Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression

Bernhard Stankewitz

CIRM: New challenges in high dimensional statistics
Marseille, December 2024

Department of Mathematics
University of Potsdam

Joint work with

Botond Szabo, Bocconi Milano

Contents

Motivation: Scalability of Gaussian process regression

Gaussian process (GP) regression

Consider i.i.d. observations from the model

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where

- $X_1, \dots, X_n \sim G$ i.i.d. on $\mathbb{R}^d$ and $\varepsilon \sim N(0, \sigma^2 I_n)$;
- $F \sim \mathrm{GP}(0, k)$ with p.s.d. kernel $k : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a GP-prior on $L^2(G)$, i.e.

$$\mathbb{E}F(x) = 0, \qquad \mathrm{Cov}(F(x), F(x')) = k(x, x'), \qquad x, x' \in \mathbb{R}^d. \tag{2}$$

Gaussian process (GP) regression

Consider i.i.d. observations from the model

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where

- $X_1, \dots, X_n \sim G$ i.i.d. on $\mathbb{R}^d$ and $\varepsilon \sim N(0, \sigma^2 I_n)$;
- $F \sim \mathrm{GP}(0, k)$ with p.s.d. kernel $k : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a GP-prior on $L^2(G)$, i.e.

$$\mathbb{E}F(x) = 0, \qquad \mathrm{Cov}(F(x), F(x')) = k(x, x'), \qquad x, x' \in \mathbb{R}^d. \tag{2}$$

Setting $K := (k(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and $k(X, x) := (k(X_i, x))_{i=1}^n \in \mathbb{R}^n$, the posterior $\Pi(\cdot|X, Y)$ is given by the GP with mean and covariance function

$$\begin{aligned} x &\mapsto k(X, x)^\top (K + \sigma^2 I_n)^{-1} Y \\ (x, x') &\mapsto k(x, x') - k(X, x)^\top (K + \sigma^2 I_n)^{-1} k(X, x'). \end{aligned} \tag{3}$$

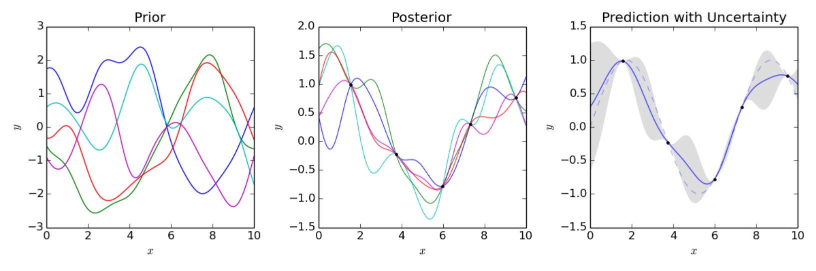


Figure 1: Gaussian Process Regression (prediction) with a squared exponential kernel. Left plot are draws from the prior function distribution. Middle are draws from the posterior. Right is mean prediction with one standard deviation shaded.

Gaussian process (GP) regression

Consider i.i.d. observations from the model

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4}$$

where

- $X_1, \dots, X_n \sim G$ i.i.d. on $\mathbb{R}^d$ and $\varepsilon \sim N(0, \sigma^2 I_n)$;
- $F \sim \mathrm{GP}(0, k)$ with p.s.d. kernel $k : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a GP-prior on $L^2(G)$, i.e.

$$\mathbb{E}F(x) = 0, \qquad \mathrm{Cov}(F(x), F(x')) = k(x, x'), \qquad x, x' \in \mathbb{R}^d. \tag{5}$$

Setting $K := (k(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and $k(X, x) := (k(X_i, x))_{i=1}^n \in \mathbb{R}^n$, the posterior $\Pi(\cdot|X, Y)$ is given by the GP with mean and covariance function

$$\begin{aligned} x &\mapsto k(X, x)^\top (K + \sigma^2 I_n)^{-1} Y \\ (x, x') &\mapsto k(x, x') - k(X, x)^\top (K + \sigma^2 I_n)^{-1} k(X, x'). \end{aligned} \tag{6}$$

Motivating problem

The computation of $(K + \sigma^2 I)^{-1}$ has a computational complexity of $O(n^3)$, which becomes infeasable for large n.

Algorithms from Probabilistic Numerics

Algorithms from probabilistic numerics (Wenger et al. [Wen+22][1])

Idea: Focussing on the posterior mean $k(X,x)^\top(K+\sigma^2 I_n)^{-1}Y$, iteratively solve $(K+\sigma^2 I_n)W = Y$ for the representer weights W.

- Consider a Bayesian updating scheme with initial believes $W^* = (K+\sigma^2 I_n)^{-1}Y \sim N(0,(K+\sigma^2 I_n)^{-1}) =: N(w_0,\Gamma_0)$.

[1] J. Wenger et al. "Posterior and computational uncertainty in Gaussian processes.". In: *Advances in Neural Information Processing Systems* (2022).

Algorithms from probabilistic numerics (Wenger et al. [Wen+22][1])

Idea: Focussing on the posterior mean $k(X,x)^\top(K+\sigma^2 I_n)^{-1}Y$, iteratively solve $(K+\sigma^2 I_n)W = Y$ for the representer weights W.

- Consider a Bayesian updating scheme with initial believes $W^* = (K+\sigma^2 I_n)^{-1}Y \sim N(0,(K+\sigma^2 I_n)^{-1}) =: N(w_0,\Gamma_0)$.
- Consecutively update by conditioning on the information projection

$$\alpha_j := s_j^\top(Y-(K+\sigma^2 I_n)w_{j-1}), \qquad j \le m, \tag{7}$$

where $s_j, j \le m$ are search directions chosen by the user. Inductively, $W^*|\alpha_m \sim N(w_m,\Gamma_m)$ with

$$\begin{aligned} w_m &= w_{m-1} + \eta_m^{-1} d_m d_m^\top Y = C_m Y, \\ \Gamma_m &= \Gamma_{m-1} - \eta_m^{-1} d_m d_m^\top = (K+\sigma^2 I)^{-1} - C_m, \end{aligned} \tag{8}$$

where $d_m = (I - C_{m-1}(K+\sigma^2 I))s_m$, $\eta_m = s_m^\top(K+\sigma^2 I)d_m$ and $C_m = \sum_{j=1}^m \eta_j^{-1} d_j d_j^\top$.

[1] J. Wenger et al. "Posterior and computational uncertainty in Gaussian processes.". In: *Advances in Neural Information Processing Systems* (2022).

- After m steps, believes are given by $N(w_m, \Gamma_m) = N(C_m Y, (K+\sigma^2)^{-1} - C_m)$. This yields the approximate Gaussian posterior $\Psi_m := \mathbb{P}^{F|W=w} N(w_m, \Gamma_m)(dw)$ with mean and covariance functions

$$x \mapsto k(X,x)^\top C_m Y \qquad (x,x') \mapsto k(x,x') - k(X,x)^\top C_m k(X,x'), \tag{9}$$

where C_m is a rank m matrix approximating $(K+\sigma^2 I_n)^{-1}$.

- After m steps, believes are given by $N(w_m, \Gamma_m) = N(C_m Y, (K + \sigma^2)^{-1} - C_m)$. This yields the approximate Gaussian posterior $\Psi_m := \mathbb{P}^{F|W=w} N(w_m, \Gamma_m)(dw)$ with mean and covariance functions

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{9}$$

where C_m is a rank m matrix approximating $(K + \sigma^2 I_n)^{-1}$.

- The approximate covariance can be split into a mathematical and a computational uncertainty

$$(x, x') \mapsto \underbrace{k(x, x') - k(X, x)^\top (K + \sigma^2 I)^{-1} k(X, x')}_{\text{Mathematical uncertainty}} + \underbrace{k(X, x)^\top \Gamma_m k(X, x')}_{\text{Computational uncertainty}}. \tag{10}$$

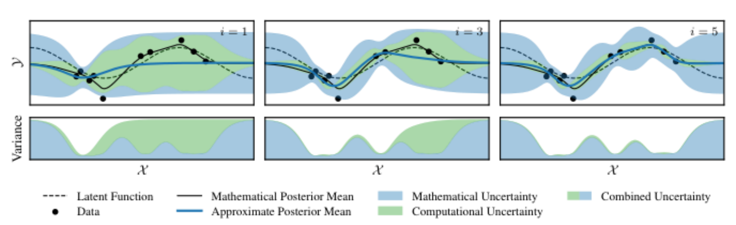


Figure 2: Mathematical and computational uncertainty. Source: [Wen+22]

Algorithm 1 GP approximation scheme

procedure $\text{ITERGP}(k, X, Y)$
    $C_0 \leftarrow 0 \in \mathbb{R}^{n \times n}$
    for $j = 1, 2, \ldots, m$ **do**
        $s_j \leftarrow \text{POLICY}()$
        $d_j \leftarrow (I - C_{j-1} K_\sigma) s_j$
        $\eta_j \leftarrow s_j^\top K_\sigma d_j$
        $C_j \leftarrow C_{j-1} + \eta_j^{-1} d_j d_j^\top$
    end for
    $\mu_m(\cdot) \leftarrow k(X, \cdot)^\top C_m Y$
    $k_m(\cdot, \cdot) \leftarrow k(\cdot, \cdot) - k(X, \cdot)^\top C_m k(X, \cdot)$
end procedure
return $\text{GP}(\mu_m, k_m)$

Policy examples

(a) $s_j := e_j$, $j \leq m \rightsquigarrow$ partial Cholesky decomposition of $K + \sigma^2 I$.

(b) $s_j := \widehat{u}_j$, $j \leq m \rightsquigarrow$ SVD of $K + \sigma^2 I$.

(c) $s_j := \tilde{u}_j$, $j \leq m \rightsquigarrow$ Lanczos approximation.

(b) $s_j := d_j^{\text{CG}}$, $j \leq m \rightsquigarrow$ *CG* applied to $K_\sigma v = Y$.

The empirical eigenvector posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top = n \sum_{j=1}^{n} \widehat{\lambda}_j \widehat{u}_j \widehat{u}_j^\top \tag{11}$$

and choose the search directions $s_j := \widehat{u}_j,\ j \leq m$.

[2] D. Nieman, B.Szabo and H. van Zanten. "Contraction rates for sparse variational approximations in Gaussian process regression". In: *Journal of Machine Learning Research* 23 (2022).

The empirical eigenvector posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top = n \sum_{j=1}^{n} \widehat{\lambda}_j \widehat{u}_j \widehat{u}_j^\top \tag{11}$$

and choose the search directions $s_j := \widehat{u}_j,\ j \leq m$. Then, the approximate posterior $\Psi_m = \Psi_m^{\mathrm{EV}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{12}$$

where $(K + \sigma^2 I_n)^{-1}$ is approximated by

$$C_m = C_m^{\mathrm{EV}} = \sum_{j=1}^{m} (\widehat{\mu}_j + \sigma^2)^{-1} \widehat{u}_j \widehat{u}_j^\top. \tag{13}$$

---

[2] D. Nieman, B.Szabo and H. van Zanten. "Contraction rates for sparse variational approximations in Gaussian process regression". In: *Journal of Machine Learning Research* 23 (2022).

The empirical eigenvector posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top = n \sum_{j=1}^{n} \widehat{\lambda}_j \widehat{u}_j \widehat{u}_j^\top \tag{11}$$

and choose the search directions $s_j := \widehat{u}_j,\ j \leq m$. Then, the approximate posterior $\Psi_m = \Psi_m^{\mathrm{EV}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{12}$$

where $(K + \sigma^2 I_n)^{-1}$ is approximated by

$$C_m = C_m^{\mathrm{EV}} = \sum_{j=1}^{m} (\widehat{\mu}_j + \sigma^2)^{-1} \widehat{u}_j \widehat{u}_j^\top. \tag{13}$$

The Ψ_m^{EV} is equivalent to the Variational Bayes posterior based on spectral inducing variables [NSZ22][2].

[2] D. Nieman, B.Szabo and H. van Zanten. "Contraction rates for sparse variational approximations in Gaussian process regression". In: *Journal of Machine Learning Research* 23 (2022).

The Lanczos posterior

For $v_0 \in \mathbb{R}$ with $\|v_0\| = 1$, consider the Krylov spaces

$$\mathcal{K}_{\tilde{m}} := \text{span}\{v_0, Kv_0, \ldots, K^{\tilde{m}-1}v_0\}, \qquad \tilde{m} = 1, 2, \ldots, n. \tag{14}$$

The Lanczos approximate eigenpairs

$$(\tilde{\mu}_j, \tilde{u}_j), \qquad \tilde{\lambda}_j := n^{-1}\tilde{\mu}_j, \qquad j = 1, \ldots, \tilde{m} \tag{15}$$

are given by the following algorithm:

The Lanczos posterior

For $v_0 \in \mathbb{R}$ with $\|v_0\| = 1$, consider the Krylov spaces

$$\mathcal{K}_{\tilde{m}} := \text{span}\{v_0, Kv_0, \ldots, K^{\tilde{m}-1}v_0\}, \qquad \tilde{m} = 1, 2, \ldots, n. \tag{14}$$

The Lanczos approximate eigenpairs

$$(\tilde{\mu}_j, \tilde{u}_j), \qquad \tilde{\lambda}_j := n^{-1}\tilde{\mu}_j, \qquad j = 1, \ldots, \tilde{m} \tag{15}$$

are given by the following algorithm:

Algorithm 3 Lanczos algorithm

1: **procedure** ITERLANCZOS($K, v_0, \tilde{m}$)
2: Initialize v_0 with $\|v_0\| = 1$.
3: Compute ONB $v_1, \ldots, v_{\tilde{m}}$ of $\mathcal{K}_{\tilde{m}}$.
4: $V \leftarrow (v_1, \ldots, v_{\tilde{m}})$.
5: $A \leftarrow n^{-1}K$.
6: Compute eigenpairs $(\tilde{\lambda}_j, \tilde{u}_j)_{j \le \tilde{m}}$ of $V^\top AV$.
7: $\tilde{u}_j \leftarrow V\tilde{u}_j$, $j \le \tilde{m}$.
8: **end procedure**
9: **return** $(\tilde{\lambda}_j, \tilde{u}_j)_{j \le \tilde{m}}$.

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{16}$$

and choose the search directions $s_j := \tilde{u}_j, j \leq m$, where $(\tilde{\mu}_j, \tilde{u}_j)_{j \leq m}$ is the Lanczos approximate eigensystem up to order m.

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{16}$$

and choose the search directions $s_j := \tilde{u}_j, j \leq m$, where $(\tilde{\mu}_j, \tilde{u}_j)_{j \leq m}$ is the Lanczos approximate eigensystem up to order m. Then, the approximate posterior $\Psi_m = \Psi_m^{\mathsf{L}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{17}$$

with

$$C_m = C_m^{\mathsf{L}} = \sum_{j=1}^{m} (\tilde{\mu}_j + \sigma^2)^{-1} \tilde{u}_j \tilde{u}_j^\top . \tag{18}$$

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{16}$$

and choose the search directions $s_j := \tilde{u}_j, j \leq m$, where $(\tilde{\mu}_j, \tilde{u}_j)_{j \leq m}$ is the Lanczos approximate eigensystem up to order m. Then, the approximate posterior $\Psi_m = \Psi_m^{\mathsf{L}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{17}$$

with

$$C_m = C_m^{\mathsf{L}} = \sum_{j=1}^{m} (\tilde{\mu}_j + \sigma^2)^{-1} \tilde{u}_j \tilde{u}_j^\top. \tag{18}$$

Randomness of the kernel matrix

Since $K = (k(X_i, X_j))_{i,j \leq n}$ is a random matrix, the spectral decomposition of K cannot be computed in advance.

The CG posterior

Conjugate gradient descent. Iteratively solve $(K + \sigma^2 I_n)w = Y$ by setting $w_0 = 0$ and for $j \geq 1$,

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{19}$$

where $\varrho(w) := (w^\top (K + \sigma^2 I_n) w)/2 - Y^\top w$, and the $(d_j^{\mathrm{CG}})_{j \geq 1}$ are conjugate search directions satisfying $(d^{\mathrm{CG}})_j^\top (K + \sigma^2 I_n) d_k^{\mathrm{CG}} = 0$, $j \neq k$.

The CG posterior

Conjugate gradient descent. Iteratively solve $(K + \sigma^2 I_n)w = Y$ by setting $w_0 = 0$ and for $j \geq 1$,

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{19}$$

where $\varrho(w) := (w^\top (K + \sigma^2 I_n) w)/2 - Y^\top w$, and the $(d_j^{\mathrm{CG}})_{j \geq 1}$ are conjugate search directions satisfying $(d^{\mathrm{CG}})_j^\top (K + \sigma^2 I_n) d_k^{\mathrm{CG}} = 0$, $j \neq k$.

For the policies $s_j := d_j^{\mathrm{CG}}, j \leq m$, Bayesian updating is equivalent to the CG-iteration and we obtain the approximate posterior Ψ_m^{CG} given by

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{20}$$

where $C_m = C_m^{\mathrm{CG}}$ is given by the implicit approximation of the inverse provided by CG.

The CG posterior

Conjugate gradient descent. Iteratively solve $(K + \sigma^2 I_n)w = Y$ by setting $w_0 = 0$ and for $j \geq 1$,

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{19}$$

where $\varrho(w) := (w^\top (K + \sigma^2 I_n) w)/2 - Y^\top w$, and the $(d_j^{\mathrm{CG}})_{j \geq 1}$ are conjugate search directions satisfying $(d^{\mathrm{CG}})_j^\top (K + \sigma^2 I_n) d_k^{\mathrm{CG}} = 0$, $j \neq k$.

For the policies $s_j := d_j^{\mathrm{CG}}, j \leq m$, Bayesian updating is equivalent to the CG-iteration and we obtain the approximate posterior Ψ_m^{CG} given by

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{20}$$

where $C_m = C_m^{\mathrm{CG}}$ is given by the implicit approximation of the inverse provided by CG.

The CG posterior

Conjugate gradient descent. Iteratively solve $(K+\sigma^2 I_n)w = Y$ by setting $w_0 = 0$ and for $j \geq 1$,

$$\varrho(w_j) = \min_{t\in\mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{19}$$

where $\varrho(w) := (w^\top (K+\sigma^2 I_n)w)/2 - Y^\top w$, and the $(d_j^{\mathrm{CG}})_{j\geq 1}$ are conjugate search directions satisfying $(d^{\mathrm{CG}})_j^\top (K+\sigma^2 I_n) d_k^{\mathrm{CG}} = 0$, $j \neq k$.

For the policies $s_j := d_j^{\mathrm{CG}}$, $j \leq m$, Bayesian updating is equivalent to the CG-iteration and we obtain the approximate posterior Ψ_m^{CG} given by

$$x \mapsto k(X,x)^\top C_m Y \qquad (x,x') \mapsto k(x,x') - k(X,x)^\top C_m k(X,x'), \tag{20}$$

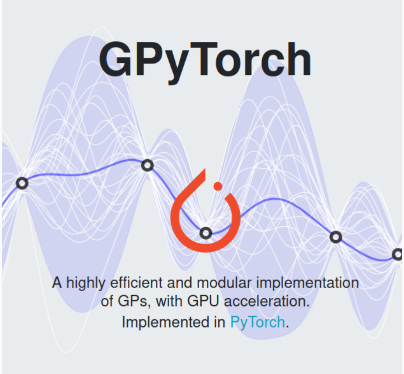
where $C_m = C_m^{\mathrm{CG}}$ is given by the implicit approximation of the inverse provided by CG.

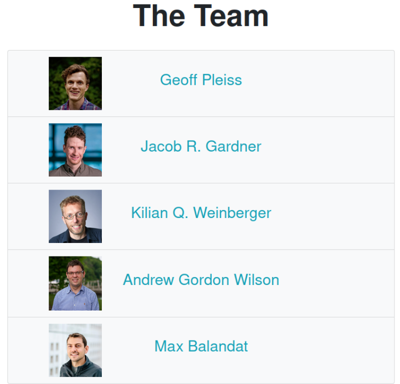
Reduction in computational complexity

The approximate inversions $C_m^{\mathrm{L}}, C_m^{\mathrm{CG}}$ have a computation cost of $O(mn^2)$, which is feasible when $m \ll n$.

GPU accelerated matrix vector multiplication

CG only relies on matrix vector multiplications, which can be GPU accelerated and makes CG particularly relevant for large scale applications, see Wang, Pleiss, Gardner, Tyree, Weinberger and Wilson [Wan+19].





Main results: Contraction of approximate posteriors

Contraction rates

For $f_0 \in \overline{\mathbb{H}}$ with $\mathbb{H} = \text{ran } T_k^{1/2}$,

$$T_k : L^2(G) \to L^2(G), \qquad f \mapsto \int f(y)k(\cdot, y)\, G(dy) = \sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{L^2(G)} \phi_j, \qquad (21)$$

let $\mathbb{P}_{f_0}$ be the measure corresponding to the data generating process

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad i = 1, \dots n. \qquad (22)$$

Contraction rates

For $f_0 \in \overline{\mathbb{H}}$ with $\mathbb{H} = \operatorname{ran} T_k^{1/2}$,

$$T_k : L^2(G) \to L^2(G), \qquad f \mapsto \int f(y)k(\cdot, y)\, G(dy) = \sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{L^2(G)} \phi_j, \tag{21}$$

let $\mathbb{P}_{f_0}$ be the measure corresponding to the data generating process

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad i = 1, \ldots n. \tag{22}$$

Consider the densities

$$\mathcal{P} := \left\{ p_f(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - f(x))^2}{2\pi\sigma^2} \right), f \in L^2(G) \right\} \tag{23}$$

with respect to $G \otimes \lambda$ and write

$$d_H(f, g) := d_H(p_f, p_g) = \sqrt{\int (\sqrt{p_f} - \sqrt{p_g})^2 \, dG \otimes \lambda}, \qquad f, g \in L^2(G) \tag{24}$$

for the Hellinger distance.

Contraction rates

For $f_0 \in \overline{\mathbb{H}}$ with $\mathbb{H} = \operatorname{ran} T_k^{1/2}$,

$$T_k : L^2(G) \to L^2(G), \qquad f \mapsto \int f(y) k(\cdot, y)\, G(dy) = \sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{L^2(G)} \phi_j, \tag{21}$$

let $\mathbb{P}_{f_0}$ be the measure corresponding to the data generating process

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad i = 1, \dots n. \tag{22}$$

Consider the densities

$$\mathcal{P} := \left\{ p_f(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - f(x))^2}{2\pi\sigma^2} \right), f \in L^2(G) \right\} \tag{23}$$

with respect to $G \otimes \lambda$ and write

$$d_H(f, g) := d_H(p_f, p_g) = \sqrt{\int (\sqrt{p_f} - \sqrt{p_g})^2\, dG \otimes \lambda}, \qquad f, g \in L^2(G) \tag{24}$$

for the Hellinger distance.

Definition 3.1 (Contraction rate)

The posterior contracts with rate $\varepsilon_n \to 0$ around the truth $f_0 \in L^2(G)$ if

$$\Pi\{d_H(\cdot, f_0) \geq M_n \varepsilon_n | X, Y\} = \Pi_n\{d_H(\cdot, f_0) \geq M_n \varepsilon_n | (X_i, Y_i)_{i=1}^n\} \xrightarrow[n \to \infty]{\mathbb{P}_{f_0}^{\otimes n}} 0.$$

For $f_0 \in L^2(G)$, define the concentration function at f_0 as

$$\varphi_{f_0}(\varepsilon) := \inf_{h \in \mathbb{H}: \|h - f_0\|_2 \leq \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}\{\|F\|_2 < \varepsilon\}, \tag{25}$$

where $\mathbb{H} = \operatorname{ran} T_k^{1/2}$ is the RKHS of the Gaussian process F.

(A1) **(CFun):** For a sequence $\varepsilon_n \to 0$, assume the concentration function at f_0 satisfies

$$\varphi_{f_0}(\varepsilon_n) \leq C_\varphi n \varepsilon_n^2 \tag{26}$$

for some $C_\varphi > 0$.

Proposition 3.2 (Standard contraction rate, Ghosal and van der Vaart [Gv17])

Assume that at some $f_0 \in \overline{\mathbb{H}}$, the contraction function inequality Equation (26) holds for a sequence $\varepsilon_n \to 0$ with $n\varepsilon_n^2 \to \infty$. Then, there exists a constant $C_1 > 0$ such that for any constant $C_2 > 0$,

$$\mathbb{E}_{f_0}^n(\Pi\{d_H(\cdot, f_0) \geq M_n \varepsilon_n | X, Y\} \mathbf{1}_{A_n}) \leq C_1 \exp(-C_2 n \varepsilon_n^2), \tag{27}$$

for n sufficiently large and a sequence $(A_n)_{n \in \mathbb{N}}$ with $\mathbb{P}_{f_0}^{\otimes n}(A_n) \to 0$.

(A2) **(SPE):** The population eigenvalues $(\lambda_j)_{j\geq 1}$ of T_k are simple, i.e., $\lambda_1 > \lambda_2 > \cdots > 0$.

(A3) **(EVD):** We assume the following decay behaviour of the population eigenvalues:

(i) There exists a convex function $\lambda : [0, \infty) \to [0, \infty)$ such that $\lambda_j = \lambda(j)$ and $\lim_{j\to\infty} \lambda(j) = 0$.
(ii) There exists a constant $C > 0$ such that, $\lambda(Cj) \leq \lambda(j)/2$ for all $j \in \mathbb{N}$.
(iii) There exists a constant $c > 0$ such that $\lambda_j \geq e^{-cj}$ for all $j \in \mathbb{N}$.

(A4) **(KLMom):** There exists a $p > 4$, such that the Karhunen-Loève coefficients $\eta_j := \langle k(\cdot, X_1), \phi_j \rangle_{\mathbb{H}} = \phi_j(X_1)$ of $k(\cdot, X_1)$ satisfy

$$\sup_{j\geq 0} \mathbb{E}|\eta_j|^p < \infty, \tag{28}$$

where ϕ_j denotes the j-th eigenfunction of the kernel operator T_k.

Theorem 3.3 (Contraction rates for EVGP, LGP and CGGP, S. and Szabo)

Under Assumptions **(SPE)**, **(EVD)**, **(KLMom)**, *let* $f_0 \in \overline{\mathbb{H}} \cap L^\infty(G)$ *satisfy the concentration function inequality from Assumption* **(CFUN)**, *for a sequences* $\varepsilon_n \to 0$ *with* $n\varepsilon_n^2 \to \infty$. *Further, let*

$$\sum_{j=m_n+1}^{\infty} \lambda_j \leq C\varepsilon_n^2 \qquad \text{and} \qquad \mathbb{E}\widehat{\lambda}_{m_n+1} \leq Cn^{-1} \tag{29}$$

hold for a sequence m_n *satisfying* $C' \log n \leq m_n = o(\sqrt{n}/\log n \wedge (n^{(p/4-1)/2} \log^{p/8-1} n))$ *for some* $C' > 0$ *sufficiently large. Then, the* EVGP, LGP *and the* CGGP *approximate posteriors based on* $m_n \log n$ *actions contract around* f_0 *with rate* ε_n, *i.e., for any sequence* $M_n \to \infty$,

$$\Psi_{m_n \log n}\{d_{\mathrm{H}}(\cdot, f_0) \geq M_n\varepsilon_n\} \xrightarrow{n\to\infty} 0 \tag{30}$$

in probability under $\mathbb{P}_{f_0}^{\otimes n}$ *and* $n \to \infty$.

Example: Polynomially decaying eigenvalues

For a suitable ONB $(\phi_j)_{j\geq 1}$ of $L^2(G)$ and $Z_j \sim N(0,1)$ i.i.d., consider the random series prior

$$F(x) = \sum_{j=1}^{\infty} \tau j^{-1/2-\alpha/d} Z_j \phi_j(x), \qquad x \in \mathbb{R}^d \tag{31}$$

where $\alpha > 0$ and τ are the regularity and scale hyperparameters of the process. Then, for any

$$f_0 \in S^\beta(L) := \{f \in L^2(G) : \|f\|_{S^\beta}^2 \leq L\} \quad \text{with} \quad \|f\|_{S^\beta}^2 := \sum_{j=1}^{\infty} j^{2\beta/d} \langle f, \phi_j \rangle^2, \tag{32}$$

with $d/2 < \beta \leq \alpha + d/2$ and an apropriate choice of τ, the approximate posterior satisfies that for any $M_n \to \infty$,

$$\Psi_{m_n}\{f : d_{\mathrm{H}}(f, f_0) \geq M_n n^{-\beta/(d+2\beta)} | X, Y\} \to 0, \tag{33}$$

in probability under $\mathbb{P}_{f_0}^{\otimes n}$ and $n \to \infty$ with $m_n \sim n^{d/(2\beta+d)} \log n$.

Simulation example

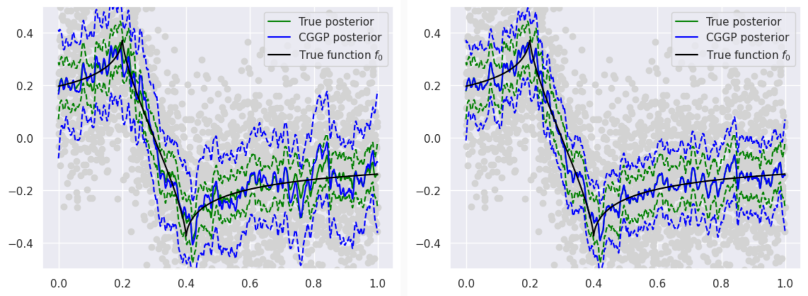


Figure 3: Simulation results for $n = 3000$, m $= 20$, 40.

Simulation example

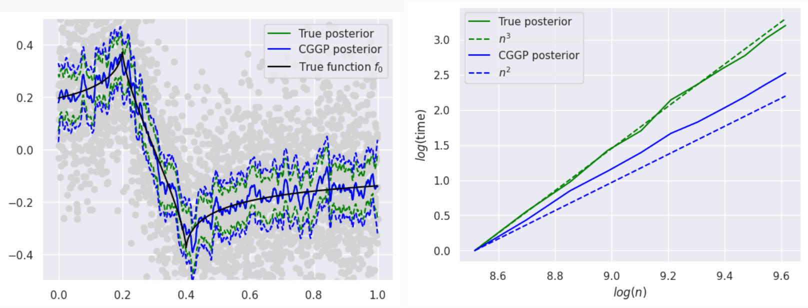


Figure 4: Simulation results for $n = 3000$, m $= 80$ and scaling of computation times.

Some conclusions

- Our theory provides new statistical guarantees for fully numerical algorithms.
- Particular relevance in the CG posterior. Default method in the `GPyTorch` library, see Gardner et al. [Gar+18].

Proof techniques

Contraction of the approximate posterior

Proposition 4.1 (Contraction of approximation, Ray and Szabó [RS19])

Under the assumptions of Proposition 3.2, let $(\Psi_{m_n})_{n\in\mathbb{N}}$ *be a sequence of distribution such that for any sequence* $M_n' \to \infty$, *there exists events* A_n' *such that*

$$\mathrm{KL}(\Psi_{m_n}, \Pi(\cdot|X, Y))\mathbf{1}_{A_n'} \leq nM_n'^2\varepsilon_n^2 \qquad \textit{and} \qquad \mathbb{P}_{f_0}^{\otimes n}(A_n') \to 1. \tag{34}$$

Then, for all sequences $M_n \to \infty$

$$\Psi_{m_n}\{d_H(\cdot, f_0) \geq M_n\varepsilon_n\} \xrightarrow[n\to\infty]{\mathbb{P}_{f_0}^{\otimes n}} 0. \tag{35}$$

Contraction of the approximate posterior

Proposition 4.1 (Contraction of approximation, Ray and Szabó [RS19])

Under the assumptions of Proposition 3.2, let $(\Psi_{m_n})_{n\in\mathbb{N}}$ *be a sequence of distribution such that for any sequence* $M_n' \to \infty$, *there exists events* A_n' *such that*

$$\mathrm{KL}(\Psi_{m_n}, \Pi(\cdot|X, Y))\mathbf{1}_{A_n'} \le nM_n'^2\varepsilon_n^2 \qquad \text{and} \qquad \mathbb{P}_{f_0}^{\otimes n}(A_n') \to 1. \tag{34}$$

Then, for all sequences $M_n \to \infty$

$$\Psi_{m_n}\{d_H(\cdot, f_0) \ge M_n\varepsilon_n\} \xrightarrow[n\to\infty]{\mathbb{P}_{f_0}^{\otimes n}} 0. \tag{35}$$

Proof sketch.

Use the dual formulation of the Kullback-Leibler divergence

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}) = \sup_{\mathbb{P}e^Z < \infty} (\mathbb{Q}Z - \log \mathbb{P}e^Z), \tag{36}$$

see Boucheron et al. [BLM13], to derive that for $\mathcal{F}_n := \{d_H(\cdot, f_0) \ge M_n\varepsilon n\}$,

$$\Psi_m(\mathcal{F}_n)\mathbf{1}_{A_n\cap A_n'} \le C\frac{\mathrm{KL}(\Psi_m, \Pi(\cdot|X, Y))\mathbf{1}_{A_n'} + e^{CnM_n^2\varepsilon_n^2/2}\Pi(\mathcal{F}_n|X, Y)\mathbf{1}_{A_n}}{nM_n^2\varepsilon_n^2}. \tag{37}$$

□

A Kullback-Leibler bound

$$
\begin{aligned}
2\,\mathrm{KL}(\Psi_m, \Pi_n(\cdot|X,Y)) &= 2\,\mathrm{KL}(N(KK_\sigma^{-1}Y, K - KK_\sigma^{-1}K), N(KC_mY, K - KC_mK)) \\
&= \mathrm{tr}(K - KK_\sigma^{-1}K)^{-1}(K - KC_mK) - n \\
&+ Y^\top(K_\sigma^{-1} - C_m)K(K - KK_\sigma^{-1}K)^{-1}K(K_\sigma^{-1} - C_m)Y \\
&+ \log\det([K - KC_mK)^{-1}[K - KK_\sigma^{-1}K]) \\
&=: (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III})
\end{aligned} \tag{38}
$$

with $K_\sigma = K + \sigma^2 I$, (III) ≤ 0 and

$$
\begin{aligned}
(\mathrm{I}) + (\mathrm{II}) &= \mathrm{tr}(K - KK_\sigma^{-1}K)^{-1}(K - KC_mK) - n + \|(K_\sigma^{-1} - C_m)Y\|^2_{K(K - KK_\sigma^{-1}K)^{-1}K} \\
&\leq \mathrm{tr}(K - KK_\sigma^{-1}K)^{-1}K(K_\sigma^{-1} - C_m^{\mathrm{EV}})K + 2\|(K_\sigma^{-1} - C_m^{\mathrm{EV}})Y\|^2_{K(K - KK_\sigma^{-1}K)^{-1}K} \\
&+ \mathrm{tr}(K - KK_\sigma^{-1}K)^{-1}K(C_m^{\mathrm{EV}} - C_m)K + 2\|(C_m - C_m^{\mathrm{EV}})Y\|^2_{K(K - KK_\sigma^{-1}K)^{-1}K},
\end{aligned} \tag{39}
$$

where $\|\cdot\|_A$ denotes the norm induced by the dot-product $\langle \cdot, A\cdot \rangle$.

Proposition 4.2 (Kullback-Leibler bound)

Under Assumptions **(SPE)**, **(EVD)**, *and* **(KLMom)**, *let* $f_0 \in \overline{\mathbb{H}} \cap L^\infty(G)$ *satisfy the concentration function inequality from Assumption* **(CFUN)** *for a sequence* $\varepsilon_n \to 0$ *with* $n\varepsilon_n^2 \to \infty$. *Additionally, let* m_n *be a sequence that satisfies* $C' \log n \leq m_n = o((\sqrt{n}/\log n) \wedge (n^{(p/4-1)/2}(\log n)^{p/8-1}))$ *for some* $C' > 0$ *sufficiently large and consider the Lanczos Algorithm 2 iterated for* $m_n \log n$ *steps initialized at* $v_0 \in \{Y/\|Y\|, Z/\|Z\|\}$, *where* Z *is a n-dimensional standard Gaussian. Then, for any sequence* $M_n \to \infty$, *the approximate posterior* Ψ_m *from Algorithm 1 based on* $m = m_n \log n$ *Lanczos actions satisfies the bound*

$$\mathrm{KL}(\Psi_{m_n \log n}, \Pi_n(\cdot|X,Y)) \leq \frac{M_n n}{\sigma^2}\Big(\varepsilon_n^2 + \sum_{j=m_n+1}^{\infty} \lambda_j + n\varepsilon_n^2 \mathbb{E}\widehat{\lambda}_{m_n+1}\Big) \tag{40}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ *and* $n \to \infty$.

Proposition 4.2 (Kullback-Leibler bound)

Under Assumptions **(SPE)**, **(EVD)**, *and* **(KLMom)**, *let* $f_0 \in \overline{\mathbb{H}} \cap L^\infty(G)$ *satisfy the concentration function inequality from Assumption* **(CFUN)** *for a sequence* $\varepsilon_n \to 0$ *with* $n\varepsilon_n^2 \to \infty$. *Additionally, let* m_n *be a sequence that satisfies* $C' \log n \leq m_n = o((\sqrt{n}/\log n) \wedge (n^{(p/4-1)/2}(\log n)^{p/8-1}))$ *for some* $C' > 0$ *sufficiently large and consider the Lanczos Algorithm 2 iterated for* $m_n \log n$ *steps initialized at* $v_0 \in \{Y/\|Y\|, Z/\|Z\|\}$, *where* Z *is a n-dimensional standard Gaussian. Then, for any sequence* $M_n \to \infty$, *the approximate posterior* Ψ_m *from Algorithm 1 based on* $m = m_n \log n$ *Lanczos actions satisfies the bound*

$$\mathrm{KL}(\Psi_{m_n \log n}, \Pi_n(\cdot|X, Y)) \leq \frac{M_n n}{\sigma^2}\Big(\varepsilon_n^2 + \sum_{j=m_n+1}^{\infty} \lambda_j + n\varepsilon_n^2 \mathbb{E}\widehat{\lambda}_{m_n+1}\Big) \tag{40}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ *and* $n \to \infty$.

Corollary 4.3 (Equivalence of LGP and CGGP)

For any integer $m \geq 1$, *the approximate posterior from Algorithm 1 based on* m *CG-actions is identical to the one resulting from the Lanczos iteration with* m *steps and starting value* $v_0 = Y/\|Y\|$. *Consequently, the bound from Proposition 4.2 also holds for the CG-approximate posterior under the same conditions.*

Lanczos bounds from Numerical analysis

Theorem 4.4 (Lanczos: Eigenvalue bound, [Saa80])

Under Assumption **(LWdf)**, *for any fixed integer* $i \le \tilde{m} < n$ *with* $\tilde{\lambda}_{i-1} > \widehat{\lambda}_i$ *if* $i > 1$, *and any integer* $\tilde{p} \le \tilde{m} - i$, *the eigenvalue approximation satisfies*

$$0 \le \widehat{\lambda}_i - \tilde{\lambda}_i \le (\widehat{\lambda}_i - \widehat{\lambda}_n)\Big(\frac{\tilde{\kappa}_i \kappa_{i,\tilde{p}} \tan(\widehat{u}_i, v_0)}{T_{\tilde{m}-i-\tilde{p}}(\gamma_i)}\Big)^2, \tag{41}$$

where $\gamma_i := 1 + 2(\widehat{\lambda}_i - \widehat{\lambda}_{i+\tilde{p}+1})/(\widehat{\lambda}_{i+\tilde{p}+1} - \widehat{\lambda}_n)$,

$$\tilde{\kappa}_i := \prod_{j=1}^{i-1} \frac{\tilde{\lambda}_j - \widehat{\lambda}_n}{\tilde{\lambda}_j - \widehat{\lambda}_i}, \qquad \kappa_{i,\tilde{p}} := \prod_{j=i+1}^{i+\tilde{p}} \frac{\widehat{\lambda}_j - \widehat{\lambda}_n}{\widehat{\lambda}_i - \widehat{\lambda}_j}, \tag{42}$$

and T_l *denotes the* l*-th Tschebychev polynomial.*

Geometric convergence

Since the Tschebychev polynomial satisfy

$$T_k(x) \ge c|x|^k, \qquad |x| \ge 1, \tag{43}$$

values $\gamma_i > 1$ guarantee geomentric convergence.

Theorem 4.5 (Lanczos: Eigenvector bound [Saa80])

Under Assumption (LWdf), for any fixed $i \le \tilde{m}$, let $(\tilde{\lambda}^*, \tilde{u}^*)$ be the approximate eigenpair from Algorithm 2 that satisfies $\widehat{\lambda}_i - \tilde{\lambda}^* = \min_{j \le \tilde{m}} \widehat{\lambda}_i - \tilde{\lambda}_j$. Then, for any integer $\tilde{p} \le \tilde{m} - i$, we have

$$\frac{1}{2}\|\tilde{u}^* \tilde{u}^{*\top} - \widehat{u}_i \widehat{u}_i^\top\|_{HS}^2 = \sin^2(\tilde{u}^*, \widehat{u}_i) \le \left(1 + \frac{\|K\|_{op}}{n\delta_i^2}\right)\left(\frac{\kappa_i \kappa_{i,\tilde{p}} \tan(\widehat{u}_i, v_0)}{T_{\tilde{m}-i-\tilde{p}}(\gamma_i)}\right)^2, \tag{44}$$

where $\delta_i^2 := \min_{\tilde{\lambda}_j \neq \tilde{\lambda}^*} |\widehat{\lambda}_i - \tilde{\lambda}_j|$, $\gamma_i := 1 + 2(\widehat{\lambda}_i - \widehat{\lambda}_{i+\tilde{p}+1})/(\widehat{\lambda}_{i+\tilde{p}+1} - \widehat{\lambda}_n)$,

$$\kappa_i := \prod_{j=1}^{i-1} \frac{\widehat{\lambda}_j - \widehat{\lambda}_n}{\widehat{\lambda}_j - \widehat{\lambda}_i}, \qquad \kappa_{i,\tilde{p}} := \prod_{j=i+1}^{i+\tilde{p}} \frac{\widehat{\lambda}_j - \widehat{\lambda}_n}{\widehat{\lambda}_i - \widehat{\lambda}_j} \tag{45}$$

and T_l denotes the l-th Tschebychev polynomial.

Challenges from spectral concentration

Theorem 4.6 (Eigenvalue concentration, Shawe-Taylor and Williams [STW02])

The empirical eigenvalues $(\widehat{\lambda}_j)_{j\le n}$ of the normalized kernel matrix K/n satisfy

(i) *For any $t > 0$ and any fixed $m \ge 1$, both*

$$\mathbb{P}\{|\widehat{\lambda}_m - \mathbb{E}\widehat{\lambda}_m| \ge t\} \le 2\exp\left(\frac{-2nt^2}{\max_x k(x,x)^4}\right) \tag{46}$$

and

$$\mathbb{P}\{|\sum_{j=m+1}^{n} \widehat{\lambda}_j - \mathbb{E}\sum_{j=m+1}^{n} \widehat{\lambda}_j| \ge t\} \le 2\exp\left(\frac{-2nt^2}{\max_x k(x,x)^4}\right). \tag{47}$$

(ii) *For any fixed $m \ge 1$,*

$$\mathbb{E}\sum_{j=1}^{m}\widehat{\lambda}_j \ge \sum_{j=1}^{m}\lambda_j \qquad \text{and} \qquad \mathbb{E}\sum_{j=m+1}^{n}\widehat{\lambda}_j \le \sum_{j=m+1}^{\infty}\lambda_j. \tag{48}$$

Challenges from spectral concentration

Proposition 4.6 (Relative perturbaton bounds, [JW23])

Under Assumptions **(SPE)** *and* **(KLMom)**, *fix* $m < m_0 \leq n$ *such that* $\lambda_{m_0} \leq \lambda_m/2$ *and further assume that*

$$\mathbf{r}_i(\Sigma) := \sum_{k \neq i} \frac{\lambda_k}{|\lambda_i - \lambda_k|} + \frac{\lambda_i}{(\lambda_{i-1} - \lambda_i) \wedge (\lambda_i - \lambda_{i+1})} \leq C\sqrt{\frac{n}{\log n}}, \quad \text{for all } i \leq m. \tag{46}$$

Then, the eigenvalues of $A = n^{-1}K$ *satisfy the relative perturbation bound*

$$\left|\frac{\widehat{\lambda}_i - \lambda_i}{\lambda_i}\right| \leq C\sqrt{\frac{\log n}{n}} \qquad \text{for all } i \leq m \tag{47}$$

with probability at least $1 - m_0^2(\log n)^{-p/4} n^{1-p/4}$.

Challenges from spectral concentration

Proposition 4.6 (Relative perturbaton bounds, [JW23])

Under Assumptions **(SPE)** *and* **(KLMom)**, *fix $m < m_0 \leq n$ such that $\lambda_{m_0} \leq \lambda_m/2$ and further assume that*

$$\mathbf{r}_i(\Sigma) := \sum_{k \neq i} \frac{\lambda_k}{|\lambda_i - \lambda_k|} + \frac{\lambda_i}{(\lambda_{i-1} - \lambda_i) \wedge (\lambda_i - \lambda_{i+1})} \leq C\sqrt{\frac{n}{\log n}}, \quad \text{for all } i \leq m. \tag{46}$$

Then, the eigenvalues of $A = n^{-1}K$ satisfy the relative perturbation bound

$$\left|\frac{\widehat{\lambda}_i - \lambda_i}{\lambda_i}\right| \leq C\sqrt{\frac{\log n}{n}} \qquad \text{for all } i \leq m \tag{47}$$

with probability at least $1 - m_0^2(\log n)^{-p/4} n^{1-p/4}$.

Martin Wahl, Bielefeld

Ongoing joint work on perturbation series for empirical eigenvalues and eigenprojectors.

Some conclusions (continued)

- Our theory provides new statistical guarantees for fully numerical algorithms.
- Particular relevance lies in the CG posterior. Default method in the `GPyTorch` library, see Gardner et al. [Gar+18].
- New interpretation of the CG posterior as a numerical approximation of a variational Bayes method.

Resources

SCAN ME Preprint

SCAN ME Author page

Thank you!

References

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. **Concentration Inequalities: A Non-asymptotic Theory of Independence.**: Oxford university press, 2013.

[Gar+18] J. Gardner et al. **"GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration".** In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

[Gv17] S. Ghosal and A. van der Vaart. **Fundamentals of nonparametric Bayesian inference.** Cambridge University Press, 2017.

[JW23] M. Jirak and M. Wahl. **"Relative perturbation bounds with applications to empirical covariance operators".** In: *Advances in Mathematics* 412 (2023), p. 108808.

[NSZ22] D. Nieman, B. Szabo, and H. van Zanten. **"Contraction rates for sparse variational approximations in Gaussian process regression".** In: *Journal of Machine Learning Research* 23 (2022), pp. 1–26.

References ii

[RS19] K. Ray and B. Szabó. **"Variational Bayes for High-Dimensional Linear Regression With Sparse Priors".** In: *Journal of the American Statistical Association* (2019). URL: https://doi.org/10.1080%2F01621459.2020.1847121.

[STW02] J. Shawe-Taylor and C. K. I. Williams. **"The stability of kernel Principal component analysis and its relation to the process eigenspectrum".** In: *Advances in Neural Information Processing Systems*. 2002.

[Saa80] Y. Saad. **"On the rates of convergence of the Lanczos and the Block-Lanczos methods".** In: *SIAM Journal on Numerical Analysis* 17.5 (1980), pp. 687–706.

[Wan+19] K. Wang et al. **"Exact Gaussian Processes on a Million Data Points".** In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[Wen+22] J. Wenger et al. **"Posterior and computational uncertainty in Gaussian processes".** In: *Advances in Neural Information Processing Systems*. 2022.