Posterior contraction rates for banded approximations of Gaussian processes

Omar Kassi[1], Bernhard Stankewitz[2], Botond Szabo[1]
[1]Bocconi University [2]University of Potsdam

$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad f \sim \mathcal{GP}(0, k)$. The posterior is also a Gaussian Process

$$\boxed{f \mid X, Y \sim \mathcal{GP}(\mu_n, c_n)}$$

with

$$\mu_n(x) = k_x^\top (K + \sigma^2 I_n)^{-1} Y,$$
$$c_n(x, x') = k(x, x') - k_x^\top (K + \sigma^2 I_n)^{-1} k_{x'}.$$

$k_x = (k(X_1, x), \ldots, k(X_n, x))^\top, \quad K = (k(X_i, X_j))_{i,j \le n}, \quad Y = (Y_i)_{i \le n}.$

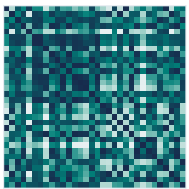
Computation: $O(n^3)$ training, $O(n^2)$ memory.

Funded by the European Union
erc European Research Council

Sparse covariance matrices

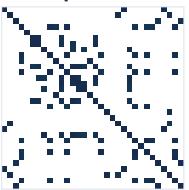
$$\widehat{K}_{ij} = k(X_i, X_j)\, \mathbf{1}\{\|X_i - X_j\| \le r\}.$$

Full K





Sparse $\widehat{K}$



m nonzeros per row

Storage	$O(nm)$
One CG step	$O(nm)$

Ordered 1D design give a banded representation.

Replace the inverse in **both** posterior formulas:

$$(K + \sigma^2 I_n)^{-1} \quad \longmapsto \quad (\widehat{K} + \sigma^2 I_n)^{-1}.$$

This gives the approximate Gaussian process $\Psi_r = \mathcal{GP}(\widehat{\mu}_r, \widehat{c}_r)$.

Contraction under controlled covariance error

Π_n: full posterior $\quad \Psi_r$: approximation $\quad \eta_n$: full-GP rate

under standard GP assumptions plus control of the approximation $\hat{K}$ in operator norm:

$$\|\widehat{K} - K\|_{\mathrm{op}} \le c < \sigma^2/2 \qquad \text{with probability tending to one.}$$

KL bound

$$\mathrm{KL}(\Psi_r, \Pi_n) \lesssim_{\mathbb{P}_{f_0}} \|\widehat{K} - K\|_{\mathrm{op}}^2 \left[\mathrm{tr}\Big(K(K + \sigma^2 I_n)^{-1}\Big) + n\eta_n^2\right].$$

Contraction. The approximation retains rate η_n if

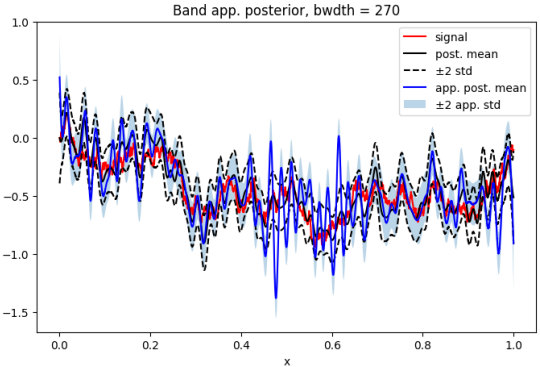
$$\mathrm{tr}\big(K(K + \sigma^2 I_n)^{-1}\big) = O_{\mathbb{P}_{f_0}}(n\eta_n^2).$$

Vanishing KL. Under this trace condition, $\|\widehat{K} - K\|_{\mathrm{op}}^2 n\eta_n^2 \to 0$ in probability also gives $\mathrm{KL}(\Psi_r, \Pi_n) \to 0$.

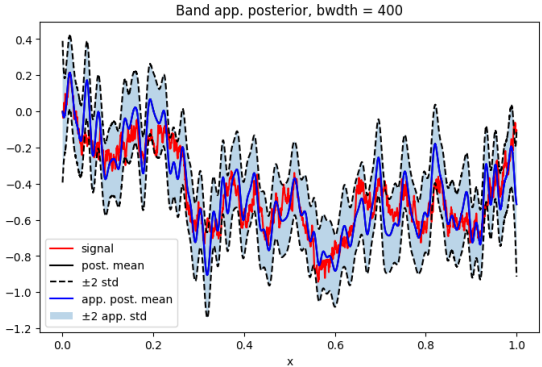
One-dimensional SE example

$d = 1$, **SE kernel,** $n = 10{,}000$, $\sigma = 1$. **Brownian truth and uniform design.**

$m = 270$ **nearest neighbours**



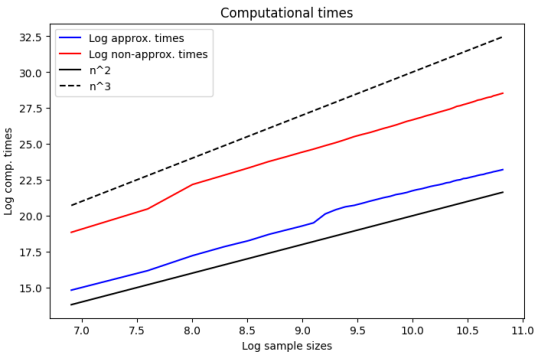
$m = 400$ **nearest neighbours**



Black: full GP **Blue**: approximation **Red**: truth

SE timings and sufficient radii

Computation times: SE kernel, $d = 1$



Blue: sparse GP. Red: full GP. Log-log axes.

Kernel-specific radius choices

Squared exponential

$$r_n = \tau_n \sqrt{2 \log n}$$

Matérn

$$r_n = 2\tau_n \log n$$

τ_n: suitably rescaled lengthscale.

Matérn sparsity gains require $\nu > \alpha$.

A suitable radius can retain the full GP contraction rate.