Bocconi

European Research Council
Established by the European Commission

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

Bernhard Stankewitz

Bernoulli-ims 11th World Congress in Probability and Statistics
Bochum, August 2024

Department of Decision Sciences
Bocconi University

Joint work with

Botond Szabo, Bocconi Milano

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

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

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

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

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 (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 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 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

$$\alpha_j := s_j^\top(Y-(K+\sigma^2 I_n)w_{j-1}), \qquad j \le m \tag{4}$$

where $s_j, j \le m$ are search directions chosen by the user.

[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 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

$$\alpha_j := s_j^\top (Y - (K+\sigma^2 I_n) w_{j-1}), \qquad j \leq m \tag{4}$$

  where $s_j, j \leq m$ are search directions chosen by the user.
- 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

$$x \mapsto k(X,x)^\top C_m Y \qquad (x,x') \mapsto k(x,x') - k(X,x)^\top C_m k(X,x'), \tag{5}$$

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

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

The Empirical eigenvector posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{6}$$

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

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{6}$$

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

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{7}$$

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

$$C_m = C_m^{\mathrm{EV}} = \sum_{j=1}^{m} (\widehat{\mu}_j + \sigma^2)^{-1} \widehat{u}_j \widehat{u}_j^\top . \tag{8}$$

[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

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{6}$$

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

$$x \mapsto k(X,x)^\top C_m Y \qquad (x,x') \mapsto k(x,x') - k(X,x)^\top C_m k(X,x'), \tag{7}$$

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

$$C_m = C_m^{\mathrm{EV}} = \sum_{j=1}^{m} (\widehat{\mu}_j + \sigma^2)^{-1} \widehat{u}_j \widehat{u}_j^\top . \tag{8}$$

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

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{9}$$

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.

The Lanczos posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{9}$$

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 $\Psi_m = \Psi_m^{\mathrm{L}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{10}$$

with

$$C_m = C_m^{\mathrm{L}} = \sum_{j=1}^{m} (\tilde{\mu}_j + \sigma^2)^{-1} \tilde{u}_j \tilde{u}_j^\top . \tag{11}$$

The Lanczos posterior

Consider the spectral decomposition of the empirical kernel matrix

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top \tag{9}$$

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 $\Psi_m = \Psi_m^{\mathrm{L}}$ is given by the mean and covariance function

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{10}$$

with

$$C_m = C_m^{\mathrm{L}} = \sum_{j=1}^{m} (\tilde{\mu}_j + \sigma^2)^{-1} \tilde{u}_j \tilde{u}_j^\top. \tag{11}$$

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.

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

$$\varrho(w_j) = \min_{t\in\mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{12}$$

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 $j \geq 1$,

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{12}$$

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

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{13}$$

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

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

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{12}$$

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

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{13}$$

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

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 et al. [Wan+19].

The CG posterior

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

$$\varrho(w_j) = \min_{t \in \mathbb{R}} \varrho(w_{j-1} + t d_j^{\mathrm{CG}}), \tag{12}$$

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

$$x \mapsto k(X, x)^\top C_m Y \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top C_m k(X, x'), \tag{13}$$

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 $m \ll n$.

A stylized approximate contraction result

Theorem (Approximate posterior contraction)

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

$$T_k : L^2(G) \to L^2(G), \qquad f \mapsto \int f(y)k(\cdot, y)\, G(dy) = \sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{L^2(G)} \phi_j, \tag{14}$$

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

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad i = 1, \dots n \tag{15}$$

Then, the *true posterior* Π_n *satisfies that for any sequence* $M_n \to \infty$,

$$\Pi_n(\{f \in L^2(G) : d(f, f_0) \geq M_n \varepsilon_n\} | X, Y) \to 0 \tag{16}$$

in probability under $\mathbb{P}_{f_0}^{\otimes n}$ *and* $n \to \infty$, *where* ε_n *is the optimal achievable rate implied by a concentration function inequality.*

A stylized approximate contraction result

Theorem (Approximate posterior contraction)

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

$$T_k : L^2(G) \to L^2(G), \qquad f \mapsto \int f(y)k(\cdot, y)\, G(dy) = \sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{L^2(G)} \phi_j, \tag{14}$$

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

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad i = 1, \dots n \tag{15}$$

Then, the approximate posterior Ψ_m *satisfies that for any sequence* $M_n \to \infty$,

$$\Psi_{m_n}(\{f \in L^2(G) : d(f, f_0) \geq M_n \varepsilon_n\} | X, Y) \to 0 \tag{16}$$

in probability under $\mathbb{P}_{f_0}^{\otimes n}$ *and* $n \to \infty$, *where* ε_n *is the optimal achievable rate implied by a concentration function inequality and* $m_n \to \infty$ *is an appropriate sequence.*

Example: Polynomially decaying eigenvalues

For an 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

$$F(x) = \sum_{j=1}^{\infty} \tau j^{-1/2-\alpha/d} Z_j \phi_j(x), \qquad x \in \mathbb{R}^d \tag{17}$$

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

$$f_0 \in S^\beta(L) := \{f \in L^2(G) : \|f\|_{S^\beta}^2 \leq L\} \quad \text{with} \quad \|f\|_{S^\beta}^2 := \sum_{j=1}^{\infty} j^{2\beta/d} \langle f, \phi_j \rangle^2, \tag{18}$$

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

$$\Psi_{m_n}\{f : d_{\mathrm{H}}(f, f_0) \geq M_n n^{-\beta/(d+2\beta)} | X, Y\} \to 0, \tag{19}$$

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

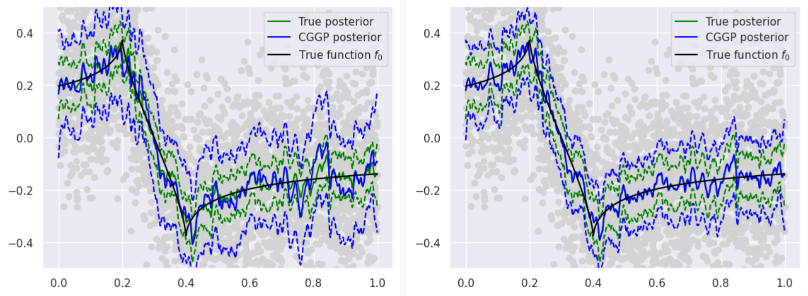
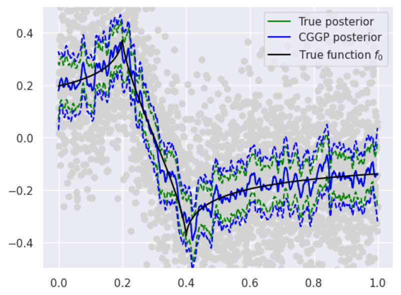


Figure 1: Simulation results for $n = 3000$, m $= 20$, 40.

Simulation example



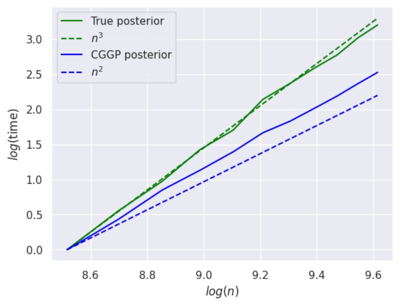


Figure 2: Simulation results for $n = 3000$, m $= 80$ and scaling of computation times.

Some conclusions

- Our theory is the first providing 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

- For $\mathrm{KL}(\Psi_{m_n}, \Pi_n(\cdot|X, Y)) \leq n\varepsilon_n^2$, the approximate posterior Ψ_{m_n} inherits the contraction rate ε_n, see Ray and Szabó [RS19].
- For the empricial eigenvector posterior with $s_j = \widehat{u}_j$, $j \leq m$ this bound is available via elementary tools.
- Analyze the Lanczos posterior as an approximation.

Lanczos bounds from Numerical analysis

Theorem (Lanczos: Eigenvalue bound, [Saa80])

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

$$0 \leq \widehat{\lambda}_i - \tilde{\lambda}_i \leq (\widehat{\lambda}_i - \widehat{\lambda}_n)\Big(\frac{\tilde{\kappa}_i \kappa_{i,\tilde{p}} \tan(\widehat{u}_i, v_0)}{T_{\tilde{m}-i-\tilde{p}}(\gamma_i)}\Big)^2, \tag{20}$$

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

$$\tilde{\kappa}_i := \prod_{j=1}^{i-1} \frac{\tilde{\lambda}_j - \widehat{\lambda}_n}{\tilde{\lambda}_j - \widehat{\lambda}_i}, \qquad \kappa_{i,\tilde{p}} := \prod_{j=i+1}^{i+\tilde{p}} \frac{\widehat{\lambda}_j - \widehat{\lambda}_n}{\widehat{\lambda}_i - \widehat{\lambda}_j}, \tag{21}$$

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

Challenges from spectral concentration

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

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

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

$$\mathbb{P}\{|\widehat{\lambda}_m - \mathbb{E}\widehat{\lambda}_m| \geq t\} \leq 2 \exp\left(\frac{-2nt^2}{\max_x k(x,x)^4}\right) \tag{22}$$

and

$$\mathbb{P}\{|\sum_{j=m+1}^{n} \widehat{\lambda}_j - \mathbb{E}\sum_{j=m+1}^{n} \widehat{\lambda}_j| \geq t\} \leq 2 \exp\left(\frac{-2nt^2}{\max_x k(x,x)^4}\right). \tag{23}$$

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

$$\mathbb{E}\sum_{j=1}^{m} \widehat{\lambda}_j \geq \sum_{j=1}^{m} \lambda_j \qquad \text{and} \qquad \mathbb{E}\sum_{j=m+1}^{n} \widehat{\lambda}_j \leq \sum_{j=m+1}^{\infty} \lambda_j. \tag{24}$$

Challenges from spectral concentration

Proposition (Relative perturbaton bounds, Jirak and Wahl [JW23])

Under appropriate assumptions, fix $m \in \mathbb{N}$ and further assume that

$$\mathbf{r}_j(T_k) := \sum_{k \neq j} \frac{\lambda_k}{|\lambda_j - \lambda_k|} + \frac{\lambda_j}{(\lambda_{j-1} - \lambda_j) \wedge (\lambda_j - \lambda_{j+1})} \leq C\sqrt{\frac{n}{\log n}}, \tag{22}$$

for all $j \leq m$. Then, the eigenvalues of K/n satisfy the relative perturbation bound

$$\left|\frac{\widehat{\lambda}_j - \lambda_j}{\lambda_j}\right| \leq C\sqrt{\frac{\log n}{n}} \qquad \text{for all } j \leq m \tag{23}$$

with high probability.

Challenges from spectral concentration

Proposition (Relative perturbaton bounds, Jirak and Wahl [JW23])

Under appropriate assumptions, fix $m \in \mathbb{N}$ and further assume that

$$\mathbf{r}_j(T_k) := \sum_{k \neq j} \frac{\lambda_k}{|\lambda_j - \lambda_k|} + \frac{\lambda_j}{(\lambda_{j-1} - \lambda_j) \wedge (\lambda_j - \lambda_{j+1})} \leq C\sqrt{\frac{n}{\log n}}, \tag{22}$$

for all $j \leq m$. Then, the eigenvalues of K/n satisfy the relative perturbation bound

$$\left|\frac{\widehat{\lambda}_j - \lambda_j}{\lambda_j}\right| \leq C\sqrt{\frac{\log n}{n}} \qquad \text{for all } j \leq m \tag{23}$$

with high probability.

Martin Wahl

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

Proof techniques

- For $\mathrm{KL}(\Psi_{m_n}, \Pi_n(\cdot|X, Y)) \leq n\varepsilon_n^2$, the approximate posterior Ψ_{m_n} inherits the contraction rate ε_n, see Ray and Szabó [RS19].
- For the empricial eigenvector posterior with $s_j = \widehat{u}_j$, $j \leq m$ this bound is available via elementary tools.
- Analyze the Lanczos posterior as an approximation.
- Establish the equivalence of the CG and the Lanczos posterior.

Some conclusions

- Our theory is the first providing 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

[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.

[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.

[RS19] K. Ray and B. Szabó. **"Variational Bayes for High-Dimensional Linear Regression With Sparse Priors".** In: *Journal of the American Statistical Association* (2019).

[STW02] J. Shawe-Taylor and C. K. I. Williams. **"The stability of kernel Principla 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.