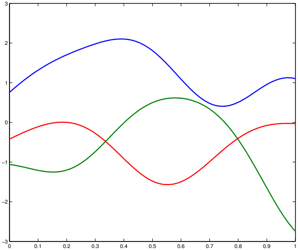


$$B_t^R = Z + B_t$$

$Z \sim \mathcal{N}(0,1)$ independent of (B_t)

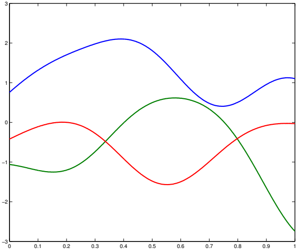
$$K(s,t) = 1 + \min(s,t)$$

Example 3 Squared-exponential (SqExp) GP



$$K(s,t) = e^{-(s-t)^2}$$

Example 3 Squared-exponential (SqExp) GP



$$K(s,t) = e^{-(s-t)^2}$$

3) GP as Gaussian random variable in Banach space $\mathbb{B}$

$$Z : \Omega \to \mathbb{B}$$

with e.g. $\mathbb{B} = (L^2[0,1], \|\cdot\|_2)$ or $\mathbb{B} = (\mathcal{C}^0[0,1], \|\cdot\|_\infty)$

Fact Under mild conditions, process def and $\mathbb{B}$-valued def are equivalent

GP $\leftrightarrow$ series GPs can be expanded into series, e.g. via Karhunen-Loève expansion

$$B(s) = \sum_{k\geq 1} \mu_k \zeta_k \varphi_k^B(s), \qquad \mu_k \sim k^{-1} = k^{-1/2-1/2}$$

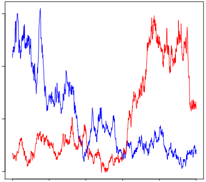
4) Transformations of GPs

- integrated Brownian motion is $\int_0^t B(s)ds$
- α–Riemann-Liouville process

$$W_t = \int_0^t (t-s)^{\alpha-1/2} dB_s$$

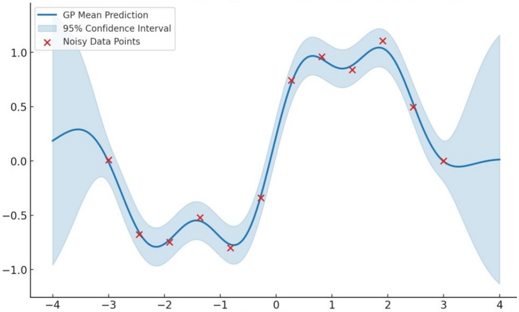
- For (Z_t) a GP, can define a random density by (!not a GP!)

$$t \to \frac{e^{Z_t}}{\int_0^1 e^{Z_s} ds}$$



Goal Use GPs for statistical inference!

- In order to: estimate unknown function f nonparametrics
- How? Being Bayesian!



[https://aicompetence.org]

Topics of this mini-course

- use of GPs to make statistical inference on functions
- are GPs statistically optimal?
- limitations of GPs and how can one overcome these?
- some works in progress and open directions

These lectures $\rightarrow$ general tools to derive properties of GPs in nonparametrics

First, let us define the setting of BNP = Bayesian nonparametrics

Statistics: standard frequentist framework

Statistical experiment

- X random object = data
- $\mathcal{P}$ model

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}.$$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Statistics: standard frequentist framework

Statistical experiment

- X random object = data
- $\mathcal{P}$ model

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}.$$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Estimator a measurable function $\hat{\theta}(X) \in \Theta$

$\hat{\theta}(X)$ is a random point in Θ

one studies $\hat{\theta}(X)$ under $X \sim P_{\theta_0}$

example $\hat{\theta}^{MLE}(X) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p_\theta(X)$

Statistics: Bayesian framework

Statistical experiment

- X random object = data
- $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ model

Bayesian setting [Do not know θ? View it as random!]

a) $\theta \sim \Pi$ prior distribution

b) $X|\,\theta \sim P_\theta$

$\Rightarrow$ joint distribution of (θ, X) is specified

c) law of $\theta|\,X$ is posterior distribution denoted $\Pi[\cdot|\,X]$

Bayesian estimator $\Pi(\cdot|\,X) \in \mathcal{M}_1(\Theta)$

$\Pi(\cdot|\,X)$ is a data–dependent measure on Θ

E0 – Example 0

$$X = (X_1, \ldots, X_n)$$

$$\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$$

Frequentist estimator

$$\hat{\theta}^{MLE}(X) = \overline{X}_n$$

Bayesian setting

a) $\theta \sim \mathcal{N}(0,1) = \Pi$ prior (say)

b) $X|\,\theta \sim \mathcal{N}(\theta, 1)^{\otimes n}$

c) $\theta|\,X \sim \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right) = \Pi[\cdot|\,X]$ posterior

$$\Pi[\cdot|\,X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Bayesian framework

Bayesian setting

a) $\theta \sim \Pi$ prior

b) $X|\,\theta \sim P_\theta$

c) $\theta|\,X \sim: \Pi[\cdot|\,X]$ posterior

All this produces a data–dependent measure $\Pi[\cdot|\,X]$

Bayesian framework

Bayesian setting

a) $\theta \sim \Pi$ prior

b) $X|\,\theta \sim P_\theta$

c) $\theta|\,X \sim: \Pi[\cdot\,|\,X]$ posterior

All this produces a data–dependent measure $\Pi[\cdot\,|\,X]$

And what if ...

... one would forget a)+b)+c) ...

... and study $\Pi[\cdot\,|\,X]$ as a 'standard' estimator??

Frequentist analysis of Bayesian procedures

Posterior distribution $\Pi[\cdot \,|\, X]$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Frequentist analysis of Bayesian procedures

Posterior distribution $\Pi[\cdot \mid X]$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

E0 - Example 0

$$\begin{aligned} X \mid \theta &\sim \mathcal{N}(\theta, 1)^{\otimes n} \\ \theta &\sim \mathcal{N}(0, 1) = \Pi \end{aligned}$$

$$\Pi[\cdot \mid X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right) \sim \overline{\theta}(X) + \frac{1}{\sqrt{n+1}}\mathcal{N}(0,1)$$

- centered close to $\hat{\theta}^{MLE}(X) = \overline{X}_n$
- *converges* at rate $1/\sqrt{n}$ towards θ_0 [see below]

Nonparametric models

Consider the problem of estimation of a function f

Nonparametric models

Consider the problem of estimation of a function f

Gaussian white noise

$$dX(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t), \qquad t \in [0,1]$$

Gaussian regression one observes design points X_i and values Y_i

$$Y_i = f(X_i) + \epsilon_i, \qquad 1 \leq i \leq n$$

Inverse problems $Y_i = \mathcal{G}(f)(X_i) + \epsilon_i, \qquad 1 \leq i \leq n$

Nonparametric models

Consider the problem of estimation of a function f

Gaussian white noise

$$dX(t) = f(t)dt + \frac{1}{\sqrt{n}} dW(t), \qquad t \in [0,1]$$

Gaussian regression   one observes design points X_i and values Y_i

$$Y_i = f(X_i) + \epsilon_i, \qquad 1 \le i \le n$$

Inverse problems $Y_i = \mathcal{G}(f)(X_i) + \epsilon_i, \qquad 1 \le i \le n$

many other settings   density estimation, classification ...

Estimating unknown f from data X is a nonparametric problem

Typical optimal minimax estimation rate, for β–smooth f in $\mathbb{R}^d$ in $\|\cdot\|_2$–loss

$$n^{-\frac{\beta}{2\beta+d}}$$

Bayesian asymptotics

Bayesian setting X data

a) $f \sim \Pi$ prior

b) $X|\, f \sim P_f = P_f^{(n)}$ model

c) $f|\, X \sim: \Pi[\cdot|\, X]$ posterior

Frequentist analysis of Bayesian procedures

- Assume there exists f_0 such that $X \sim P_{f_0}$
- study the behaviour of $\Pi[\cdot|\, X]$ under P_{f_0}:
  - convergence to f_0
  - goal: find $\varepsilon_n \to 0$ as fast as possible with, as $n \to \infty$,

$$\Pi[\{f : \|f - f_0\|_2 \leq \varepsilon_n\}|\, X] \to^{P_{f_0}} 1$$

Bayesian dominated framework

Experiment. $X = X^{(n)}$, $\mathcal{P} = \{P_f^{(n)},\ f \in \mathcal{F}\}$, $(\mathcal{F}, \mathbb{F})$ measure space

Dominated framework. Suppose there exists a dominating measure $\mu^{(n)}$

$$dP_f^{(n)} = p_f^{(n)}(\cdot) d\mu^{(n)}$$

Bayesian setting.

a) $f \sim \Pi$ prior distribution

b) $X \,|\, f \sim P_f^{(n)}$

c) $f \,|\, X \sim: \Pi[\cdot \,|\, X]$ posterior

Bayes formula. For any measurable set $B \in \mathbb{F}$,

$$\Pi(B \,|\, X^{(n)}) = \frac{\int_B p_f^{(n)}(X^{(n)}) d\Pi(f)}{\int p_f^{(n)}(X^{(n)}) d\Pi(f)}.$$

Remark. $\Pi[B] = 0 \Rightarrow \Pi[B \,|\, X] = 0$

Special case: Gaussian regression

Observe $(X, Y) = (X_i, Y_i)_{1 \le i \le n}$, with $X_i \overset{iid}{\sim} P_X$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ and 'true' f_0

$$Y_i = f_0(X_i) + \sigma\epsilon_i$$

Prior distribution Π. Let $f \sim \Pi = GP(0, k)$

Special case: Gaussian regression

Observe $(X, Y) = (X_i, Y_i)_{1 \le i \le n}$, with $X_i \overset{iid}{\sim} P_X$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and 'true' f_0

$$Y_i = f_0(X_i) + \sigma\epsilon_i$$

Prior distribution Π. Let $f \sim \Pi = GP(0, k)$

Posterior $\Pi[\cdot \,|\, X, Y]$ is a GP

$$x \mapsto K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}\boldsymbol{y}, \qquad \textit{mean}$$

$$(x,z) \mapsto k(x,z) - K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}K_{\boldsymbol{f}z}, \qquad \textit{covariance}$$

Here we denote $\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f} = (f(X_1), \ldots, f(X_n))^T$,
$K_{x\boldsymbol{f}} = \mathrm{cov}_\Pi(f(x), \boldsymbol{f}) = (k(x, X_1), \ldots, k(x, X_n))$, $K_{\boldsymbol{ff}} = \mathrm{cov}_\Pi(\boldsymbol{f}, \boldsymbol{f}) = [k(X_i, X_j)]_{1 \le i,j \le n}$.

Special case: Gaussian regression

Observe $(X, Y) = (X_i, Y_i)_{1\le i\le n}$, with $X_i \overset{iid}{\sim} P_X$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and 'true' f_0

$$Y_i = f_0(X_i) + \sigma\epsilon_i$$

Prior distribution Π. Let $f \sim \Pi = GP(0, k)$

Posterior $\Pi[\cdot\,|\,X, Y]$ is a GP

$$x \mapsto K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}\boldsymbol{y}, \qquad mean$$

$$(x, z) \mapsto k(x, z) - K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}K_{\boldsymbol{f}z}, \qquad covariance$$

Here we denote $\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f} = (f(X_1), \ldots, f(X_n))^T$,
$K_{x\boldsymbol{f}} = \mathrm{cov}_\Pi(f(x), \boldsymbol{f}) = (k(x, X_1), \ldots, k(x, X_n))$, $K_{\boldsymbol{ff}} = \mathrm{cov}_\Pi(\boldsymbol{f}, \boldsymbol{f}) = [k(X_i, X_j)]_{1\le i,j\le n}$.

- Analytic expression, allows direct 'computations'
- Inconvenient: works only for Gaussian regression

Convergence rate

Convergence rate. The posterior converges at rate ε_n for distance d at f_0 if

$$E_{\theta_0} \Pi(\theta : \ d(\theta, \theta_0) \leq \varepsilon_n | X) \longrightarrow 1 \qquad (n \to \infty)$$

It is an upper bound: we look for the smallest possible ε_n.

What happens in a nonparametric framework ?

- [Ghosal, Ghosh, van der Vaart 00], [Ghosal, van der Vaart 07]

First examples

Fixed design regression [van der Vaart, van Zanten 08, 09]

$$Y_i = f(t_i) + \epsilon_i, \qquad 1 \le i \le n, \qquad \epsilon_i \sim \mathcal{N}(0,1) \quad \text{iid}$$

True function. Let $f_0 \in \mathcal{C}^\beta[0,1]$

Loss function. $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g(t_i)^2$

Prior. *Brownian motion + Gaussian*

$W_t = B_t + Z_0$, with $Z_0 \sim \mathcal{N}(0,1)$

Then as $n \to \infty$,

$$E_{f_0} \Pi \left[f : \|f - f_0\|_n \le \varepsilon_n \,|\, X \right] \to 1,$$

$$\varepsilon_n \sim n^{-\frac{1}{4} \wedge \frac{\beta}{2}} = \begin{cases} n^{-1/4} & \text{if } \beta \ge 1/2 \\ n^{-\beta/2} & \text{if } \beta \le 1/2 \end{cases}$$

First examples

Fixed design regression (followed)

Prior. *Riemann-Liouville process with parameter* $\alpha > 0$

$W_t = \int_0^t (t-s)^{\alpha-1/2} dB_s + \sum_{k=0}^{\lceil \alpha \rceil} Z_k t^k$, with $Z_k \sim \mathcal{N}(0,1)$ iid

Then

$$E_{f_0} \Pi \left[f : \ \|f - f_0\|_n \leq \varepsilon_n | \, X \right] \to 1,$$

where

$$\varepsilon_n \approx n^{-\frac{\alpha \wedge \beta}{2\alpha+1}} = \begin{cases} n^{-\frac{\alpha}{2\alpha+1}} & \text{if } \beta \geq \alpha \\ n^{-\frac{\beta}{2\alpha+1}} & \text{if } \beta \leq \alpha \end{cases}$$

First examples

Density estimation [van der Vaart, van Zanten 08, 09]

$$X_1, \ldots, X_n \qquad \sim \ f \qquad \text{iid}$$

True density. Let $f_0 \in \mathcal{C}^\beta[0,1]$ with $f_0 > 0$.

Loss function. Hellinger distance $h(f,g)^2 = \int(\sqrt{f} - \sqrt{g})^2$

Prior. Consider the distribution on continuous functions induced by

$$t \to \frac{e^{W_t}}{\int_0^1 e^{W_u} du}$$

with W_t either *Brownian motion* or *Riemann-Liouville process with parameter* α

Then, for ε_n as before,

$$E_{f_0} \Pi\left[h(f, f_0) \leq \varepsilon_n | X\right] \to 1.$$

Theory: Bayesian nonparametrics

Setting $X = X^{(n)}$, $\mathcal{P} = \{P_f^{(n)},\ f \in \mathcal{F}\}$ [not necessarily iid]

Π prior distribution on $\mathcal{F}$

Goal For some distance d and rate ε_n [with $n\varepsilon_n^2 \to \infty$]

$$E_{f_0}\Pi\left[f :\ d(f, f_0) > M\varepsilon_n | X\right] \to 0$$

Theory: Bayesian nonparametrics

Setting $\quad X = X^{(n)}$, $\mathcal{P} = \{P_f^{(n)},\ f \in \mathcal{F}\}$ [not necessarily iid]

Π prior distribution on $\mathcal{F}$

Goal For some distance d and rate ε_n [with $n\varepsilon_n^2 \to \infty$]

$$E_{f_0}\Pi\left[f\ :\ d(f, f_0) > M\varepsilon_n | X\right] \to 0$$

Key condition $\quad$ The prior puts enough mass on neighborhoods of f_0

$$\Pi(B_{KL}(f_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}$$

Theory: Bayesian nonparametrics

Setting $X = X^{(n)}$, $\mathcal{P} = \{P_f^{(n)},\ f \in \mathcal{F}\}$ [not necessarily iid]

Π prior distribution on $\mathcal{F}$

Goal For some distance d and rate ε_n [with $n\varepsilon_n^2 \to \infty$]

$$E_{f_0}\Pi\left[f :\ d(f, f_0) > M\varepsilon_n | X\right] \to 0$$

Key condition The prior puts enough mass on neighborhoods of f_0

$$\Pi(B_{KL}(f_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}$$

$$B_{KL}(f_0, \varepsilon_n) = \left\{ K_n(f_0, f) \leq n\varepsilon_n^2,\ V_n(f_0, f) \leq n\varepsilon_n^2 \right\}$$

$$K_n(f_0, f) := \int p_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} d\mu^{(n)}, \quad V_n(f_0, f) := \int p_{f_0}^{(n)} \log^2 \frac{p_{f_0}^{(n)}}{p_f^{(n)}} d\mu^{(n)}$$

Theory: Bayesian nonparametrics

Theorem, generic [Ghosal Ghosh van der Vaart 00]

If $\mathcal{F}_n \subset \mathcal{F}$ and $c > 0$ such that, for d such that **(T0)** is verified,

$$\log N(\varepsilon_n, \mathcal{F}_n, d_n) \leq dn\varepsilon_n^2 \qquad \textit{entropy}$$

$$\Pi(\,\mathcal{F}_n^c\,) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{remaining mass}$$

$$\Pi(\,B_{KL}(f_0, \varepsilon_n)\,) \geq e^{-cn\varepsilon_n^2} \qquad \textit{prior mass}$$

Then for $M > 0$ large enough,

$$E_{f_0}\Pi[f :\; d(f, f_0) \leq M\varepsilon_n | X] \to 1$$

Fractional posteriors

$$\Pi_\rho[B|\,X] = \frac{\int_B \left(p_f^{(n)}(X)\right)^\rho d\Pi(f)}{\int \left(p_f^{(n)}(X)\right)^\rho d\Pi(f)}$$

Theorem, fractional post. Suppose, for $\varepsilon_n > 0$, $\rho \in (0,1)$ and $n\rho\varepsilon_n^2 \to \infty$,

$$\Pi(B_{KL}(f_0, \varepsilon_n)) \geq e^{-n\rho\varepsilon_n^2}.$$

Then there exists $C > 0$ such that as $n \to \infty$,

$$\Pi_\rho\left(f : \frac{1}{n}D_\rho(p_f^{(n)}, p_{f_0}^{(n)}) \geq \frac{C\rho}{1-\rho}\varepsilon_n^2 \,|\, X\right) = o_P(1).$$

$$D_\rho(p,q) = D_\alpha(f,g) = -\frac{1}{1-\alpha}\log\left(\int p^\alpha q^{1-\alpha} d\mu\right) \qquad \rho - \text{Rényi divergence}$$

[T. Zhang 06], [Bhattacharya et al. 19] [L'Huillier, Travis, C. and Ray 23]

1 Introduction

2 Statistical properties of GPs I
- Contraction rates for GPs
- Adaptation to smoothness
- Variable selection
- UQ and other topics

3 Statistical properties of GPs II

4 Scalable GPs: approximations and surrogates

2. Statistical properties of GPs I

Verifying the prior mass conditions

Key condition The prior puts enough mass on neighborhoods of f_0

$$\Pi(B_{KL}(f_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}$$

In many models, one can show

$$B_{\|\cdot\|_2}(f_0, \varepsilon_n) := \{f : \ \|f - f_0\|_2 \leq \varepsilon_n\} \subset B_{KL}(f_0, \varepsilon_n)$$

or

$$B_{\|\cdot\|_\infty}(f_0, \varepsilon_n) := \{f : \ \|f - f_0\|_\infty \leq \varepsilon_n\} \subset B_{KL}(f_0, \varepsilon_n)$$

Example Gaussian white noise model $dX(t) = f(t)dt + dW(t)/\sqrt{n}$

$$B_{KL}(f_0, \varepsilon_n) = B_{\|\cdot\|_2}(f_0, \varepsilon_n)$$

A first 'hands-on' example

Model Gaussian white noise $dX(t) = f(t)dt + dW(t)/\sqrt{n}$

Regularity of f_0 Suppose f_0 is β–smooth: for any $k \geq 1$

$$|f_{0,k}| \leq Lk^{-1/2-\beta}$$

Series GP prior Π on f For (φ_k) ONB of $L^2[0,1]$ and $\alpha > 0$,

$$W(t) = \sum_{k=1}^{\infty} k^{-1/2-\alpha} \zeta_k \varphi_k(t)$$

Theorem For any $\rho \in (0,1)$,

$$\Pi_\rho \left[\{ f : \ \|f - f_0\|_2 \leq M\varepsilon_n \} \,|\, X \right] \to 0$$

where the ρ–posterior convergence rate is

$$\varepsilon_n = n^{-\frac{\alpha \wedge \beta}{2\alpha+1}}$$

This result can be extended

- to the standard posterior $\rho = 1$
- to other models (regression, density,...)
- Sobolev regularity for f_0 ...

By generic result on ρ–posteriors, it suffices to check

$$\Pi(\|f - f_0\|_2 \leq \varepsilon_n) \geq e^{-cn\varepsilon_n^2}$$

$\Rightarrow$ ρ–posterior converges at rate ε_n

[$D_\rho(\mathcal{N}(a,1), \mathcal{N}(b,1)) = \rho \cdot (a-b)^2/2$]

Proof. Let $(\delta_k)_{k\geq 1}$ verify $\sum_{k=1}^{\infty} \delta_k^2 \leq (D\varepsilon_n)^2$.

Case $\alpha > \beta$. By independence,

$$\Pi(\|f - f_0\|_2 \leq D\varepsilon_n) = \Pi\left[\sum_{k\geq 1}(f_k - f_{0,k})^2 \leq (D\varepsilon_n)^2\right]$$
$$\geq \Pi\left[\forall\, k \geq 1,\ \ |f_k - f_{0,k}| \leq \delta_k\right] \geq \prod_{k\geq 1} \Pi\left[|f_k - f_{0,k}| \leq \delta_k\right]$$

For any $k \geq 1$,

$$\Pi\left[|f_k - f_{0,k}| \leq \delta_k\right] = P\left[|\sigma_k \zeta_k - f_{0,k}| \leq \delta_k\right]$$
$$\geq \int_{(f_{0,k}-\delta_k)/\sigma_k}^{(f_{0,k}+\delta_k)/\sigma_k} e^{-x^2/2} dx/\sqrt{2\pi}$$

By symmetry, without loss of generality assume $f_{0,k} \geq 0$ in the sequel

Let $N_\alpha := \lfloor n^{\frac{1}{1+2\alpha}} \rfloor$ and

$$\delta_k = \begin{cases} 1/\sqrt{n}, & 1 \le k \le N_\alpha, \\ 2Lk^{-1/2-\beta}, & k > N_\alpha \end{cases}$$

Case $\alpha > \beta$, $k \le N_\alpha$

$$\begin{aligned} \Pi\left[|f_k - f_{0,k}| \le \delta_k\right] &\gtrsim \int_{(f_{0,k}-\delta_k)/\sigma_k}^{(f_{0,k}+\delta_k)/\sigma_k} e^{-x^2/2} dx \\ &\gtrsim \frac{\delta_k}{\sigma_k} \exp\left\{-\frac{1}{2\sigma_k^2}(f_{0,k}+\delta_k)^2\right\} \\ &\gtrsim \frac{\delta_k}{\sigma_k} \exp\left\{-\frac{1}{\sigma_k^2}(f_{0,k}^2 + n^{-1})\right\} \\ &\gtrsim \frac{\delta_k}{\sigma_k} \exp\left\{-\frac{C}{\sigma_k^2}(Lk^{-1/2-\beta})^2\right\} \\ &\gtrsim \frac{1}{\sqrt{n}} \exp\left\{-C(k^{\alpha-\beta})^2\right\} \end{aligned}$$

Case $\alpha > \beta$, $k \leq N_\alpha$ (followed)

Using $\sum_{k=1}^{N} k^q \lesssim N^{q+1}$ for any $q > 0$ and integer N,

$$\prod_{k=1}^{N_\alpha} \Pi\left[|f_k - f_{0,k}| \leq \delta_k\right] \geq \exp\left\{-N_\alpha \log(\sqrt{n}/C_0) - C_2 N_\alpha^{2(\alpha-\beta)+1}\right\}$$

$$\geq \exp\left\{-C_3 N_\alpha^{2(\alpha-\beta)+1}\right\} \geq \exp\left\{-C_3 n\varepsilon_n^2\right\}$$

$$\geq \exp\left\{-C_3 n(D\varepsilon_n)^2\right\},$$

noticing that $N_\alpha^{2(\alpha-\beta)+1} \leq n\varepsilon_n^2$

$$n\varepsilon_n^2 = nn^{-\frac{2\beta}{2\alpha+1}} = n^{\frac{2(\alpha-\beta)+1}{2\alpha+1}}$$

Case $\alpha > \beta$, $k > N_\alpha$

$$[f_{0,k} - \delta_k, f_{0,k} + \delta_k] \supset \left[-Lk^{-1/2-\beta}, Lk^{-1/2-\beta}\right] \qquad \text{choice } \delta_k$$

$$\Pi\left[|f_k - f_{0,k}| \le \delta_k\right] \ge \Pi\left[|f_k| \le Lk^{-1/2-\beta}\right] \ge \Pi\left[|\zeta_k| \le Lk^{\alpha-\beta}\right].$$

$$\begin{aligned}
\prod_{k>N_\alpha} \Pi\left[|f_k - f_{0,k}| \le \delta_k\right] &\ge \prod_{k>N_\alpha} \left(1 - 2\overline{\Phi}(Lk^{\alpha-\beta})\right) \\
&\ge \exp\left\{\sum_{k>N_\alpha} \log\left(1 - 2e^{-(Lk^{\alpha-\beta})^2/2}\right)\right\} \\
&\ge \exp\left\{-2\sum_{k>N_\alpha} e^{-(Lk^{\alpha-\beta})^2/2}\right\} = 1 + o(1)
\end{aligned}$$

Case $\alpha > \beta$ (conclusion)

$$\Pi(\|f - f_0\|_2 \le D\varepsilon_n) \ge \exp(-Cn(D\varepsilon_n)^2)(1 + o(1)) \ge \exp(-C'n(D\varepsilon_n)^2)$$

Case $\alpha \leq \beta$ Since $N_\alpha/n \leq \varepsilon_n^2 = n^{2\alpha/(2\alpha+1)}$

$$\bigcap_{k=1}^{N_\alpha} \left\{(f_k - f_{0,k})^2 \leq D^2/(2n)\right\} \ \cap \ \left\{f : \sum_{k>N_\alpha} (f_k - f_{0,k})^2 \leq (D\varepsilon_n)^2/2\right\}$$
$$\subset \ \left\{f : \ \|f - f_0\|_2^2 \leq (D\varepsilon_n)^2\right\}$$

Case $\alpha \leq \beta$, $1 \leq k \leq N_\alpha$ With $\delta_k = D/\sqrt{2n}$,

$$\begin{aligned}
\Pi\left[|f_k - f_{0,k}| \leq \delta_k\right] &\gtrsim \ \frac{\delta_k}{\sigma_k} \exp\left\{-\frac{C}{\sigma_k^p}(f_{0,k}^2 + n^{-1})\right\} \\
&\gtrsim \ \frac{\delta_k}{\sigma_k} \exp\left\{-C(L^2 + (\sigma_k^{-1}/\sqrt{n})^2\right\} \\
&\gtrsim D\frac{k^{1/2+\alpha}}{\sqrt{n}} \exp\left\{-C_3\right\}.
\end{aligned}$$

with $|f_{0,k}| \leq Lk^{-1/2-\beta} \leq Lk^{-1/2-\alpha} = L\sigma_k$ for $\alpha \leq \beta$; $\sigma_k^{-1} \leq \sqrt{n}$ for $k \leq N_\alpha$,

Lemma As soon as $N_\alpha = \lfloor n^{\frac{1}{1+2\alpha}} \rfloor \geq 2$, it holds

$$\prod_{k=1}^{N_\alpha} \frac{k^{1/2+\alpha}}{\sqrt{n}} \geq e^{-(1/2+\alpha)N_\alpha}$$

$$\begin{aligned}\sum_{k=1}^{N_\alpha} \log k &\geq \int_1^{N_\alpha} \log(x)dx \\ &\geq N_\alpha \log N_\alpha - (N_\alpha - 1)\end{aligned}$$

Using the Lemma and $N_\alpha \lesssim n\varepsilon_n^2$

$$\prod_{k=1}^{N_\alpha} \Pi\left[|f_k - f_{0,k}| \le \delta_k\right] \ge \exp\left\{-C_5 D^2 N_\alpha\right\} \ge \exp\left\{-C_6 n(D\varepsilon_n)^2\right\}.$$

Case $\alpha \le \beta$, $k > N_\alpha$

Since $\sum_{k>N_\alpha} f_{0,k}^2 \le L^2 N_\alpha^{-2\alpha} \lesssim \varepsilon_n^2$ and $\sum_{k>N_\alpha} \sigma_k^2 \lesssim \varepsilon_n^2$,

$$\begin{aligned}
&\Pi\left[\sum_{k>N_\alpha} (f_k - f_{0,k})^2 \le (D\varepsilon_n)^2/2\right] \\
&\ge \Pi\left[\sum_{k>N_\alpha} f_k^2 \le (D\varepsilon_n)^2/4\right] \\
&\ge \Pi\left[\sum_{k>N_\alpha} \left(f_k^2 - \sigma_k^2 E[\zeta_k^2]\right) \le (D\varepsilon_n)^2/8\right],
\end{aligned}$$

By Markov's inequality,

$$
\begin{aligned}
\Pi & \left[\sum_{k > N_\alpha} \left(f_k^2 - \sigma_k^2 E[\zeta_k^2] \right) > (D\varepsilon_n)^2/8 \right] \\
& = P \left[\sum_{k > N_\alpha} \sigma_k^2 \left(\zeta_k^2 - E[\zeta_k^2] \right) > (D\varepsilon_n)^2/8 \right] \\
& \leq \frac{64}{(D\varepsilon_n)^4} \operatorname{Var} \left[\sum_{k > N_\alpha} \sigma_k^2 \zeta_k^2 \right] \\
& \leq \frac{64}{(D\varepsilon_n)^4} \operatorname{Var} \left[\zeta_1^2 \right] \sum_{k > N_\alpha} \sigma_k^4 \\
& \leq \frac{C_7}{(D\varepsilon_n)^4} N_\alpha^{-1-4\alpha}.
\end{aligned}
$$

Since $N_\alpha^{-1-4\alpha} \lesssim N_\alpha^{-1} \varepsilon_n^4$, this is a $o(1)$

Putting together the above bounds in both regimes of k's

$$\begin{aligned}
\Pi\left[\|f-f_0\|_2^2 \le (D\varepsilon_n)^2\right] \\
&\ge (1-o(1)) \cdot \exp\left\{-C_6 n (D\varepsilon_n)^2\right\} \\
&\ge \exp\left\{-C_7 n (D\varepsilon_n)^2\right\}
\end{aligned}$$

This concludes the proof of the Theorem!

'Direct prior mass' approach

- This is a typical 'qualitative' proof by prior mass
- It works in some generality under series GPs

For more general GPs and more general models, in general necessary to control

$$\Pi[\|f - f_0\|_\infty \leq \varepsilon_n] \geq \exp(-Cn\varepsilon_n^2)$$

For this we will use tailored tools for GPs $\rightarrow$ The 'RKHS approach'

GPs: RKHS

$W = (W_t : t \in T)$ centered GP

Covariance kernel $K(s,t) = E(W_s W_t)$

GPs: RKHS

$W = (W_t : t \in T)$ centered GP

Covariance kernel $K(s,t) = E(W_s W_t)$

Reproducing Kernel Hilbert Space $\mathbb{H}$ (RKHS) associated to W.

Define a norm $\|\cdot\|_{\mathbb{H}}$ via

$$\langle \sum_{i=1}^{p} a_i K(s_i,\cdot)\,,\, \sum_{j=1}^{q} b_j K(t_j,\cdot)\,\rangle_{\mathbb{H}} = \sum_{i,j} a_i b_j K(s_i, t_j)$$

Then one sets

$$\mathbb{H} = \overline{\mathrm{Vect}\{K(s,\cdot), \quad s \in T\}}^{\mathbb{H}}$$

GPs: RKHS $\mathbb{H}$, examples

Brownian motion (B_t) $\qquad \mathbb{H} = \{ \int_0^{\cdot} f(u)du, \quad f \in L^2[0,1] \}$

with inner product $\langle f, g \rangle_{\mathbb{H}} = \int_0^1 f' g'$

Sketch of proof $\qquad s \wedge \cdot = \int_0^{\cdot} 1\!\!1_{[0,s]}(u)du$

Series prior $\sum_{k \geq 1} \sigma_k \zeta_k \varphi_k$ $\qquad \mathbb{H} = \{h = (h_k) \in \ell^2, \quad \sum_{k \geq 1} \sigma_k^{-2} h_k^2 < +\infty\}$

with inner product $\langle f, g \rangle_{\mathbb{H}} = \sum_{k \geq 1} \sigma_k^{-2} f_k g_k$

GPs, the RKHS approach

Key Fact For g in the support of W in $\mathbb{B}$, and all $\varepsilon > 0$,

$$e^{-\varphi_g(\varepsilon/2)} \leq P(\|W - g\|_{\mathbb{B}} < \varepsilon) \leq e^{-\varphi_g(\varepsilon)}$$

where φ_g is the concentration function of W at g

Concentration function. Let g be in the support of W in $\mathbb{B}$. For $\varepsilon > 0$, set

$$\varphi_g(\varepsilon) \quad = \underbrace{\inf_{h \in \mathbb{H}:\ \|h-g\|_{\mathbb{B}} < \varepsilon} \frac{\|h\|_{\mathbb{H}}^2}{2}}_{\text{Approximation}} \quad \underbrace{- \log P(\|W\|_{\mathbb{B}} < \varepsilon)}_{\text{Small ball probability}}$$

Idea of proof: “Girsanov-Cameron-Martin" change of variable formula if $g \in \mathbb{H}$

GPs, the RKHS approach: the two terms

Lower-bound in Key fact shows that to prove prior mass condition, it is enough to

- either know an equivalent or get an upper-bound of

$$\varphi_0(\varepsilon) = -\log P[\|W\|_{\mathbb{B}} < \varepsilon] \qquad \text{small ball probability}$$

  - can borrow existing results from probability literature!
  - [Li, Linde 90'] small ball $\varphi_0(\varepsilon)$ is tightly connected to entropy of $\mathbb{H}_1$

- bound from above the term

$$\inf_{h\in\mathbb{H}:\ \|h-g\|_{\mathbb{B}}<\varepsilon} \|h\|_{\mathbb{H}}^2 \qquad \text{approximation term}$$

  - if $g \in \mathbb{H}$, this term is constant [take $h = g$ (!)]
  - if $g \notin \mathbb{H}$, one approximates it by h's in g

Example [small ball probability in $\mathbb{B} = L^2[0,1]$] Brownian motion (B_t)

$$-\log \mathbb{P}(\|B\|_2 < \varepsilon) \asymp \varepsilon^{-2} \quad (\varepsilon \to 0)$$

- using K-L expansion, BM is GP series prior with $\sigma_k \asymp k^{-1} = k^{-1/2-1/2}$
- we already proved $-\log \mathbb{P}(\|B\|_2 < \varepsilon_n) \leq n\varepsilon_n^2 \asymp \varepsilon_n^{-2}$ ...
- ... for $\varepsilon_n = n^{-1/4}$ [enough fo our needs!]

Example [small ball probability in $\mathbb{B} = \mathcal{C}^0[0,1]$] Brownian motion (B_t)

$$-\log \mathbb{P}(\|B\|_\infty < \varepsilon) \asymp \varepsilon^{-2} \quad (\varepsilon \to 0)$$

- can be proved directly,
- or by using link with entropy of $\mathbb{H}_1$

GPs, the RKHS approach [van der Vaart, van Zanten 08]

Consider a nonparametric problem with unknown function $f_0 \in \mathbb{B}$

Prior Π = law of a Gaussian process W on $\mathbb{B}$, with RKHS $\mathbb{H}$

Suppose

- f_0 is in the support in $\mathbb{B}$ of the prior
- the norm $\|\cdot\|$ on $\mathbb{B}$ *combines correctly* with the testing distance d

Let ε_n be a solution of the equation

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$$

Then the posterior contracts at rate ε_n: for large enough M,

$$E_{f_0}\Pi(d(f, f_0) > M\varepsilon_n | X) \to 0$$

GPs, theory via RKHS

Ingredients of proof [checking the '3 conditions' in Generic Theorem]

- prior mass
  the [Fact] links $P(\|W - w\|_{\mathbb{B}} < \varepsilon)$ and concentration function

- sieves
  [Borell 75]'s inequality
  Let $\mathbb{B}_1$ and $\mathbb{H}_1$ unit balls $\mathbb{B}$ and $\mathbb{H}$ associated to W

  $$P(W \notin M\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M)$$

  Suggests to set $\Theta_n = \sqrt{n}\varepsilon_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$

- entropy
  can link entropy of $\mathbb{H}_1$ and small ball probability

Using $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$: example of BM released at 0

Prior on f consider Brownian motion released at 0

$$W_t = B_t + Z$$

with $Z \sim \mathcal{N}(0,1)$ independent of (B_t)

View it as Gaussian random variable in $\mathbb{B} = (\mathcal{C}^0[0,1], \|\cdot\|_\infty)$

RKHS $\mathbb{H} = \{ \ c + \int_0^{\cdot} f(u)du, \quad c \in \mathbb{R}\,,\ f \in L^2[0,1] \ \}, \qquad \langle f, g \rangle_{\mathbb{H}} = \int_0^1 f'g'$

- The small ball term: as before $\varphi_0(\varepsilon) \asymp \varepsilon^2$
- Approximation term: need to find

$$\inf_{h \in \mathbb{H}:\ \|h - f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2$$

Using $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$: example of BM released at 0

Let $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ be the RKHS of Brownian motion released at 0
Suppose $f_0 \in \mathcal{C}^\beta[0,1]$, for some $\beta > 0$

$$\inf_{h \in \mathbb{H}:\, \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{\frac{2\beta-2}{\beta} \wedge 0}.$$

- If $\beta \geq 1$ then $f_0 \in \mathbb{H}$ (!) [the 'inf' is a constant in this case]
- If $\beta < 1$, one can extend f_0 to $\mathbb{R}$ while keeping the Hölder property

  Let $\phi_\sigma(u) = \phi(u/\sigma)/\sigma$, for $\sigma > 0$ and ϕ Gaussian density

$$h_\sigma(t) := (\phi_\sigma * w_0)(t) = \int_{\mathbb{R}} \phi_\sigma(t-u) w_0(u) du$$

Properties of the convolution $h_\sigma(t) = \int_{\mathbb{R}} \phi_\sigma(t-u) w_0(u) du$

- Approximation of f_0

$$\begin{aligned} |\phi_\sigma * w_0(t) - w_0(t)| &= |\int \phi_\sigma(u)(w_0(t-u) - w_0(t)) du \\ &\lesssim \int \phi_\sigma(u) |u|^\beta du \lesssim \sigma^\beta \int |v|^\beta \phi(v) dv \lesssim \sigma^\beta. \end{aligned}$$

- it belongs to $\mathbb{H}$ since $\|h_\sigma\|_{\mathbb{H}}^2 = \int_0^1 (h_\sigma)'(t)^2 dt$ and

$$\begin{aligned} |(h_\sigma)'(t)| &= |\int w_0(t-u) \frac{1}{\sigma^2} \phi'(u/\sigma) du| \\ &= |\int (w_0(t-u) - w_0(t)) \frac{1}{\sigma^2} \phi'(u/\sigma) du| \quad \left(\text{as } \int \phi' = 0\right) \\ &\lesssim \sigma^{-2} \int |u|^\beta |\phi'(u/\sigma)| du \lesssim \sigma^{\beta-1}. \end{aligned}$$

  so that $\|h_\sigma\|_{\mathbb{H}}^2 \lesssim \sigma^{2\beta-2}$

The result follows by taking $\sigma \asymp \varepsilon^{1/\beta}$.

Using $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$: example of BM released at 0

Putting everything together, one gets

- Small ball probability $\varphi_0(\varepsilon) \asymp \varepsilon^{-2}$
- Approximation term $\lesssim \varepsilon^{\frac{2\beta-2}{\beta}\wedge 0}$

So $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$ if

$$\varepsilon_n^{-2} + \varepsilon_n^{\frac{2\beta-2}{\beta}\wedge 0} \le n\varepsilon_n^2$$

That is,

$$\varepsilon_n \ge n^{-\frac{\beta}{2}\wedge\frac{1}{4}}$$

This gives $\varepsilon_n \ge n^{-\frac{\alpha\wedge\beta}{2\alpha+1}}$ with $\alpha = 1/2$ ['regularity' of Brownian motion!]

GP posterior rates: examples (continued)

Some more examples, for β–smooth f_0 [slight variations on smoothness cond's]

- GP series prior with parameter $\alpha > 0$ [the 'hands-on' proof]

$$\varepsilon_n \leq n^{-\frac{\alpha\wedge\beta}{2\beta+1}}$$

- Riemann-Liouville α–process [mentioned before]

$$\varepsilon_n \leq n^{-\frac{\alpha\wedge\beta}{2\beta+1}}$$

- Matern α–process (Z_t), zero-mean with $E[Z_x Z_y] = \int e^{i\lambda(x-y)} m(\lambda) d\lambda$,

$$m(\lambda) = \frac{1}{(1+\lambda^2)^{\frac{1}{2}+\alpha}} \qquad \text{spectral density}$$

$$\varepsilon_n \leq n^{-\frac{\alpha\wedge\beta}{2\beta+1}}$$

Question: these are upper bounds, can one do better?

GP posterior rates: lower bounds

Question: these are upper bounds, can one do better?

The answer is ... no! in general

Lower bound result for α–series priors in white noise [C. 08]

$$E_{f_0}\Pi[\|f - f_0\|_2 \leq \zeta_n | X] \to 0$$

$$\zeta_n \gtrsim \begin{cases} n^{-\frac{\alpha}{2\alpha+1}} & \text{for any } \alpha < \beta \\ n^{-\frac{\beta}{2\alpha+1}} & \text{for some } \beta-\text{smooth } f_0,\ \text{if } \alpha < \beta \text{ [up to logs]} \end{cases}$$

GPs reach consistency but are optimal only under matched smoothness

GP posterior rates: lower bounds (continued)

Squared-exponential GP SqExp

Centered Gaussian process Z_t with covariance

$$E(Z_t Z_s) = e^{-(s-t)^2/L}$$

[van der Vaart, van Zanten 11] show that, for fixed L, there are regular functions f_0 for which the rate is at best *logarithmic*

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

Intuition: this is because SqExp is ‘infinitely smooth’!

GP posterior rates: lower bounds (continued)

Squared-exponential GP SqExp

Centered Gaussian process Z_t with covariance

$$E(Z_t Z_s) = e^{-(s-t)^2/L}$$

[van der Vaart, van Zanten 11] show that, for fixed L, there are regular functions f_0 for which the rate is at best *logarithmic*

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

Intuition: this is because SqExp is 'infinitely smooth'!

However, the use of SqExp is quite widespread and gives very good results in practice when the parameter L is "well chosen" ...

Adaptation to smoothness

A procedure is adaptive to smoothness if it achieves the optimal minimax rate $n^{-\beta/(2\beta+1)}$ for β–smooth f_0, simultaneously for any $\beta > 0$

As such GPs are too 'rigid' to get adaptation to smoothness

Idea(s): Tune an extra parameter to make them more flexible

- Idea 1: estimate α
- Idea 2: rescaling of paths

Adaptation to smoothness: idea 1, estimating α

Hierarchical Bayes Consider the hierarchical GP series prior Π

$$\alpha \sim \mathrm{Exp}(1)$$
$$f \,|\, \alpha \sim \Pi_\alpha \qquad \text{law of} \quad \sum_{k\geq 1} k^{-\frac{1}{2}-\alpha}\zeta_k\varphi_k(\cdot).$$

Empirical Bayes Projecting the white noise model onto the basis (φ_k), setting $Y_k = \int \varphi_k(u)dX^{(n)}(u)$ and $Y = (Y_k)$, the marginal distribution of $Y \,|\, \alpha$ is

$$Y \,|\, \alpha \sim \bigotimes_{k=1}^{\infty} \mathcal{N}\left(0, k^{-1-2\alpha} + \frac{1}{n}\right),$$

This gives a log-marginal likelihood

$$\ell_n(\alpha) = -\frac{1}{2}\sum_{k=1}^{\infty}\left(\log\left(1 + \frac{n}{k^{1+2\alpha}}\right) - \frac{n^2}{k^{1+2\alpha}+n}Y_k^2\right).$$

$$\hat{\alpha} = \underset{\alpha\in[0,\log n]}{\operatorname{argmax}}\ \ell_n(\alpha)$$

Adaptation to smoothness: idea 1, estimating α

[Knapik, Szabó, van der Vaart, van Zanten 16]

Theorem Suppose f_0 is β-Sobolev smooth in the white noise model

Hierarchical Bayes. The hierarchical prior verifies

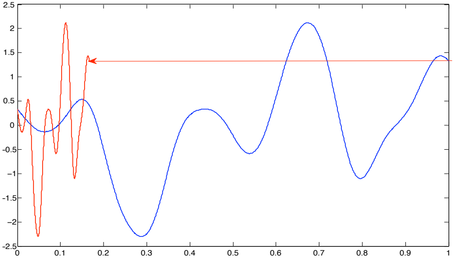
$$E_{f_0}\Pi[\|f-f_0\|_2 > (\log n)^l n^{-\frac{\beta}{2\beta+1}} \mid X] = o(1).$$

Empirical Bayes. The plug-in posterior $\Pi_{\hat{\alpha}}[\cdot \mid X]$ verifies

$$E_{f_0}\Pi_{\hat{\alpha}}[\|f-f_0\|_2 > (\log n)^l n^{-\frac{\beta}{2\beta+1}} \mid Y] = o(1).$$

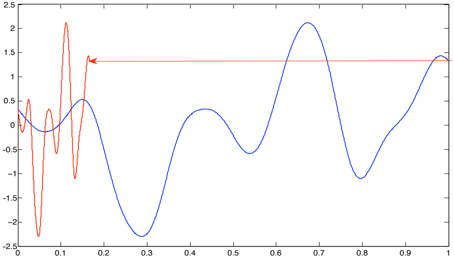
Adaptation to smoothness: idea 2, shrinking of paths

[van der Vaart, van Zanten 09]



Adaptation to smoothness: idea 2, shrinking of paths

[van der Vaart, van Zanten 09]



Prior Π: consider the process $t \to Z_{At}$

- $A \sim \pi_A$ Gamma distribution
- $u \to Z_u$ centered GP with squared-exponential kernel

Intuition Taking A large 'accelerates time' $\to$ makes path 'rougher'

Adaptation to smoothness: idea 2, shrinking of paths

White noise model prior $t \to Z_{At}$ leads to smoothness adaptation

$$E_{f_0}\Pi[\|f - f_0\|_2 > (\log n)^l n^{-\frac{\beta}{2\beta+1}} \,|\, X] = o(1).$$

Density estimation Set $t \to \dfrac{e^{Z_{At}}}{\int_0^1 e^{Z_{Au}}\,du}$

Then the posterior is also smoothness-adaptive up to a log factor

$$E_{f_0}\Pi\left[h(f, f_0) > (\log n)^l n^{-\frac{\beta}{2\beta+1}} \mid X\right] \to 1$$

Classification Similar results hold for estimating the classification function

$$x \to P[Y = 1 | X = x]$$

Idea of proof $$\Pi[\|f - f_0\|_2 \le \varepsilon_n] = \int \Pi[\|f - f_0\|_2 \le \varepsilon_n | A = a] d\pi_A(a)$$

Variable selection & dimension reduction with GPs

Variable selection

Setting: Density estimation or Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad X_i \sim G$$

Variable selection

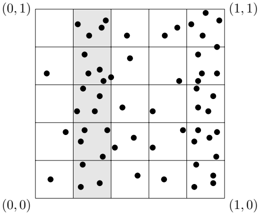
Setting: Density estimation or Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,\sigma^2), \quad X_i \sim G$$

If some variables have no effect on the response, i.e.:

$$f_0(x_1,\dots,x_D) \;=\; f_0(x_1,\dots,x_d), \quad D > d$$

we can still approximate the true parameter BUT...



Variable selection

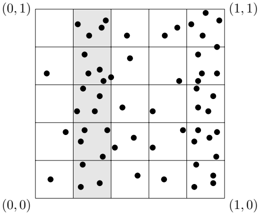
Setting: Density estimation or Regression with random design:

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad X_i \sim G$$

If some variables have no effect on the response, i.e.:

$$f_0(x_1, \ldots, x_D) = f_0(x_1, \ldots, x_d), \quad D > d$$

we can still approximate the true parameter BUT...



The contraction rate is suboptimal, of order $n^{-\frac{\beta}{2\beta+D}}$ instead of $n^{-\frac{\beta}{2\beta+d}}$.
This phenomenon is called the *curse of dimensionality*.

Variable selection

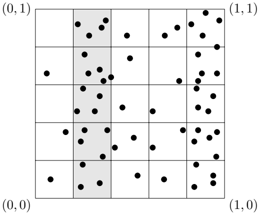
Setting: Density estimation or Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad X_i \sim G$$

If some variables have no effect on the response, i.e.:

$$f_0(x_1, \ldots, x_D) \;=\; f_0(x_1, \ldots, x_d), \quad D > d$$

we can still approximate the true parameter BUT...



The contraction rate is suboptimal, of order $n^{-\frac{\beta}{2\beta+D}}$ instead of $n^{-\frac{\beta}{2\beta+d}}$.
This phenomenon is called the *curse of dimensionality*.

Solutions:

- Hierarchical extension of Gaussian priors: the active variables are randomly selected.
- Freezing of paths

Given a *sparsity pattern* $\gamma \in \{0,1\}^D$, a γ-sparse rescaled squared exponential GP is defined as

$$W^{a,\gamma} := (W_0(ax_\gamma), x \in \mathbb{R}^D),$$

- $x_\gamma = (x_j, \gamma(j) = 1, j = 1, \ldots, D)$
- $a > 0$ rescaling parameter
- W_0 standard squared exponential GP in $\mathbb{R}^{|\gamma|}$.

[Jiang and Tokdar 2021]

Given a *sparsity pattern* $\gamma \in \{0,1\}^D$, a γ-sparse rescaled squared exponential GP is defined as

$$W^{a,\gamma} := (W_0(ax_\gamma), x \in \mathbb{R}^D),$$

- $x_\gamma = (x_j, \gamma(j) = 1, j = 1, \dots, D)$
- $a > 0$ rescaling parameter
- W_0 standard squared exponential GP in $\mathbb{R}^{|\gamma|}$.

Prior $\Pi \sim W^{A,\Gamma}$:

- $A^{|\Gamma|}$ Gamma distribution
- $\mathbb{P}(\Gamma = \gamma) = q(|\gamma|)/\binom{D}{|\gamma|}$, q probability vector on $[\![0, D]\!]$.

[Jiang and Tokdar 2021]

Given a *sparsity pattern* $\gamma \in \{0,1\}^D$, a γ-sparse rescaled squared exponential GP is defined as

$$W^{a,\gamma} := (W_0(ax_\gamma), x \in \mathbb{R}^D),$$

- $x_\gamma = (x_j, \gamma(j) = 1, j = 1, \dots, D)$
- $a > 0$ rescaling parameter
- W_0 standard squared exponential GP in $\mathbb{R}^{|\gamma|}$.

Prior $\Pi \sim W^{A,\Gamma}$:

- $A^{|\Gamma|}$ Gamma distribution
- $\mathbb{P}(\Gamma = \gamma) = q(|\gamma|)/\binom{D}{|\gamma|}$, q probability vector on $[\![0, D]\!]$.

For regression with Gaussian random design, if $f_0 \in H^\beta(\mathbb{R}^D) \cap L^2(G)$ has only d active variables, then for M large enough,

$$E_{f_0}\Pi\left[\|f - f_0\|_{L^2(G)} \geq M(\log n)^{\vartheta(\beta,d)} n^{-\frac{\beta}{2\beta+d}} \mid (X, Y)\right] \underset{n\to\infty}{\to} 0$$

Proof. (Idea)

Suppose f_0 has a sparsity pattern $\gamma \in \{0,1\}^D$ with $d := |\gamma|$.

Prior mass condition: $\Pi\left(\left\|W^{A,\Gamma} - f_0\right\|_\infty \le 2\varepsilon_n\right) \;\ge\; \exp(-n\varepsilon_n^2)$.

For T_n a carefully chosen constant, we can write,

$$
\begin{aligned}
\Pi\left(\left\|W^{A,\gamma} - f_0\right\|_\infty \le 2\varepsilon_n\right) \;&\ge\; \int_0^\infty \Pi(\|W^{a,\gamma} - f_0\|_\infty \le 2\varepsilon_n) \frac{d(A|\Gamma=\gamma)}{d\lambda}(a)da \\
&\ge\; \int_{T_n}^{2T_n} \exp\left(-\varphi_{\mathbf{f_0}}^{a,\gamma}(\varepsilon_n)\right) \frac{d(A|\Gamma=\gamma)}{d\lambda}(a)da \\
&\ge\; \exp\left(-C \cdot \varepsilon_n^{-d/\beta} \log(1/\varepsilon_n)^{d+1}\right) \\
\text{(Up to mult. constant on } \varepsilon_n\text{)} \quad &\ge\; \exp\left(-\frac{1}{2} n\varepsilon_n^2\right).
\end{aligned}
$$

Taking into account the prior on the sparsity pattern,

$$
\begin{aligned}
\Pi\left(\left\|W^{A,\Gamma} - f_0\right\|_\infty \le 2\varepsilon_n\right) \;&\ge\; \mathbb{P}(\Gamma = \gamma) \cdot \Pi\left(\left\|W^{A,\gamma} - f_0\right\|_\infty \le 2\varepsilon_n\right) \\
&\ge\; \exp\left(-n\varepsilon_n^2\right).
\end{aligned}
$$

Freezing of paths

The previous solution adds an extra layer to the model.
But a clever use of the rescaling step can serve the same purpose.

Freezing of paths

The previous solution adds an extra layer to the model.
But a clever use of the rescaling step can serve the same purpose.

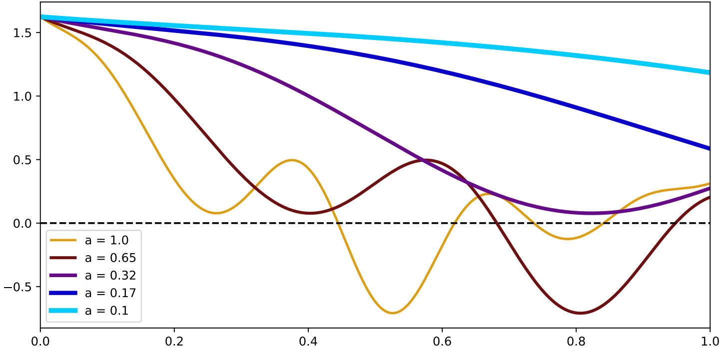
- [Castillo & Randrianarisoa 2024] propose a multi-bandwidth rescaling parameter (one for each coordinate), with a prior that encourages small length scales.

Freezing of paths

The previous solution adds an extra layer to the model.
But a clever use of the rescaling step can serve the same purpose.

- [Castillo & Randrianarisoa 2024] propose a multi-bandwidth rescaling parameter (one for each coordinate), with a prior that encourages small length scales.

A vanishing length scale in coordinate i 'freezes' the path in this direction.



Freezing of paths

Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad X_i \sim G$$

W standard squared exponential GP in $\mathbb{R}^D$.

Prior $\Pi \sim W^A$:

- A_j i.i.d. *exponential* distributions (places some probability mass near zero)
- $W^A = (W(A_1 x_1, \ldots, A_D x_D) : x = (x_1, \ldots, x_D) \in \mathbb{R}^D)$

Freezing of paths

Regression with random design:

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad X_i \sim G$$

W standard squared exponential GP in $\mathbb{R}^D$.

Prior $\Pi \sim W^A$:

- A_j i.i.d. *exponential* distributions (places some probability mass near zero)
- $W^A = (W(A_1 x_1, \ldots, A_D x_D) : x = (x_1, \ldots, x_D) \in \mathbb{R}^D)$

If $f_0 \in \mathcal{C}^\beta([0,1]^D)$, $\|f_0\|_\infty \leq Q$ has only d active variables, then for M large enough,

$$E_{f_0} \Pi_\rho \left[\|f - f_0\|_{L^2(G)} \geq M (\log n)^{\vartheta(\beta, d)} n^{-\frac{\beta}{2\beta + d}} \mid (X, Y) \right] \underset{n \to \infty}{\to} 0,$$

where $\Pi_\rho(\cdot | X, Y)$ is the fractional posterior of order $\rho < 1$.

Subspace selection

[Tokdar & Zhu & Ghosh 2010]

More general setting: f_0 depends only on a d-dimensional subspace of $\mathbb{R}^D$.

Subspace selection

[Tokdar & Zhu & Ghosh 2010]

More general setting: f_0 depends only on a d-dimensional subspace of $\mathbb{R}^D$.

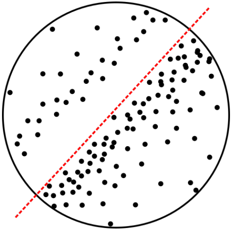


Figure: Sample from a 2-dimensional distribution whose density depends only on a one-dimensional subspace.

Subspace selection

[Tokdar & Zhu & Ghosh 2010]

More general setting: f_0 depends only on a d-dimensional subspace of $\mathbb{R}^D$.

Define,

$$W_x^{a,d,q} := W(a\,\mathrm{Diag}(d) \cdot q(x)).$$

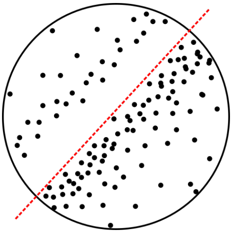


Figure: Sample from a 2-dimensional distribution whose density depends only on a one-dimensional subspace.

Subspace selection

[Tokdar & Zhu & Ghosh 2010]

More general setting: f_0 depends only on a d-dimensional subspace of $\mathbb{R}^D$.

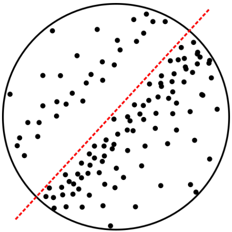


Figure: Sample from a 2-dimensional distribution whose density depends only on a one-dimensional subspace.

Define,

$$W_x^{a,d,q} := W(a\,\mathrm{Diag}(d) \cdot q(x)).$$

A hierarchical prior with stochastic subspace selection is:

$$\Pi \sim W^{A,\Gamma,\Theta},$$

where,

- A prior on the rescaling parameter a,
- Γ prior on the dimension of the subspace d,
- Θ prior on the isometry q.

Subspace selection

[Tokdar & Zhu & Ghosh 2010]

More general setting: f_0 depends only on a d-dimensional subspace of $\mathbb{R}^D$.

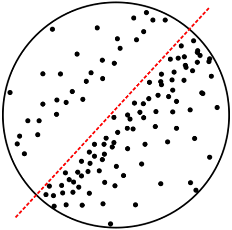


Figure: Sample from a 2-dimensional distribution whose density depends only on a one-dimensional subspace.

Define,

$$W_x^{a,d,q} := W(a\,\mathrm{Diag}(d) \cdot q(x)).$$

A hierarchical prior with stochastic subspace selection is:

$$\Pi \sim W^{A,\Gamma,\Theta},$$

where,

- A prior on the rescaling parameter a,
- Γ prior on the dimension of the subspace d,
- Θ prior on the isometry q.

$\rightarrow$ Same adaptive contraction rates.

Problem 1: How the ambient dimension D affects the contraction rate?

Growing dimension setting and subspace recovery

Problem 1: How the ambient dimension D affects the contraction rate?

Problem 2: Can we recover the relevant subspace or the sparsity pattern?

Growing dimension setting and subspace recovery

Problem 1: How the ambient dimension D affects the contraction rate?

Problem 2: Can we recover the relevant subspace or the sparsity pattern?

To address problem 1, we let the ambient dimension D grow with the number of observations n.

Growing dimension setting and subspace recovery

Problem 1: How the ambient dimension D affects the contraction rate?

Problem 2: Can we recover the relevant subspace or the sparsity pattern?

To address problem 1, we let the ambient dimension D grow with the number of observations n. (This means a new experiment and a new prior for each n.)

Growing dimension setting and subspace recovery

Problem 1: How the ambient dimension D affects the contraction rate?

Problem 2: Can we recover the relevant subspace or the sparsity pattern?

To address problem 1, we let the ambient dimension D grow with the number of observations n. (This means a new experiment and a new prior for each n.)

- **Variable selection:** [Jiang & Tokdar 2021]

  In the regression setting, with d active variables and $\log(D_n) \leq O\left(n^{\frac{d}{2d+\beta}}\right)$,
  - posterior contraction at near minimax rates to the true parameter f_0,
  - posterior consistency for the sparsity pattern.

Growing dimension setting and subspace recovery

Problem 1: How the ambient dimension D affects the contraction rate?

Problem 2: Can we recover the relevant subspace or the sparsity pattern?

To address problem 1, we let the ambient dimension D grow with the number of observations n. (This means a new experiment and a new prior for each n.)

- **Variable selection:** [Jiang & Tokdar 2021]

  In the regression setting, with d active variables and $\log(D_n) \leq O\left(n^{\frac{d}{2d+\beta}}\right)$,
  - posterior contraction at near minimax rates to the true parameter f_0,
  - posterior consistency for the sparsity pattern.

- **Subspace selection:** [preprint Odin & Bachoc & Lagnoux 2024]

  If the true parameter depends only on a subspace of dimension d,
  - Posterior consistency at near minimax rates with $D_n \leq O\left(n^{\frac{d}{2d+\beta}}\right)$,
  - With fixed ambient dimension D, the posterior contracts to the true subspace if d is known.

Statistical properties of GPs: other topics

Adaptation to smoothness (continued)

Hierarchical GPs with 1 parameter

- enable adaptation to global smoothness
- also enable variable selection
- based on impossibility results for plain GPs [Agapiou Wang 22]

  → one can conjecture that they are not adaptive to
  - spatially inhomogeneous smoothness
  - or more generally to 'local smoothness'

  For this could use heavy-tailed process [Agapiou C. 24]

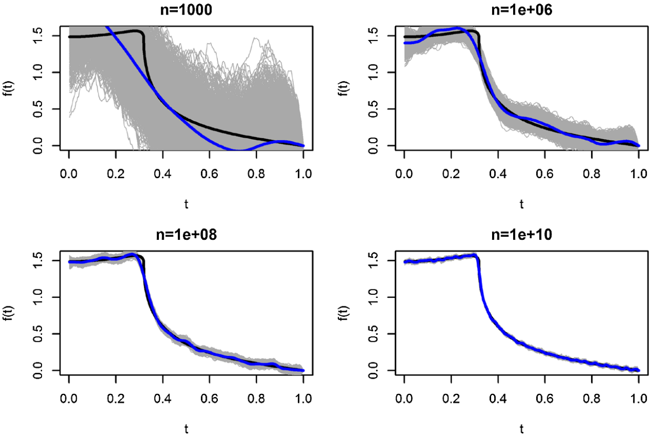
Adaptation to structure/geometry

→ next Section

Uncertainty quantification

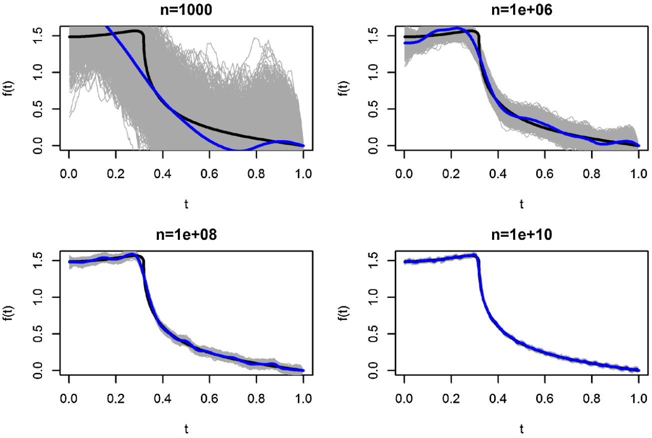
→ next 2 slides

Statistical properties of GPs: uncertainty quantification



[Szabó, van der Vaart, van Zanten 15]

Statistical properties of GPs: uncertainty quantification

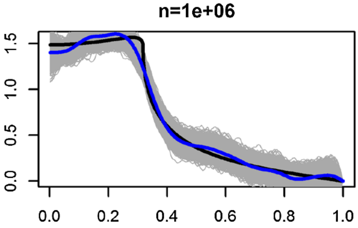


[Szabó, van der Vaart, van Zanten 15]

Are credible sets $\mathcal{C}(X)$ also confidence sets?

$$\Pi[\mathcal{C}(X) \,|\, X] \approx 1-\alpha \qquad \overset{??}{\Rightarrow} \qquad P_{f_0}[f_0 \in \mathcal{C}(X)] \approx 1-\alpha$$

Statistical properties of GPs: uncertainty quantification



Ideally, a credible set $\mathcal{C}_n = \mathcal{C}_n(X)$ (i.e. $\Pi[\mathcal{C}_n(X) \mid X] = 1 - \alpha$) should have

- coverage $P_{f_0}[f_0 \in \mathcal{C}_n] \approx 1 - \alpha$
- adaptive optimal minimax diameter $\mathrm{Diam}_d(\mathcal{C}_n) \approx n^{-\beta/(2\beta+1)}$

However, this is known to be impossible, unless more is assumed on f

This becomes possible under self-similarity type assumptions on f_0

Intuition Self-similarity enables to 'estimate regularity' of f_0

Statistical properties of GPs: uncertainty quantification

A few works in this direction for $f \sim$ GP

- [Szabó, van der Vaart, van Zanten 15]
  UQ: adaptive L^2 confidence sets under self-similarity for series GPs
- [Sniekers, van der Vaart 15]
  UQ: pointwise confidence sets under self-similarity for Brownian motion
- [Hadji, Szabó 21]
  UQ: adaptive L^2 confidence sets under self-similarity for SqExp

One may also be interested in estimating functionals $f \to \psi(f)$, BvM-type results

- [C. 12], [C. & Rousseau 15]
  Bernstein–von Mises theorems for smooth functionals ($\asymp$ Bayesian CLT)

$$\mathcal{L}(\psi(f) \in \cdot \,|\, X) \approx \mathcal{N}(\hat{\psi}, \mathcal{I}_{\psi}^{-1}/n)(\cdot)$$

[implies quantile credible sets are (asymp) confidence sets]

1 Introduction

2 Statistical properties of GPs I

3 Statistical properties of GPs II
- GPs and geometry: the intrinsic approach
- GPs and geometry: the extrinsic approach
- Deep GPs

4 Scalable GPs: approximations and surrogates

3. Statistical properties of GPs II

GPs and geometry

Intrinsic GP

Goal: Infer a regression function over a known manifold $\mathcal{M}$ with adaptation to regularity.

Image source: [Rosa & Terenin & Borovitskiy & Rousseau 2023].

Intrinsic GP

Goal: Infer a regression function over a known manifold $\mathcal{M}$ with adaptation to regularity.

Idea: Extend the squared exponential GP with random rescaling to a non-Euclidean input space.

Image source: [Rosa & Terenin & Borovitskiy & Rousseau 2023].

Intrinsic GP

Goal: Infer a regression function over a known manifold $\mathcal{M}$ with adaptation to regularity.

Idea: Extend the squared exponential GP with random rescaling to a non-Euclidean input space.

$\rightarrow$ If the manifold is Riemannian, replace the Euclidean distance by the *geodesic distance* ρ in

$$K(x, y) \ := \ \exp(-\rho(x, y)^2), \quad x, y \in \mathcal{M}^2.$$

Image source: [Rosa & Terenin & Borovitskiy & Rousseau 2023].

Intrinsic GP

Goal: Infer a regression function over a known manifold $\mathcal{M}$ with adaptation to regularity.

Idea: Extend the squared exponential GP with random rescaling to a non-Euclidean input space.

$\rightarrow$ If the manifold is Riemannian, replace the Euclidean distance by the *geodesic distance* ρ in

$$K(x, y) \;:=\; \exp(-\rho(x, y)^2), \quad x, y \in \mathcal{M}^2.$$

Problem: In most cases, $K(\cdot, \cdot)$ fails to be positive definite.

Image source: [Rosa & Terenin & Borovitskiy & Rousseau 2023].

Intrinsic GP

Goal: Infer a regression function over a known manifold $\mathcal{M}$ with adaptation to regularity.

Idea: Extend the squared exponential GP with random rescaling to a non-Euclidean input space.

$\rightarrow$ If the manifold is Riemannian, replace the Euclidean distance by the *geodesic distance* ρ in

$$K(x, y) \ := \ \exp(-\rho(x, y)^2), \quad x, y \in \mathcal{M}^2.$$

Problem: In most cases, $K(\cdot, \cdot)$ fails to be positive definite.

- Instead, build a positive definite kernel from a linear operator.

Image source: [Rosa & Terenin & Borovitskiy & Rousseau 2023].

Positive definite kernel via the Laplacian

[Castillo, Kerkyacharian, Picard 14]

On $\mathcal{M}$ compact Riemannian manifold of dimension d without boundary.

$\rightarrow$ Laplacian $\Delta_{\mathcal{M}}$ linear operator on functions on $\mathcal{M}$ with discrete spectrum

$$(-\Delta_{\mathcal{M}})\varphi_p \;=\; \lambda_p \varphi_p$$

$$0 \le \lambda_1 \le \lambda_2 \le \cdots$$

Positive definite kernel via the Laplacian

[Castillo, Kerkyacharian, Picard 14]

On $\mathcal{M}$ compact Riemannian manifold of dimension d without boundary.

$\rightarrow$ Laplacian $\Delta_{\mathcal{M}}$ linear operator on functions on $\mathcal{M}$ with discrete spectrum

$$(-\Delta_{\mathcal{M}})\varphi_p = \lambda_p \varphi_p$$

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$$

For clarity, we suppose that the eigenspaces $\mathcal{H}_{\lambda_p}$ are of dimension one.

Positive definite kernel via the Laplacian

[Castillo, Kerkyacharian, Picard 14]

On $\mathcal{M}$ compact Riemannian manifold of dimension d without boundary.

$\rightarrow$ Laplacian $\Delta_{\mathcal{M}}$ linear operator on functions on $\mathcal{M}$ with discrete spectrum

$$(-\Delta_{\mathcal{M}})\varphi_p \;=\; \lambda_p \varphi_p$$

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$$

For clarity, we suppose that the eigenspaces $\mathcal{H}_{\lambda_p}$ are of dimension one.

We have the decomposition,

$$L^2(\mathcal{M}) \;=\; \bigoplus_{p \geq 1} \mathcal{H}_{\lambda_p},$$

and the orthogonal projectors $P_{\mathcal{H}_{\lambda_p}}$ on $\mathcal{H}_{\lambda_p}$ are kernel operators $Q_p(x, y)$,

$$Q_p(x,y) \;=\; \varphi_p(x)\varphi_p(y). \qquad \text{(Positive definite)}$$

Positive definite kernel via the Laplacian

Consequence: The mixed kernel $K(x,y) := \sum_{p\geq 1} \sigma_p Q_p(x,y)$ is positive definite and is associated with the Gaussian process

$$W := \sum_{p\geq 1} \sqrt{\sigma_p}\zeta_p\varphi_p, \quad (\zeta_p) \text{ i.i.d. } \mathcal{N}(0,1).$$

Image source: [Lang & Motschan-Armen 2023]

Positive definite kernel via the Laplacian

Consequence: The mixed kernel $K(x,y) := \sum_{p\geq 1} \sigma_p Q_p(x,y)$ is positive definite and is associated with the Gaussian process

$$W \ := \ \sum_{p\geq 1} \sqrt{\sigma_p}\zeta_p\varphi_p, \quad (\zeta_p) \text{ i.i.d. } \mathcal{N}(0,1).$$

Question: How to choose the σ_ps and how to rescale the sample paths? (multiplicative rescaling has no sense on manifolds)

Image source: [Lang & Motschan-Armen 2023]

Positive definite kernel via the Laplacian

Consequence: The mixed kernel $K(x,y) := \sum_{p\geq 1} \sigma_p Q_p(x,y)$ is positive definite and is associated with the Gaussian process

$$W \; := \sum_{p\geq 1} \sqrt{\sigma_p}\zeta_p\varphi_p, \quad (\zeta_p) \text{ i.i.d. } \mathcal{N}(0,1).$$

Question: How to choose the σ_ps and how to rescale the sample paths? (multiplicative rescaling has no sense on manifolds)

Consider solutions of the heat equation on $\mathcal{M}$ and use the time t as a scale parameter.

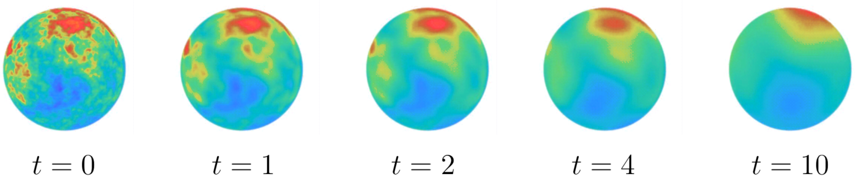


Image source: [Lang & Motschan-Armen 2023]

Positive definite kernel via the Laplacian

Note that

$$\Delta_{\mathcal{M}}\left(e^{-\lambda_p t}\varphi_p\right) = -\lambda_p e^{-\lambda_p t}\varphi_p$$
$$\frac{\partial}{\partial t} e^{-\lambda_p t}\varphi_p = -\lambda_p e^{-\lambda_p t}\varphi_p$$

This is a special solution of the heat equation on $\mathcal{M}$

$$\Delta_{\mathcal{M}} f = \frac{\partial}{\partial t} f$$

Positive definite kernel via the Laplacian

Note that

$$\Delta_{\mathcal{M}}\left(e^{-\lambda_p t}\varphi_p\right) = -\lambda_p e^{-\lambda_p t}\varphi_p$$
$$\frac{\partial}{\partial t} e^{-\lambda_p t}\varphi_p = -\lambda_p e^{-\lambda_p t}\varphi_p$$

This is a special solution of the heat equation on $\mathcal{M}$

$$\Delta_{\mathcal{M}} f = \frac{\partial}{\partial t} f$$

Let $\zeta_p \sim \mathcal{N}(0,1)$ i.i.d. A GP "random solution of the heat equation" is

$$W^t := \sum_{p\geq 1} e^{-\lambda_p t/2}\zeta_p\varphi_p$$

The associated family of covariance kernels is

$$P_t(x,y) = \sum_{p\geq 1} e^{-\lambda_p t}\varphi_p(x)\varphi_p(y) \qquad \Delta_{\mathcal{M}}\text{-Heat Kernel}$$

Positive definite kernel via the Laplacian

Let $\zeta_p \sim \mathcal{N}(0,1)$ i.i.d. A GP "random solution of the heat equation" is

$$W^t := \sum_{p \geq 1} e^{-\lambda_p t/2} \zeta_p \varphi_p$$

The associated family of covariance kernels is

$$P_t(x,y) = \sum_{p \geq 1} e^{-\lambda_p t} \varphi_p(x) \varphi_p(y) \qquad \Delta_{\mathcal{M}}\text{-Heat Kernel}$$

Subgaussian estimates:

$$\frac{c_1 e^{-c' \frac{\rho^2(x,y)}{t}}}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} \leq P_t(x,y) \leq \frac{c_2 e^{-c' \frac{\rho^2(x,y)}{t}}}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}}$$

- The heat kernel is a natural geometric generalization of the squared-exponential kernel
- The time t is a natural candidate for a scale parameter.

Adaptation on manifolds

White noise model: [Castillo, Kerkyacharian, Picard 14]

$$dX^{(n)}(x) = f(x)dx + dZ(x), \quad x \in \mathcal{M}$$

Prior Π :

- $W^T = \sum_p e^{-\lambda_p T/2} \zeta_p \varphi_p$ with $T \sim t^{-a} e^{-t^{-d/2} \log^q(1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Adaptation on manifolds

White noise model: [Castillo, Kerkyacharian, Picard 14]

$$dX^{(n)}(x) = f(x)dx + dZ(x), \quad x \in \mathcal{M}$$

Prior Π :

- $W^T = \sum_p e^{-\lambda_p T/2} \zeta_p \varphi_p$ with $T \sim t^{-a} e^{-t^{-d/2} \log^q(1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Suppose $f_0 \in B^{\beta}_{2,\infty}(\mathcal{M})$. Then for M large enough, as $n \to \infty$,

$$E_{f_0} \Pi \left[\|f - f_0\|_2 \geq M \left(\frac{\log n}{n} \right)^{\beta/(2\beta+d)} \mid X \right] \to 0.$$

The rate is sharp

Adaptation on manifolds

White noise model: [Castillo, Kerkyacharian, Picard 14]

$$dX^{(n)}(x) = f(x)dx + dZ(x), \quad x \in \mathcal{M}$$

Prior Π :

- $W^T = \sum_p e^{-\lambda_p T/2}\zeta_p\varphi_p$ with $T \sim t^{-a}e^{-t^{-d/2}\log^q(1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Suppose $f_0 \in B^\beta_{2,\infty}(\mathcal{M})$. Then for M large enough, as $n \to \infty$,

$$E_{f_0}\Pi\left[\|f - f_0\|_2 \geq M\left(\frac{\log n}{n}\right)^{\beta/(2\beta+d)} \mid X\right] \to 0.$$

The rate is sharp for small enough ρ, there exists f_0 in $B^\beta_{2,\infty}(\mathcal{M})$,

$$\Pi\left[\|f - f_0\|_2 \leq \rho\left(\frac{\log n}{n}\right)^{\beta/(2\beta+d)} \mid X\right] \to 0$$

Extrinsic GP

[Yang & Dunson 2016]

Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$X_i \sim G$, with support in $\mathcal{M}$.

Suppose:

- Unknown manifold $\mathcal{M}$
- Embedded in $\mathbb{R}^D$
- Known intrinsic dimension d

Extrinsic GP

[Yang & Dunson 2016]

Regression with random design:

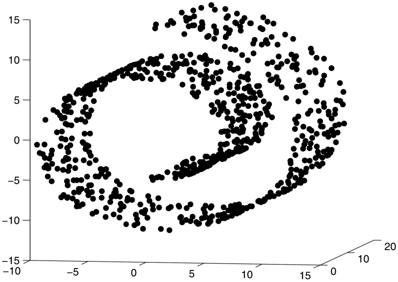
$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$X_i \sim G$, with support in $\mathcal{M}$.

Suppose:

- Unknown manifold $\mathcal{M}$
- Embedded in $\mathbb{R}^D$
- Known intrinsic dimension d

$\rightarrow$ Extrinsic approach



Extrinsic GP

[Yang & Dunson 2016]

Regression with random design:

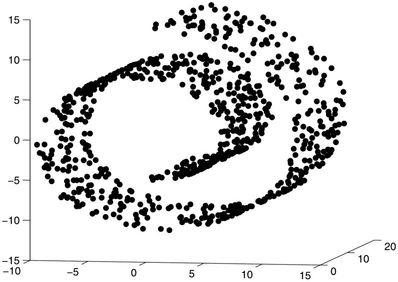
$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$X_i \sim G$, with support in $\mathcal{M}$.

Suppose:

- Unknown manifold $\mathcal{M}$
- Embedded in $\mathbb{R}^D$
- Known intrinsic dimension d

$\rightarrow$ Extrinsic approach



We consider the squared exponential GP $(W_x : x \in \mathcal{M})$ with

$$K(x, y) \;=\; \exp(-\|x - y\|^2/2), \qquad (\|\cdot\| \text{ is the Euclidean norm in } \mathbb{R}^D)$$

Extrinsic GP

For $a > 0$ length scale parameter, $(W_x^a : x \in \mathcal{M})$ with

$$K^a(x,y) \;=\; \exp(-a^2\|x-y\|^2/2), \qquad (\|\cdot\| \text{ is the Euclidean norm in } \mathbb{R}^D)$$

Prior Π :

- W^A with A^d Gamma distribution
- W^A is a GP on $\mathcal{M}$

Extrinsic GP

For $a > 0$ length scale parameter, $(W_x^a : x \in \mathcal{M})$ with

$$K^a(x,y) \;=\; \exp(-a^2\|x-y\|^2/2), \qquad (\|\cdot\| \text{ is the Euclidean norm in } \mathbb{R}^D)$$

Prior Π :

- W^A with A^d Gamma distribution
- W^A is a GP on $\mathcal{M}$

Suppose $\mathcal{M}$ is a compact γ-differentiable submanifold of $\mathbb{R}^D$ of dimension d.
If $f_0 \in \mathcal{C}^\beta(\mathcal{M})$ with $\beta \le \min\{2, \gamma - 1\}$, then for M large enough,

$$E_{f_0}\Pi\left[\|f^Q - f_0^Q\|_{L^2(G)} \ge M(\log n)^{\vartheta(\beta,d)} n^{-\frac{\beta}{2\beta+d}} \mid (X,Y)\right] \underset{n\to\infty}{\to} 0$$

where $f^Q := (f \vee -Q) \wedge Q$ is the truncated version of f.

Extrinsic GP

- Estimating the intrinsic dimension d leads to an empirical Bayes procedure.
- The restriction $\beta \leq 2$ is caused by approximating the geodesic distance by the Euclidean distance.

Comparison Intrinsic vs Extrinsic

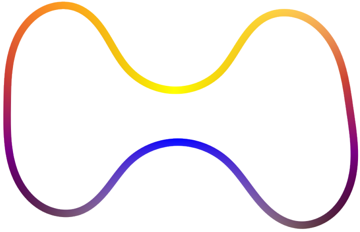


Figure: Realization of an intrinsic GP.

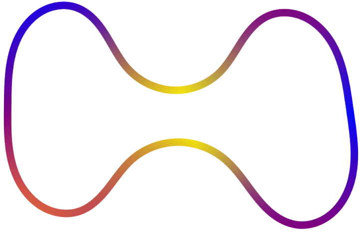


Figure: Realization of an extrinsic GP.

Comparison Intrinsic vs Extrinsic

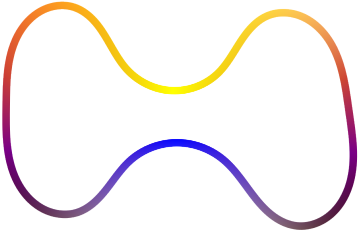


Figure: Realization of an intrinsic GP.

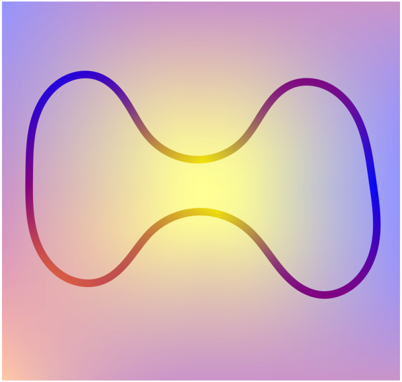


Figure: Realization of an extrinsic GP.

Comparison Intrinsic vs Extrinsic

[Rosa & Terenin & Borovitskiy & Rousseau 2023]

Regression with random design:

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$X_i \sim G$, with support in a d-dimensional manifold $\mathcal{M}$.

- **Prior Π_{int} :** W an intrinsic Riemannian Matérn Gaussian process with smoothness parameter $\nu > d/2$
- **Prior Π_{ext} :** W' an extrinsic Matérn Gaussian process with smoothness parameter $\nu > d/2$

Comparison Intrinsic vs Extrinsic

[Rosa & Terenin & Borovitskiy & Rousseau 2023]

Regression with random design:

$$Y_i \;=\; f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$X_i \sim G$, with support in a d-dimensional manifold $\mathcal{M}$.

- **Prior $\Pi_{\mathbf{int}}$:** W an intrinsic Riemannian Matérn Gaussian process with smoothness parameter $\nu > d/2$
- **Prior $\Pi_{\mathbf{ext}}$:** W' an extrinsic Matérn Gaussian process with smoothness parameter $\nu > d/2$

If $f_0 \in H^\beta(\mathcal{M}) \cap B^\beta_{\infty,\infty}(\mathcal{M})$, for $\beta > d/2$, then for M, M' large enough,

$$E_{f_0} \Pi_{\text{int}} \left[\|f - f_0\|^2_{L^2(G)} \;\leq\; M \cdot n^{-\frac{2\min(\beta,\nu)}{2\nu+d}} \mid X, Y \right] \underset{n\to\infty}{\to} 0,$$

$$E_{f_0} \Pi_{\text{ext}} \left[\|f - f_0\|^2_{L^2(G)} \;\leq\; M \cdot n^{-\frac{2\min(\beta,\nu)}{2\nu+d}} \mid X, Y \right] \underset{n\to\infty}{\to} 0.$$

Deep Gaussian processes

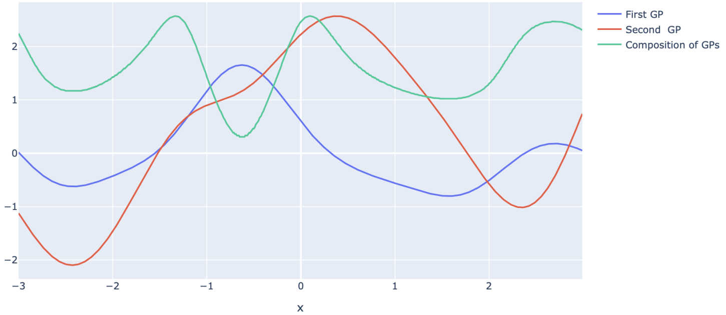
Deep Gaussian processes

A deep GP is a composition of GPs: for some $q \geq 1$ and g_q Gaussian processes,

$$f = g_q \circ \cdots \circ g_0$$

The GPs at each 'layer' are multidimensional $g_i = (g_{ij})_j$ and g_{ij} multivariate

[Damianou, Lawrence 13]



Remark $\quad f = \Psi(g_q) \circ \cdots \circ \Psi(g_0) \qquad \Psi(x) = (-K) \vee (x \wedge K)$

Deep GPs: sampling

What do people do in practice?

- [Damianou et al. 13]
  Automatic Relevance Detection (ARD) kernel on each layer of deep GP

$$K(s,t) = e^{-\sum_{i=1}^{d} A_i^2 (s_i - t_i)^2}$$

  - this is independent product of SqExp GPs with inv. lengthscales A_i
  - A_i determined by empirical Bayes-type criterion / variational Bayes (VB)
  - many follow-up works using VB/inducing points

- [Sauer et al. 24]
  `deepgp` R package
  - ARD kernel on each layer
  - with random (Gamma) A_i (hierarchical Bayes)
  - uses Vecchia approximation + MCMC

Deep Gaussian processes, theory

[Finocchio & Schmidt-Hieber, JMLR 23] To make the deep GP 'adaptive'

- discrete model selection prior [determines 'active' directions]
- draw regularity (e.g. lengthscale) at random
- condition sample paths to be smooth enough

Optimal minimax posterior contraction rate (up to logs), adaptive to unknown regularities and compositional structure $\rightarrow$ Great foundational result on deep GPs!

Deep Gaussian processes, theory

[Finocchio & Schmidt-Hieber, JMLR 23] To make the deep GP 'adaptive'

- discrete model selection prior [determines 'active' directions]
- draw regularity (e.g. lengthscale) at random
- condition sample paths to be smooth enough

Optimal minimax posterior contraction rate (up to logs), adaptive to unknown regularities and compositional structure $\rightarrow$ Great foundational result on deep GPs!

However, practical simulation from $\Pi[\cdot \,|\, X, Y]$ not so simple

- need to condition on GP verifying restrictions (e.g. bounded $\mathcal{C}^{\beta_i}$ norms)
- need to sample from complex iterative model selection posterior

[Moriarty-Osborne & Teckentrup 25] direct approach for posterior mean

- use recursion to study posterior mean of deep GP
- optimal rate for posterior mean in noiseless case

Conor's talk yesterday!

Questions

- Theoretical support without discrete variable selection?
- for (/close to) above practically used deep GP priors
- possibly for deep compositional structure, and high dimensional regression

[C. & Randrianarisoa 25]   ANR GAP-project paper!

New idea

Perform simultaneous

- smoothness adaptation
- *soft* variable selection

through appropriate prior on inverse lengthscale parameters A_i (!)

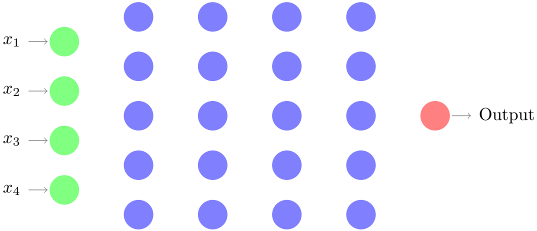
- $A_i \to \infty$ appropriately fast $\to$ rescaling allowing for adaptation

  Shrinking of paths

- $A_i \to 0$ fast enough $\to$ the GP is near-constant on that coordinate

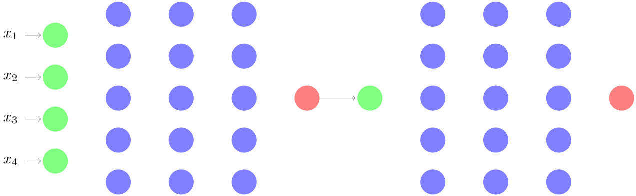
  Freezing of paths (!)

Link with standard neural networks



In the infinite width limit, get Gaussian Process (GP)

Link with standard neural networks



In the infinite width limit, get deep Gaussian Process (GP) [here 2 layers]

Simple model: dimension reduction

Suppose the regression function f_0 can be written

$$f_0(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_{d^*}})$$

where

- $1 \leq d^* \leq d$
- g is β–Hölder

Target estimation rate is $n^{-\frac{\beta}{2\beta+d^*}}$

Single GP

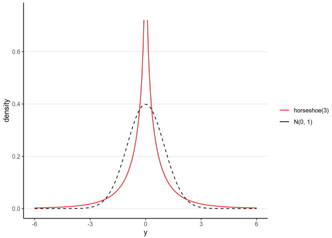
Prior Π Denote $W^A(x) = W(A_1 x_1, \dots, A_d x_d)$

$$\begin{aligned} f \mid A &\sim W^A \\ A_i &\overset{\text{i.i.d.}}{\sim} \pi \end{aligned}$$

with π prior on inverse lengthscale

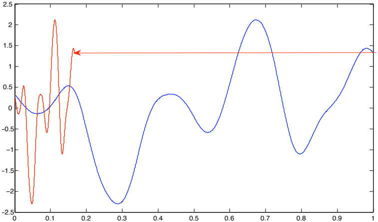
Consider two choices of π

- exponential prior $\mathcal{E}(\lambda)$
- (half-)horseshoe prior $\mathcal{H}(\tau)$

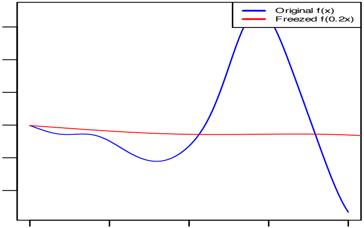


Stretching vs Freezing

Stretching of paths inverse lengthscale A is large [van der Vaart, van Zanten 09]



Freezing of paths inverse lengthscale A is close to 0



Contraction rates [1-layer variable selection]

$$f_0(x_1, \dots, x_d) = g(x_{i_1}, \dots, x_{i_{d^*}})$$

Fractional posterior i.e. "ρ–posterior"

$$\Pi^\rho[B|\, X, Y] = \frac{\int_B \prod_{1\le i\le n} p_f(X_i, Y_i)^{\rho} d\Pi(f)}{\int \prod_{1\le i\le n} p_f(X_i, Y_i)^{\rho} d\Pi(f)}, \qquad 0 < \rho < 1$$

Theorem 1. Suppose f_0 is β–Hölder as above. For Π GP prior with prior π on random inverse-lengthscales either exponential or horseshoe. For $0 < \rho < 1$, as $n \to \infty$

$$E_{f_0}\Pi_\rho \left[f : \ \|f - f_0\|_{L^2(\mu)} \ge \varepsilon_n |\, X, Y \right] \to 0$$

$$\varepsilon_n = (\log n)^c n^{-\frac{\beta}{2\beta + d^*}}$$

Optimal contraction rate (up to log) achieved without discrete model selection prior $\to$ soft selection automatically achieved by small A's [freezing of paths]

Compositional structures and deep GPs

Compositional structure

Compositional class. Let $I := [-1, 1]$. Consider $f : I^d \to I$ with

$$f = h_q \circ \cdots \circ h_0,$$

where $h_i : I^{d_i} \to I^{d_{i+1}}$, with $d_0 = d$, $d_{q+1} = 1$ and

- writing $h_i = (h_{ij})$ for $1 \le j \le d_{i+1}$,
- assume all h_{ij}'s depend at most on $t_i \le d_i$ variables
- and $\|h_{ij}\|_{\beta_i,\infty} \le K$

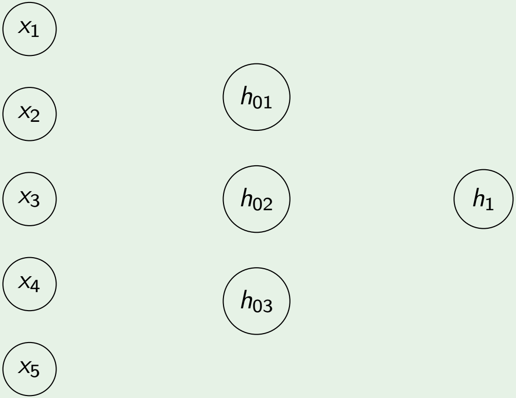
Defines a class $\mathcal{F} = \mathcal{F}(\lambda, \beta, K)$ with

$$\lambda = (q, d_1, \ldots, d_q, t_0, \ldots, t_q), \quad \beta = (\beta_0, \ldots, \beta_q).$$

Graph representation

Example

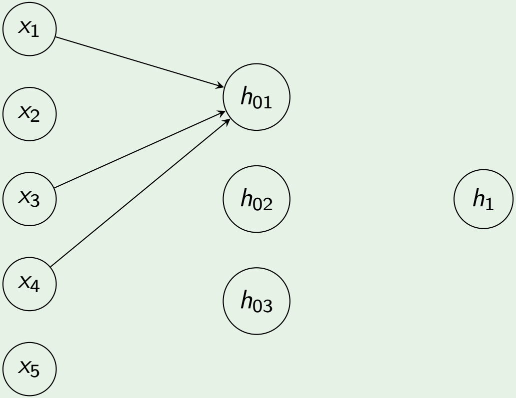
Ex: $f^*(x_1, \ldots, x_5) = h_1\left(h_{01}(x_1, x_3, x_4), h_{02}(x_1, x_4, x_5), h_{03}(x_2)\right)$



Graph representation

Example

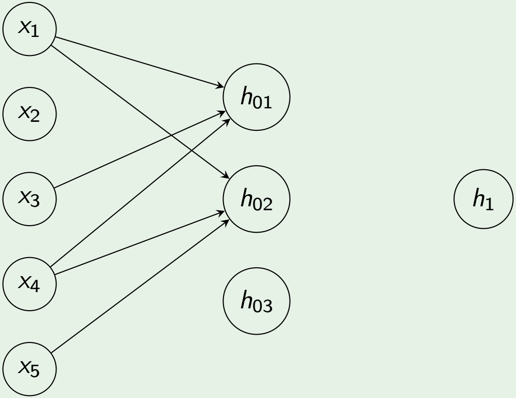
Ex: $f^*(x_1, \ldots, x_5) = h_1\left(h_{01}(x_1, x_3, x_4), h_{02}(x_1, x_4, x_5), h_{03}(x_2)\right)$



Graph representation

Example

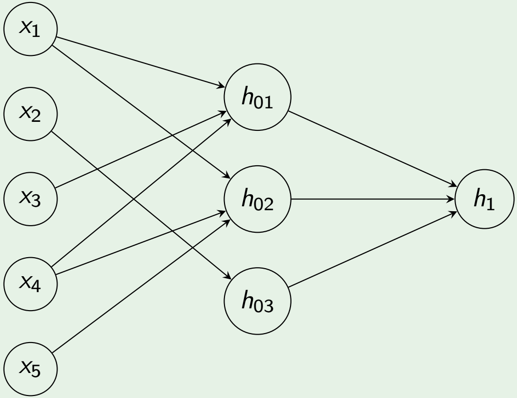
<u>Ex:</u> $f^*(x_1, \ldots, x_5) = h_1\left(h_{01}(x_1, x_3, x_4), h_{02}(x_1, x_4, x_5), h_{03}(x_2)\right)$



Graph representation

Example

<u>Ex:</u> $f^*(x_1, \dots, x_5) = h_1\left(h_{01}(x_1, x_3, x_4), h_{02}(x_1, x_4, x_5), h_{03}(x_2)\right)$



Minimax rate for compositional classes

Compositional class. Let $f = h_q \circ \cdots \circ h_0$ where $h_i = (h_{ij})$ for $1 \leq j \leq d_{i+1}$,

- assume all h_{ij}'s depend at most on $t_i \leq d_i$ variables
- and $\|h_{ij}\|_{\beta_i,\infty} \leq K$

Minimax rate. [if $t_i \leq \max(d_0, \ldots, d_{i-1})$] minimax rate over class $\mathcal{F}$ for $L^2(\mu)$ loss

$$r_n^* \asymp \max_{i=0,\ldots,q} n^{-\frac{\beta_i^*}{2\beta_i^*+t_i}}, \qquad \beta_i^* = \beta_i \cdot \prod_{j=i+1}^{q} (\beta_j \wedge 1)$$

- t_i 'effective dimension' at depth i
- β_i^* 'effective regularity' at depth i

Deep Horseshoe Gaussian Process

Let q_{max}, d_{max} deterministic 'large' [overparametrised]

Set $q = q_{max}$ and $d_1 = \cdots = d_q = d_{max}$

The *Deep Horseshoe GP* Deep–HGP is defined as the hierarchical prior

$$
\begin{aligned}
A_{ij} \mid q, d_1, \ldots, d_q &\overset{\text{i.i.d.}}{\sim} \pi^{\otimes d_i} \\
g_{ij} \mid q, d_1, \ldots, d_q, A_{ij} &\overset{\text{i.i.d.}}{\sim} W^{A_{ij}} \\
f \mid q, d_1, \ldots, d_q, g_{ij} &= \Psi(g_q) \circ \cdots \circ \Psi(g_0).
\end{aligned}
$$

with $\pi = \pi_\tau$ horseshoe prior with parameter τ [or π an exponential $\mathcal{E}(\lambda)$ prior]

Remark (for theory): could also take q and d_i's random

Contraction rates for deep HGP

Suppose $f_0 \in \mathcal{F} = \mathcal{F}(\lambda, \beta, K)$, and $q \leq q_{max}, d_i \leq d_{max}$

Theorem 2. For Π Deep-HGP prior, take π either exponential or horseshoe prior on inverse lengthscale. For $0 < \rho < 1$,

$$E_{f_0}\Pi_\rho\left[f:\ \|f - f_0\|_{L^2(\mu)} \geq \varepsilon_n | X, Y\right] \to 0$$

$$\varepsilon_n = (\log n)^{c_2} r_n^* = (\log n)^{c_2} \max_i n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}}$$

Deep HGP prior achieves optimal rate, with simultaneous adaptation to smoothness and compositional structure

More generally for prior π on scalings it suffices that

$$\pi(x) \gtrsim x^{c_1} \exp\left(-C_2 x^{d^*} \log^{c_2} x\right)$$

Allows for $\pi = \mathrm{Ga}(\lambda)$ as in R package deepgp

Comments on deep HGP vs deep model selection GP

There are two main simplifications with deep HGP

1. One does not need to separately
   - sample of regularities
   - sample of the graph and dimension structure

   This is simultaneously done with the horseshoe prior
   - large A_{ij} picks correct 'regularity' scaling
   - small A_{ij} makes variable i of jth layer (nearly) inactive

2. One does not need to
   - restrict paths to be β–Hölder
   - penalise the complexity of the prior

   Use the ρ–posterior $\to$ no need for entropy condition

Comments on deep HGP vs deep model selection GP

There are two main simplifications with deep HGP

1. One does not need to separately
   - sample of regularities
   - sample of the graph and dimension structure

   This is simultaneously done with the horseshoe prior
   - large A_{ij} picks correct 'regularity' scaling
   - small A_{ij} makes variable i of jth layer (nearly) inactive

2. One does not need to
   - restrict paths to be β–Hölder
   - penalise the complexity of the prior

   Use the ρ–posterior $\to$ no need for entropy condition

So far results do not differenciate horseshoe and exponential choices for π

High dimensional regression

$$f_0(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_{d^*}})$$

- Want to allow $d \to \infty$ and possibly also $d^* \to \infty$ (slowly)
- Better, fully non-asymptotic result in terms of d, d^*, n [in paper!]
- Special case $\tau = \mathcal{E}(\lambda)$ prior on inverse lengthscale
  The posterior rate ε_n should verify (up to logs)

$$n\varepsilon_n^2 \geq C \left[d + \lambda n^{\frac{1}{2\beta + d^*}} \right]$$

Suboptimality already for fixed d^* and $d = n^a$ for any $a > 0$

Contraction rates for HGP II

$$f_0(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_{d^*}})$$

Setting: high ambient dimension $d = d_n \to \infty$, $d^* \leq d$ with

$$d = d_n = O(n^c)\ ,\ d^* = d_n^* = O(\log^{1/2-\delta} n)$$

Theorem 3 Let Π be Deep-HGP as before, with horseshoe $\mathcal{H}(\tau)$ prior and

$$\tau = (nd^2)^{-1}$$

Consider ρ–posterior for $\rho \in (0, 1)$

- High dimensional variable selection ($f_0 : [-1, 1]^d \to \mathbb{R}$, $\mathcal{C}^\beta$) One-layer HGP attains optimal $L^2(\mu)$-rate (up to small order multiplicative factor)

$$\varepsilon_n^2 = K^2 C^{d^*} (\log n)^{\frac{2\beta(1+d^*)}{2\beta+d^*}} n^{-\frac{2\beta}{2\beta+d^*}}$$

- Compositional models with input layer $d \to \infty$ as above Deep–HGP attains compositional $L^2(\mu)$-rate (up to small order multiplicative factor)

Dependence on dimension d, d^*

Above results obtained extending earlier results on GP concentration functions

- Small ball probability
- Approximation term

that were 'up to C(d)' constants

If we want to allow for 'large d' (or even $d \to \infty$) $\to$ need to understand how these results depend on dimension

Hard work = getting dimension–dependent versions of small ball proba + approximation term, e.g. small ball proba, for $\bar{A} = A_1 \cdots A_d$
(in fact, want to apply this to effective dimension d^*)

Summary on Deep GPs

- consider a simple Deep GP prior, close in spirit to [Damianou et al. (13)]; that can be simulated using e.g. `deepgp` R package

- prior on lengthscales allows for simultaneous adaptation to *smoothness* and *structure*

- get (near) optimal rates and adaptation for (fractional) posterior

Remark. Results for $\rho = 1$ (standard posterior) possible via additional prior on noise variance

Remark: Results for classical posteriors (no tempering)

$$Y_i = f(X_i) + \epsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2)$$

If noise variance τ^2 unknown, one may put a prior on it $\tau^2 \sim \pi_\tau$

Set $\Pi = \Pi_f \otimes \pi_\tau$ and $\tilde{\Pi} = \Pi_f \otimes \tilde{\pi}_\tau$, with

$$\Pi_f = \text{HT DNN prior as before}$$

$$\pi_\tau = \text{Gamma}\left(\frac{1-\rho}{2}n + 1, \rho\right), \quad \tilde{\pi}_\tau = \text{Exp}(1)$$

Fact $$\Pi[\cdot \times \mathbb{R} | X, Y] = \tilde{\Pi}_\rho[\cdot \times \mathbb{R} | X, Y]$$

1 Introduction

2 Statistical properties of GPs I

3 Statistical properties of GPs II

4 Scalable GPs: approximations and surrogates
- Variational Bayes
- Vecchia GPs
- Future directions

4. Scalable GPs: approximations and surrogates

Computating GP posteriors

Even in Gaussian regression model where posterior is 'explicit' (conjugacy)

Some issues with posterior sampling

- Computation time of the posterior for training $O(n^3)$
  Becomes impractical for large data sets

Computating GP posteriors

Even in Gaussian regression model where posterior is 'explicit' (conjugacy)

Some issues with posterior sampling

- Computation time of the posterior for training $O(n^3)$
  Becomes impractical for large data sets
- Standard MCMC methods computationally too costly for large data sets

Scalable approaches

- Variational Bayes
- Vecchia approximation
- probabilistic numerics methods, distributed GPs, other sparse/low rank approximation of the covariance/precision matrix (e.g. banding),...

Variational Bayes

Motivation: in nonparametrics/high dimensions $\Pi[\cdot \mid X]$ can be

- a complex distribution with 'dependencies'
- expensive to sample from
  using 'classical' methods such as MCMC

Variational Bayes (VB) approach

- propose a family of 'simple' distributions $\mathcal{Q}$ for θ
- solve the optimisation problem

$$Q^* = \arg\min_{Q \in \mathcal{Q}} \ \mathrm{KL}(Q\,,\, \Pi[\cdot \mid X]), \qquad \mathrm{KL}(q,p) = \int q \log \frac{q}{p}$$

Idea: this is an 'optimisation' problem
for which (sometimes) fast algorithms are available

The ELBO

The KL to be minimised in Q can be rewritten

$$K(Q,\Pi(\cdot|\,X)) = \int \log\left(\frac{q(\theta)}{p_\theta(X)\pi(\theta)/D_X}\right) q(\theta)d\mu$$
$$= \int \log\left(\frac{q(\theta)}{p_\theta(X)\pi(\theta)}\right) q(\theta)d\mu + \log D_X,$$

$$\log D_X = K(Q,\Pi(\cdot|\,X)) + \int \log\left(\frac{p_\theta(X)\pi(\theta)}{q(\theta)}\right) q(\theta)d\mu$$

- $\log D_X$ is called evidence (indep of Q,θ)
- $\int \log\left(p_\theta(X)\pi(\theta)/q(\theta)\right) q(\theta)d\mu$ is the Evidence Lower BOund (ELBO)

$$\boxed{\log D_X \geq \mathrm{ELBO}(Q)}$$

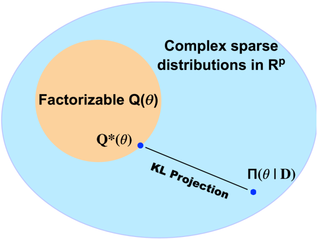
Minimising the $K(Q,\Pi(\cdot|\,X))$ in Q is equivalent to maximising the ELBO

Variational Bayes, trade-offs

Variational Bayes (VB) approach

- propose a family of 'simple' distributions $\mathcal{Q}$ for θ
- solve the optimisation problem

$$Q^* = \underset{Q \in \mathcal{Q}}{\arg\min}\ \mathrm{KL}(Q\,,\,\Pi[\cdot\,|\,X]), \qquad \mathrm{KL}(q,p) = \int q \log \frac{q}{p}$$



[Credit: Kolyan Ray]

Example Common choice for $\mathcal{Q}$

mean-field (factorisable) distributions

$$Q(\theta) = Q_1(\theta_1) \otimes \cdots \otimes Q_p(\theta_p)$$

Trade-offs Simple class $\mathcal{Q}$ $\leftrightarrow$ faster computations

Much faster than standard MCMC $\leftrightarrow$ possibly less 'informative' Q^* than $\Pi[\cdot\,|\,X]$

Variational Bayes: theoretical challenges

What we (start to) understand:

- For well-chosen $\mathcal{Q}$ (e.g. mean-field), can show under some conditions that

  VB-posterior Q^* converges at same rate as $\Pi[\cdot \mid X]$

  This is a global rate result
  - [Alquier, Ridgway 20], [Bhattacharya et al. 20], [Zhang, Gao 20] Nonparametric models
  - [Ray, Szabó 22] high-dimensional models, logistic regression

However, one may be interested in other aspects of the posterior

- posterior for functionals? [Wang, Blei 20]
- uncertainty quantification for functionals? [C., L'Huillier, Ray, Travis 25+]
- model selection?

Variational Bayes & (sparse) GPs

Variational GPs

Variational class: using inducing variables method [Titsias (09)]

Let f follow a GP prior

- Take $u_1, ..., u_m$ linear functionals of f [e.g. $u_i = f(z_i)$ or $u_i = \int a_i \cdot f$ ...]
- Then $f|(u_1, ..., u_m)$ follows again a GP with updated mean and covariance

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u},$$
$$(x, z) \mapsto k(x, z) - K_{x\boldsymbol{u}} K_{\boldsymbol{u}\boldsymbol{u}}^{-1} K_{\boldsymbol{u}z}.$$

where

$$K_{x\boldsymbol{u}} = \mathrm{cov}_\Pi(f(x), \boldsymbol{u}) = K_{\boldsymbol{u}x}^T$$
$$K_{\boldsymbol{u}\boldsymbol{u}} = [\mathrm{cov}_\Pi(u_i, u_j)]_{1 \le i,j \le m}$$

Idea: Keep $f|(u_1, ..., u_m)$ as for the prior; assume $(u_1, \ldots, u_m) \sim \mathcal{N}(\mu, \Sigma)$, then optimise in μ, Σ

Variational GPs (cont)

The variational class is the collection of $Q_{\mu,\Sigma} \in \mathcal{Q}$ with

- $(u_1, \ldots, u_m) \sim \mathcal{N}(\mu, \Sigma)$
- $f\,|(u_1, \ldots, u_m)$ same distribution as under GP prior

These are GPs, with mean and covariance given by

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} \mu,$$
$$(x, z) \mapsto k(x, z) - K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} (K_{\boldsymbol{uu}} - \Sigma) K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{u}z},$$

Facts:

- There exists an optimal μ', Σ' [Titsias 09]
- $Q^* = Q_{\mu',\Sigma'}$ is a particular rank-m approximation of $\Pi(\cdot|\boldsymbol{x}, \boldsymbol{y})$

VB GPs: examples of inducing variables

Inducing point methods

- $f(z_1), ..., f(z_m)$ with $z_i \in \{x_1, ..., x_n\}$
  Computational complexity $O(m^2 n)$ after selecting the points z_i

VB GPs: examples of inducing variables

Inducing point methods

- $f(z_1), ..., f(z_m)$ with $z_i \in \{x_1, ..., x_n\}$
  Computational complexity $O(m^2 n)$ after selecting the points z_i

Population spectral features method

- $u_j = \int f \psi_j dG_x$, for ψ_j eigenfunctions of the covariance kernel k
  Computational complexity: $O(m^2 n)$

VB GPs: examples of inducing variables

Inducing point methods

- $f(z_1), ..., f(z_m)$ with $z_i \in \{x_1, ..., x_n\}$
  Computational complexity $O(m^2 n)$ after selecting the points z_i

Population spectral features method

- $u_j = \int f \psi_j dG_x$, for ψ_j eigenfunctions of the covariance kernel k
  Computational complexity: $O(m^2 n)$

Sample spectral features method

- $u_j = [f(x_1), ..., f(x_n)]\hat{u}_j$, where $\hat{u}_j$ is the jth eigenvector of $K_{\boldsymbol{ff}}$.

VB posterior contraction

[Nieman, Szabó, van Zanten 22]

- $Y_i = f_0(X_i) + \varepsilon_i$
- $h(p_f, p_{f_0})$ Hellinger distance between densities in model
- $R_{\boldsymbol{ff}}$ be the covariance matrix of $\boldsymbol{f}|\boldsymbol{u}$

Denote by Q^* the variational posterior

Theorem For $f_0 : \mathcal{X} \mapsto \mathbb{R}$ assume that

$$\begin{aligned} &(CondGP) \quad \varphi_{\boldsymbol{f_0}}(\varepsilon_n) \leq n\varepsilon_n^2 \\ &(CondVB) \quad E_X tr(R_{\boldsymbol{ff}}) \leq Cn\varepsilon_n^2,\ E_X\|R_{\boldsymbol{ff}}\| \leq C. \end{aligned}$$

Then

$$Q^*(h(p_f, p_{f_0}) \leq M\varepsilon_n) \xrightarrow{P_{f_0}} 1$$

Idea of (Cond VB): how well the prior distribution of $\boldsymbol{f}|\boldsymbol{u}$ approximates the prior f

VB inducing points, minimax contraction rates

[Nieman, Szabó, van Zanten 22] Gaussian regression model

- For $f_0 \in C^\beta([0,1]^d)$, β-Matérn covariance kernel, and $m \geq n^{\frac{d}{d+2\beta}}$ the contraction rate is $n^{-\beta/(d+2\beta)}$ for the population spectral features method.

- For $f_0 \in C^\beta([0,1])$, squared exponential kernel (with rescaling parameter $b = n^{-1/(1+2\beta)}$), and $m \geq n^{\frac{1}{1+2\beta}}$ the contraction rate is $n^{-\beta/(1+2\beta)}(\log n)^{5/4}$ both for the sample and population spectral features methods.

- For $f_0 \in S^\beta([0,1]^d)$, β-regular sequence prior $\Pi = \sum_{k\geq 1} k^{-1/2-\beta}\zeta_k\varphi_k$, $\zeta_k \overset{iid}{\sim} \mathcal{N}(0,1)$ and $m \geq n^{\frac{d}{d+2\beta}}$, the posterior mean concentrates with rate $n^{-\beta/(d+2\beta)}$ for the DPP-inducing points method.

VB inducing points, minimax contraction rates

[Nieman, Szabó, van Zanten 22] Gaussian regression model

- For $f_0 \in C^\beta([0,1]^d)$, β-Matérn covariance kernel, and $m \geq n^{\frac{d}{d+2\beta}}$ the contraction rate is $n^{-\beta/(d+2\beta)}$ for the population spectral features method.

- For $f_0 \in C^\beta([0,1])$, squared exponential kernel (with rescaling parameter $b = n^{-1/(1+2\beta)}$), and $m \geq n^{\frac{1}{1+2\beta}}$ the contraction rate is $n^{-\beta/(1+2\beta)}(\log n)^{5/4}$ both for the sample and population spectral features methods.

- For $f_0 \in S^\beta([0,1]^d)$, β-regular sequence prior $\Pi = \sum_{k\geq 1} k^{-1/2-\beta}\zeta_k\varphi_k$, $\zeta_k \overset{iid}{\sim} \mathcal{N}(0,1)$ and $m \geq n^{\frac{d}{d+2\beta}}$, the posterior mean concentrates with rate $n^{-\beta/(d+2\beta)}$ for the DPP-inducing points method.

Inducing variables GPs for Linear Inverse Problems [Randrianarisoa, Szabó 23]

Thibault Randrianarisoa's talk this afternoon!

Vecchia GPs

Vecchia Approximations of GPs

Consider a *mother* Gaussian process Z on $\mathcal{X}_n$ with joint density

$$p(Z_{\mathcal{X}_n}) = p(Z_{X_1}) \prod_{i=2}^{n} p(Z_{X_i} | Z_{X_j, j<i}).$$

Vecchia Approximations of GPs

Consider a *mother* Gaussian process Z on $\mathcal{X}_n$ with joint density

$$p(Z_{\mathcal{X}_n}) = p(Z_{X_1}) \prod_{i=2}^{n} p(Z_{X_i} | Z_{X_j, j<i}).$$

The Vecchia approximations of Gaussian Processes (Vecchia GPs) replace each conditional set $\{X_j, j < i\}$ with a much smaller parent set $\mathrm{pa}(X_i)$

$$p(\hat{Z}_{\mathcal{X}_n}) = p(\hat{Z}_{X_1}) \prod_{i=2}^{n} p(\hat{Z}_{X_i} | \hat{Z}_{\mathrm{pa}(X_i)}),$$

Vecchia Approximations of GPs

Consider a *mother* Gaussian process Z on $\mathcal{X}_n$ with joint density

$$p(Z_{\mathcal{X}_n}) = p(Z_{X_1}) \prod_{i=2}^{n} p(Z_{X_i} | Z_{X_j, j<i}).$$

The Vecchia approximations of Gaussian Processes (Vecchia GPs) replace each conditional set $\{X_j, j < i\}$ with a much smaller parent set $\mathrm{pa}(X_i)$

$$p(\hat{Z}_{\mathcal{X}_n}) = p(\hat{Z}_{X_1}) \prod_{i=2}^{n} p(\hat{Z}_{X_i} | \hat{Z}_{\mathrm{pa}(X_i)}),$$

such that the **conditional distributions given parent sets** remain unchanged (with $K_{\cdot,\cdot}$ denoting the covariance matrices):

$$\begin{aligned} &[\hat{Z}_{X_i} \mid \hat{Z}_{\mathrm{pa}(X_i)} = z] \overset{d}{=} [Z_{X_i} \mid Z_{\mathrm{pa}(X_i)} = z] \\ \sim &N\Big(K^T_{\mathrm{pa}(X_i),X_i} K^{-1}_{\mathrm{pa}(X_i),\mathrm{pa}(X_i)} z,\ K_{X_i,X_i} - K^T_{\mathrm{pa}(X_i),X_i} K^{-1}_{\mathrm{pa}(X_i),\mathrm{pa}(X_i)} K_{\mathrm{pa}(X_i),X_i}\Big). \end{aligned}$$

Evaluating $p(\hat{Z}_{\mathcal{X}_n})$ takes $O(nm^3)$ time if $\mathrm{card}(\mathrm{pa}(X_i)) \leq m, \forall i$.

Vecchia Approximations of GPs

[Szabó, Zhu 25+]

Fact. It is possible to choose parent sets of cardinality of order $\binom{\alpha+d}{\alpha}$ to achieve minimax rates

Theorem Let $f_0 \in C^\beta$ and Z a Matérn GP with smoothness parameter β. The Vecchia GP $\hat{Z}$ with above choice of parent set achieves a posterior contraction rate of

$$\varepsilon_n \asymp n^{-\frac{2\beta}{2\beta+d}}$$

Adaptation to smoothness is also possible by suitable rescaling

Botond Szabó's talk on Vecchia GPs this afternoon!

Conclusion and some work in progress/open problems

Gaussian processes are a flexible tool for nonparametrics

In these lectures, we have presented theory including the following results:

- GPs achieve optimal minimax rates in L^2
  - Generic tools via the concentration function quantify rates
  - RKHS $\mathbb{H}$ quantifies both approximation and complexity (small ball)
- By choosing hyperparameters via empirical or hierarchical Bayes
  - adaptation to global smoothness (e.g. small lengthscale)
  - variable selection (e.g. large lengthscale)
  - adaptation to structure (compositional, geometric) e.g. DeepGPs

Current/future directions

- Theory and implementation for scalable versions
- Deep Vecchia GPs [with Thibault R., Botond S., Yichen Zhu]
- Local rates via flexible GPs (e.g. deep GPs) [with Thibault R., Botond S., Yichen Zhu]
- theory/implementation for heavy tailed processes

Thank you!