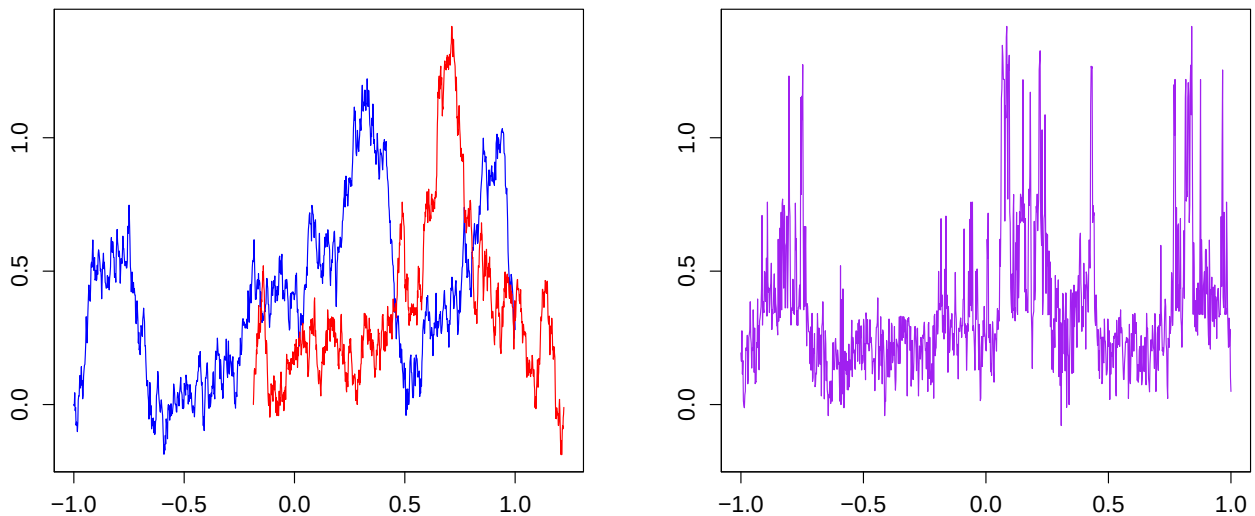
Posterior contraction for deep Gaussian process priors

Gianluca Finocchio
(joint with J. Schmidt-Hieber)

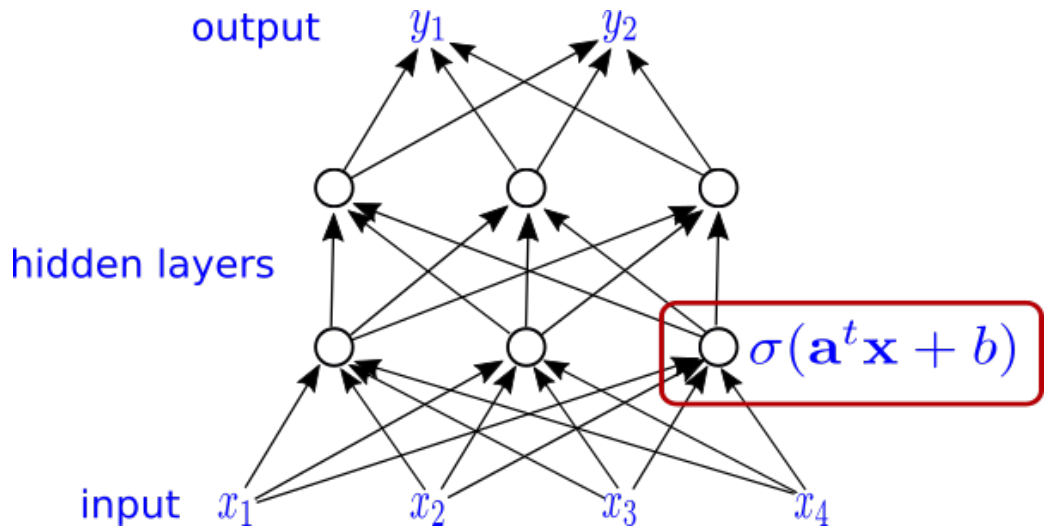
ISBA 2022

deep Gaussian processes

- Gaussian process (GP) priors are popular in machine learning
- it seems natural to compose GPs
- Example: iterated Brownian motion $W_1(W_2(t))$ with W_1, W_2 Brownian motions
  - not a GP
  - smoothness is almost $1/4$



neural networks



$$f(\mathrm{x}) = W_L \sigma W_{L-1} \ldots \sigma W_1 \sigma W_0 \mathrm{x}$$

- $\sigma(x)$ is the activation function, e.g. ReLU $\sigma(x) = \max(x, 0)$
- L is the network depth or number of hidden layers
- $L = 1$ shallow, $L > 1$ deep
- matrices W_i are the free parameters

random initialization

initialization is crucial for success of deep learning

- initialization schemes : draw network weights independently

A wide, randomly initialized shallow neural network is approximately a Gaussian process with covariance

$$K(\mathrm{x}, \mathrm{x}') = \text{const.} \times E\left[\sigma(w^\top \mathrm{x})\sigma(w^\top \mathrm{x}')\right]$$

- expectation with respect to distribution of w
- constant is determined by the variance of the weights in output layer

wide shallow networks vs. GP priors

wide shallow networks	**GP priors**
decrease $-$(log)-likelihood using (S)GD initialized with GP	posterior $\propto e^{\text{log-likelih.}} \times$ GP
overparametrized	'underparametrized'
behaves in NTK regime like kernel regression	for Gaussian model, posterior is also Gaussian and centered around Tikhonov minimizer
bad extrapolator	good extrapolator
no UQ	credible sets

deep neural networks vs. deep GPs

- is a randomly initialized wide deep neural network a deep Gaussian process?
- depends on the way we form the wide limit
- typically, it will again be a GP

stabilization during deep learning

gradient descent for deep networks can end up in vanishing or exploding gradient regimes

$\rightarrow$ can be circumvented using batch normalization, this means essentially that

- if X denotes the distribution of the design
- and $H_i(\mathrm{X})$ denotes the distribution of the outputs from the i-th hidden layer
- we force $H_i(\mathrm{X})$ to have mean zero and variance one

does something similar happen for deep Gaussian process priors?

statistical model

Nonparametric regression model: Observe n independent pairs $(\mathsf{X}_i, Y_i) \in [-1,1]^d \times \mathbb{R}$ with

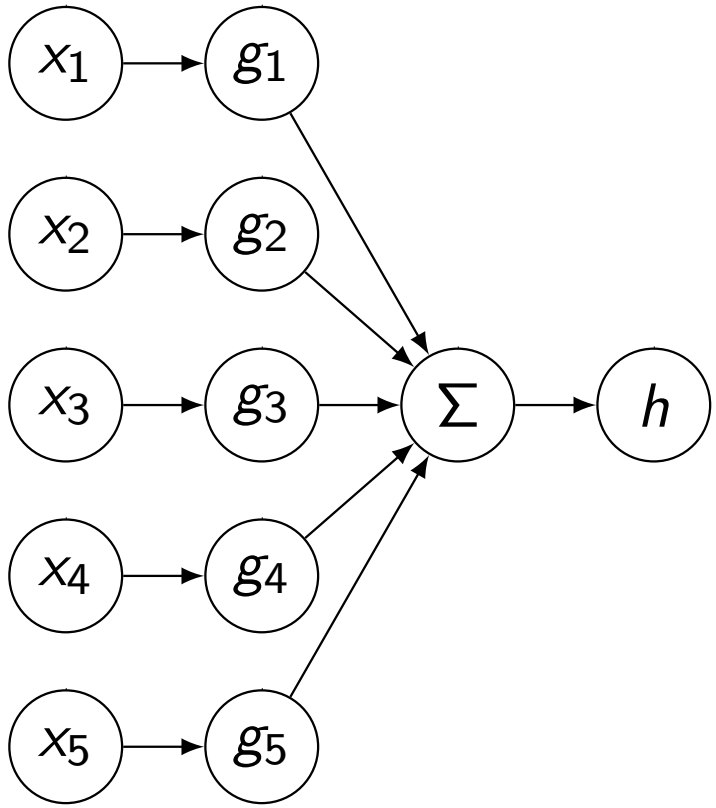
$$Y_i = f(\mathsf{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

- $\varepsilon_i \sim \mathcal{N}(0,1)$
- unknown regression function f

we need more structure than smoothness on the regression function to see any interesting effects of deep GP priors

composition structure

Generalized additive models are included



$$f(x_1, \ldots, x_d) = h\left(\sum_{i=1}^{d} g_i(x_i)\right)$$

general function composition

- We assume that

$$f = g_q \circ \ldots \circ g_0$$

with

  - $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$.
  - each of the d_{i+1} components of g_i is β_i-smooth and depends only on t_i variables
  - t_i can be much smaller than d_i
  - effective smoothness $\beta_i^* := \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1)$

- the minimax estimation rate is (up to log n-factors)

$$\max_{i=0,\ldots,q} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}$$

- the rate depends on the pairs (t_i, β_i^*), $\quad i = 0, \ldots, q$.
- Schmidt-Hieber (2017) $\rightsquigarrow$ DNNs can achieve this rate

heuristic

- composition assumption means that the target function f has a modular structure, that is, it is built successively from simpler functions
- it has been argued that such a structure is present in many examples where deep learning is state of the art

construction of a deep Gaussian process prior

hierarchical prior construction:

1. put a prior on composition graphs
   - to avoid overfitting, large graphs should get little prior weight
   - sample size dependent
2. put GP prior on each edge
3. regularization of sample paths

choice of Gaussian process

- pick a family $\widetilde{G}^{(\beta,r)}$ of centered GPs on $[-1,1]^r$
- concentration function (vdVaart & van Zanten '08)

$$\varphi_f^{(\beta,r)}(u) := \underbrace{\inf_{g:\|g-f\|_\infty \le u} \|g\|^2_{\mathbb{H}(\beta,r)}}_{\text{RKHS approx.}} - \log \underbrace{\mathbb{P}\big(\big\|\widetilde{G}^{(\beta,r)}\big\|_\infty \le u\big)}_{\text{small ball prob.}},$$

- find $\varepsilon_n(\alpha,\beta,r)$ such that for all Hölder-β functions on $[-1,1]^r$ and any $0 < \alpha \le 1$,

$$\varphi_f^{(\beta,r)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) \le n\varepsilon_n(\alpha,\beta,r)^2.$$

- "typical" rate

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

prior on composition graphs

- larger graphs can explain the data better
- to avoid overfitting, one needs to downweight large models
- to do this in a Bayesian framework is quite standard nowadays
- start with a distribution γ on composition graphs
- define prior π on graph η as

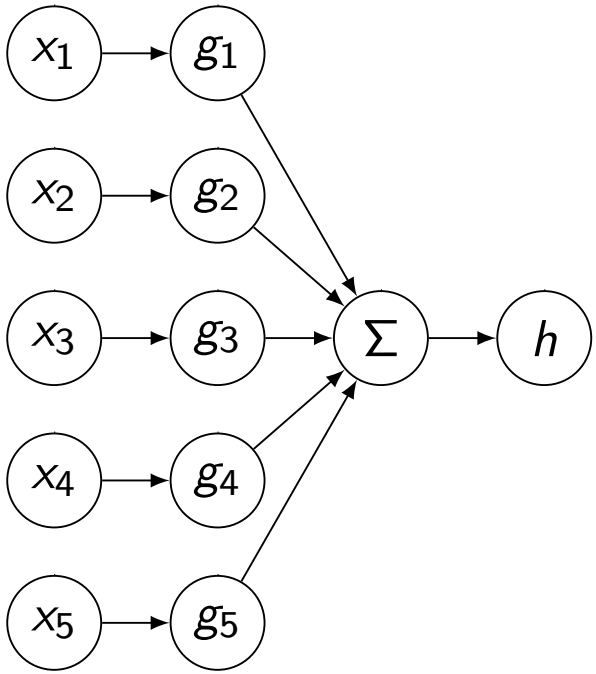
$$\pi(\eta) \propto e^{-n\varepsilon_n(\eta)^2}\gamma(\eta)$$

- with

$$\varepsilon_n(\eta) = \max_{i=0,\dots,q} \underbrace{\varepsilon_n(\alpha_i, \beta_i, t_i)}_{\text{defined through concentration function}}$$

- and $\alpha_i := \prod_{\ell=i+1}^{q}(\beta_\ell \wedge 1)$

GP prior on the nodes



- given a composition graph
- draw each node in i-th layer from a GP conditioned to have
  - sample paths in $[-1, 1]$
  - sample paths that lie in an $\varepsilon_n(\alpha_i, \beta_i, t_i)$-sup norm ball around a Hölder-β_i function

on the conditioning

is conditioning necessary:

- has a similar flavor as batch normalization in deep learning
- is mainly needed for the theory to work
- compositions of functions can have quite unexpected behavior
  - Kolmogorov-Arnold representation theorem shows that continuous functions $[0,1]^d \to \mathbb{R}$ can be written as compositions of univariate functions
- for us it is still unclear whether this is really necessary
- van der Vaart and van Zanten (2009) $\rightsquigarrow$ inverse Gamma bandwidth leads to adaptation

main result

Theorem:

- consider nonparametric regression with composition structure assumption on regression function

$$f = g_q \circ \ldots \circ \underbrace{g_i}_{\beta_i\text{-smooth depending on } t_i\text{variables}} \circ \ldots \circ g_0$$

- under weak regularity assumptions and suitable choices of the GPs, posterior contracts (up to log n-factors) with the minimax estimation rate

$$\max_{i=0,\ldots,q} n^{-\frac{\alpha_i \beta_i}{2\alpha_i \beta_i + t_i}}, \quad \text{with } \alpha_i := \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1).$$

- Giordano, Ray, Schmidt-Hieber (2022) $\rightsquigarrow$ simple GP priors cannot achieve this rate!

summary

- comparison of GP priors and neural nets
- constructed a hierarchical deep GP prior
- derived nearly optimal posterior contraction rates
- proof extends the Bayesian nonparametrics theory for GPs developed by van der Vaart, van Zanten (2008)
- many open questions:
  - is posterior computable ? If so, how does it compare with deep learning?
  - are path constraints on GPs necessary ?

Thank you for your attention!