Fixed-domain Asymptotics of Covariance Matrices and Preconditioning

Jie Chen

IBM Thomas J. Watson Research Center

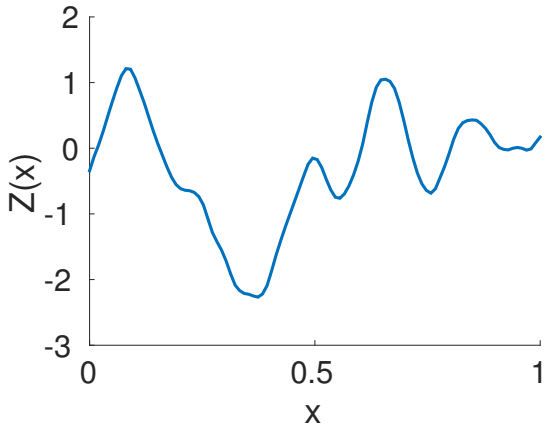
Presented at Preconditioning Conference, August 1, 2017

Gaussian processes

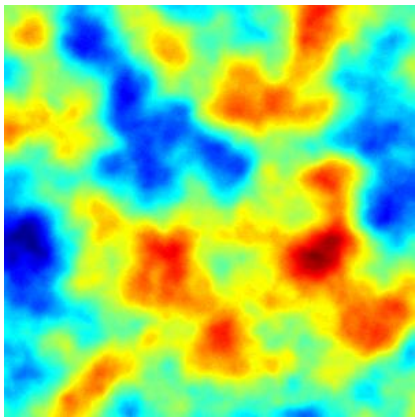
- Gaussian processes (GP) are stochastic models whereby observations are jointly Gaussian
- Notation: site $\boldsymbol{x} \in \mathbb{R}^d$,
  (random) observation $Z(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$,
  mean function $\mu(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$,
  covariance function $k(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
- For any collection of distinct sites $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the random vector $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, K)$, where

$$\boldsymbol{z} = \begin{bmatrix} Z(\boldsymbol{x}_1) \\ \vdots \\ Z(\boldsymbol{x}_n) \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu(\boldsymbol{x}_1) \\ \vdots \\ \mu(\boldsymbol{x}_n) \end{bmatrix},$$

$$K = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}.$$



1D example

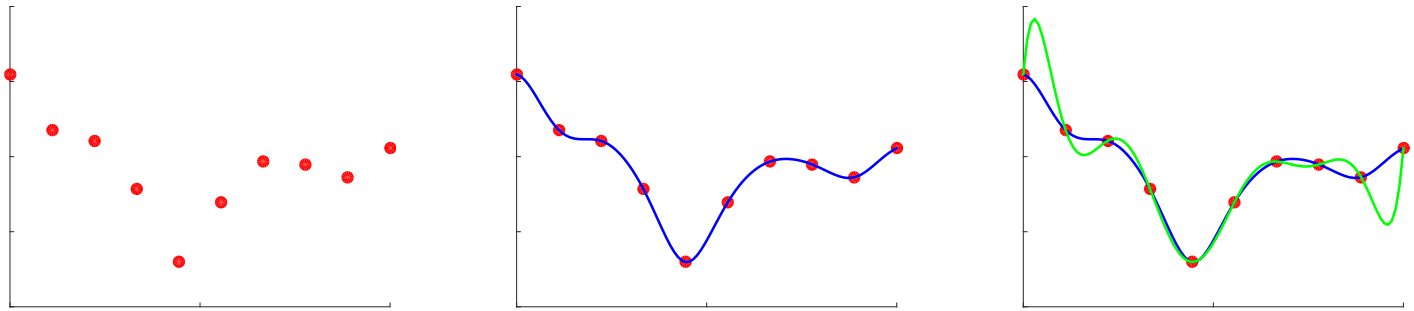


2D example

Gaussian processes

GP models may be used for:

- Sampling: Simulate random observations
- Kriging: Interpolate observations
- Model selection: What is the right interpolation?



(a) Sampling (b) Kriging (c) Model selection

- Uncertainty quantification:

$$\text{Observation} = \text{Physical model (diff. eqn.)} + \text{GP noise}$$

All calculations involve the covariance matrix K

Gaussian processes

Assume zero-mean for simplicity

- Sampling:

$$\boldsymbol{z} = \text{Cholesky}(K) \cdot \boldsymbol{y} \quad \text{or} \quad = K^{\frac{1}{2}} \boldsymbol{y}, \quad \text{where } \boldsymbol{y} \sim \mathcal{N}(\mathbf{0}_n, I_n)$$

- Kriging:

$$\widehat{Z}(\boldsymbol{x}_0) = \boldsymbol{w}_0^T K^{-1} \boldsymbol{z} \quad \text{where} \quad \boldsymbol{w}_0 = [k(\boldsymbol{x}_0, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}_0, \boldsymbol{x}_n)]^T$$
$$\text{Var}\{\widehat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\} = \text{Var}\{Z(\boldsymbol{x}_0)\} - \boldsymbol{w}_0^T K^{-1} \boldsymbol{w}_0$$

- Log-likelihood function (Assume kernel is parameterized by $\boldsymbol{\theta}$):

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2} \boldsymbol{z}^T K(\boldsymbol{\theta})^{-1} \boldsymbol{z} - \frac{1}{2} \log \det K(\boldsymbol{\theta}) - \frac{n}{2} \log 2\pi$$

- Evaluating $\mathcal{L}(\boldsymbol{\theta})$ is often a subroutine inside an optimization problem (e.g., maximum likelihood MLE, maximum a posteriori MAP, etc)

What is special about covariance matrices?

- Symetric positive definite
- Fully dense
- Increasingly ill conditioned

Positive definiteness

- K must be pd because by definition, it is *covariance*
- A bivariate function $k(\cdot,\cdot)$ is strictly pd if $\sum \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) > 0$ for all $\boldsymbol{\alpha} \neq \mathbf{0}$
- *Stationary* kernel: Simplify $k(\boldsymbol{y}, \boldsymbol{z})$ as $k(\boldsymbol{x})$, where $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{z}$

Bochner's Theorem (1D)

A function k with $k(0) = 1$ is pd if and only if it is a characteristic function.

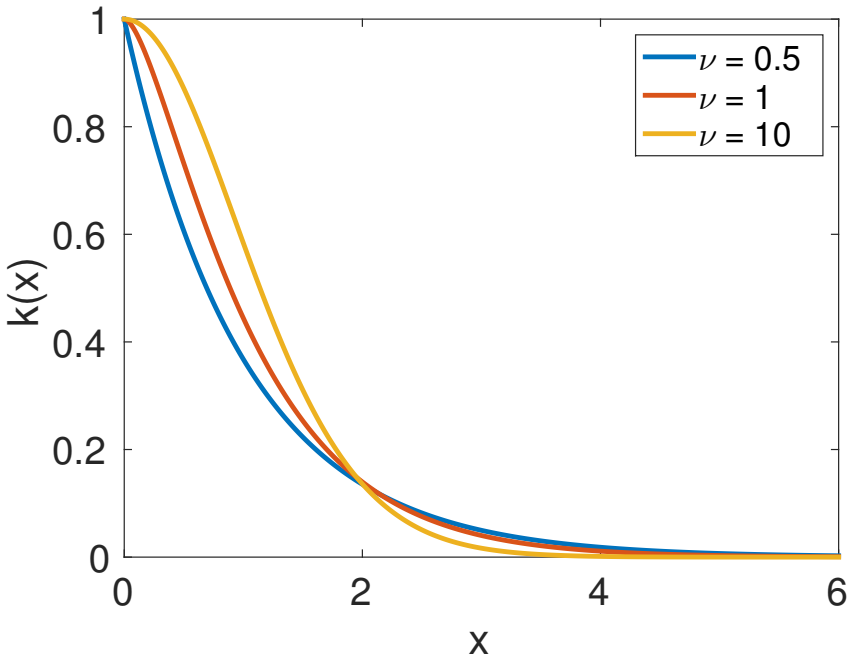
$$k(x) = \mathbb{E}[e^{\mathbf{i}x\Omega}] = \int_{\mathbb{R}} e^{\mathbf{i}x\omega}\, d\underbrace{F(\omega)}_{\text{cdf}}; \quad \text{if } F' = f, \text{ then } k(x) = \int_{\mathbb{R}} e^{\mathbf{i}x\omega} \underbrace{f(\omega)}_{\text{pdf}}\, d\omega.$$

Example: Matérn covariance functions

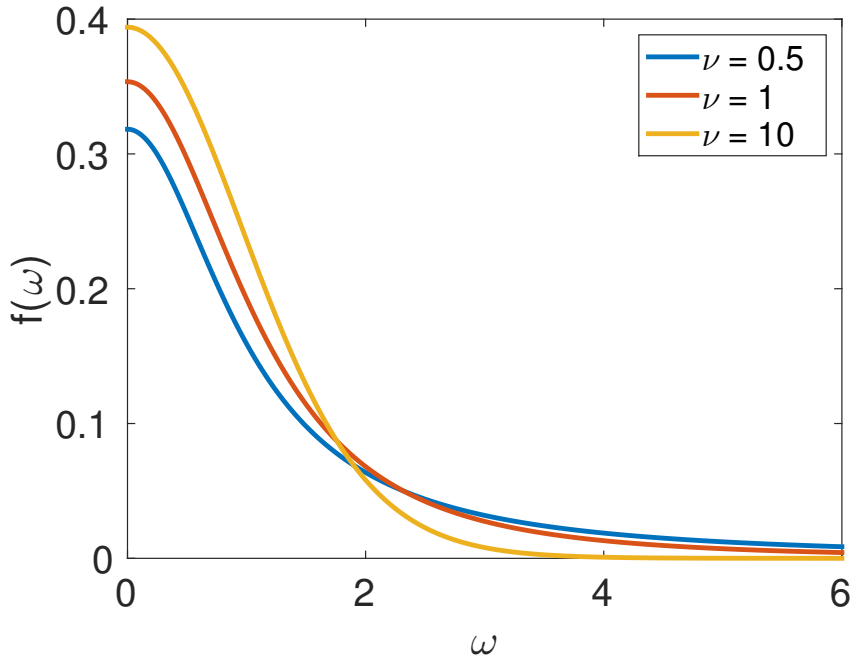
$$k(\boldsymbol{x}) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\boldsymbol{x}\|}{\ell} \right)^{\nu} \mathsf{K}_{\nu} \left(\frac{\sqrt{2\nu}\|\boldsymbol{x}\|}{\ell} \right)$$

$$f(\boldsymbol{\omega}) = \frac{(2\nu)^{\nu}\Gamma(\nu + d/2)}{\pi^{d/2}\ell^{2\nu}\Gamma(\nu)} \left(\frac{2\nu}{\ell^2} + \|\boldsymbol{\omega}\|^2 \right)^{-(\nu+d/2)} > 0$$

Positive definiteness



(a) $k(x)$



(b) $f(\omega)$

Matérn covariance function and spectral density (length scale $\ell = 1$)

Ill conditioning

Basic tools for calculating condition number

- In what follows we always assume that k is stationary and continuous
- Quadratic form of K is related to Fourier integral

$$\boldsymbol{a}^T K \boldsymbol{a} = \sum_{i,j} a_i a_j k(\boldsymbol{x}_i - \boldsymbol{x}_j) = \int_{\mathbb{R}^d} f(\boldsymbol{\omega}) \left| \sum_j a_j e^{\mathbf{i}\boldsymbol{\omega}^T \boldsymbol{x}_j} \right|^2 d\boldsymbol{\omega}.$$

- Quadratic form is also the variance of a linear combination of the random observations

$$\boldsymbol{a}^T K \boldsymbol{a} = \mathrm{Var} \left\{ \sum_j a_j Z(\boldsymbol{x}_j) \right\}.$$

Ill conditioning

Theorem

Assume the observation domain has a finite parameter. Then, the condition number $\kappa(K)$ grows faster than linearly in n.

Proof.

- Observation sites $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ become denser and denser in a fixed domain; hence one may pick two sites $\boldsymbol{y}$ and $\boldsymbol{z}$ increasingly close, such that $\mathrm{Var}\{Z(\boldsymbol{y}) - Z(\boldsymbol{z})\} \to 0$.
- Therefore, minimum eigenvalue of K tends to 0.
- There exists $r > 0$ such that $k(\boldsymbol{x}) \geq \frac{1}{2}k(\mathbf{0})$ for all $\|\boldsymbol{x}\| \leq r$. The domain may be covered by balls of diameter r. Let the number of these balls be B.
- One of the balls must contain at least $m \geq n/B$ observations.
- Hence, the sum of these observations, divided by $\sqrt{m}$, has variance at least $\frac{1}{2}k(\mathbf{0})m \geq \frac{k(\mathbf{0})}{2B}n$
- Therefore, maximum eigenvalue of K grows at least linearly in n. □

Ill conditioning

In practice, the condition number may grow *much* faster than linearly.

- Consider a regular grid $\in [0,1]^d$ with size $n_1 \times \cdots \times n_d$. Let $n = n_1 n_2 \cdots n_d$.
- Use vector indices.
- When restricted on grid, the Fourier integral may be rewritten as an integral in $[-\pi, \pi]^d$:

$$\boldsymbol{a}^T K \boldsymbol{a} = \int_{[-\pi,\pi]^d} f_{\boldsymbol{n}}(\boldsymbol{\omega}) \left| \sum_{\boldsymbol{j}} a_{\boldsymbol{j}} e^{\mathbf{i}\boldsymbol{\omega}^T \boldsymbol{j}} \right|^2 d\boldsymbol{\omega},$$

where

$$f_{\boldsymbol{n}}(\boldsymbol{\omega}) = n \sum_{\boldsymbol{l} \in \mathbb{Z}^d} f(\boldsymbol{n} \circ (\boldsymbol{\omega} + 2\pi \boldsymbol{l})), \quad \boldsymbol{\omega} \in [-\pi, \pi]^d.$$

- If f is radially decreasing,

$$\sup f_{\boldsymbol{n}} = f_{\boldsymbol{n}}(\mathbf{0}) \sim n f(\mathbf{0}) \quad \text{and} \quad \inf f_{\boldsymbol{n}} = f_{\boldsymbol{n}}(\boldsymbol{\pi}) \sim 2^d n f(\boldsymbol{n} \circ \boldsymbol{\pi}).$$

- Thus, $\kappa(K)$ increases in a polynomial rate if f decreases so.

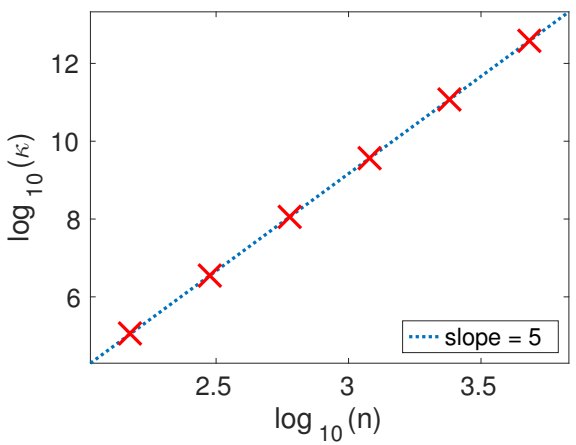
Ill conditioning

Theorem

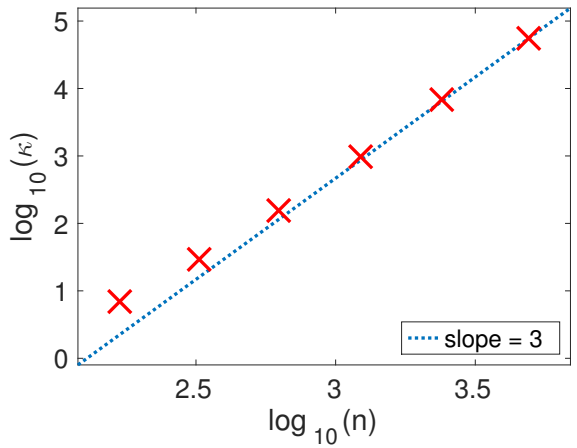
For anisotropic Matérn covariance functions

$$k(\boldsymbol{x}) = \frac{\left(\sqrt{2\nu}r\right)^{\nu} \mathsf{K}_{\nu}\left(\sqrt{2\nu}r\right)}{2^{\nu-1}\Gamma(\nu)} \quad \text{where} \quad r = \sqrt{\frac{x_1^2}{\ell_1^2} + \cdots + \frac{x_1^2}{\ell_1^2}},$$

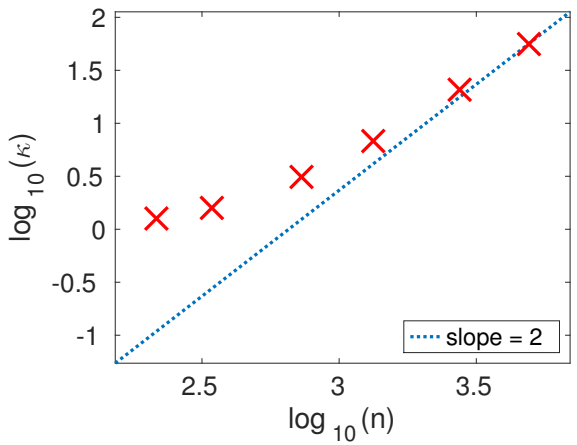
the condition number $\kappa(K)$ grows as $(\ell_1^2 n_1^2 + \cdots + \ell_d^2 n_d^2)^{\nu+d/2}$. Therefore, if the grid has the same size along each dimension, then κ grows as $n^{1+2\nu/d}$.



(a) $d = 1$, $\nu = 2$



(b) $d = 2$, $\nu = 2$



(c) $d = 3$, $\nu = 1.5$

Why preconditioning?

- (Obvious reason:) Improve convergence speed of iterative solves
- (Additional reason:) Improve parameter estimates

Parameter estimation

- A covariance function has a vector $\boldsymbol{\theta}$ of parameters. E.g., in Matérn, the parameters are ν and ℓ.
- There are several approaches for estimating $\boldsymbol{\theta}$.
- Maximum likelihood estimation approach:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad \text{where} \quad \mathcal{L} \equiv -\frac{1}{2}\boldsymbol{z}^T K(\boldsymbol{\theta})^{-1}\boldsymbol{z} - \frac{1}{2}\log\det K(\boldsymbol{\theta}) - \frac{n}{2}\log 2\pi$$

- Estimating equation approach:

$$\widehat{\boldsymbol{\theta}} \quad \text{solves} \quad \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{where} \quad \mathbb{E}[\boldsymbol{h}] = \mathbf{0}.$$

- Effectiveness of the estimation:

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathcal{G}\{\boldsymbol{h}\}^{-1}) \quad \text{if} \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, K(\boldsymbol{\theta})).$$

  where $\mathcal{G}\{\boldsymbol{h}\} \equiv \mathbb{E}[\nabla \boldsymbol{h}] \cdot \mathrm{Var}\{\boldsymbol{h}\}^{-1} \cdot \mathbb{E}[\nabla \boldsymbol{h}]$ is the Godambe information matrix.
- When the estimating equations are score equations $\boldsymbol{h} = \nabla\mathcal{L}$,

$$\mathcal{G}\{\boldsymbol{h}\} = -\nabla^2\mathcal{L}.$$

Parameter estimation

- To bypass the trace calculation, one may use approximate score equations

$$h_i = \frac{1}{2}\boldsymbol{z}^T K^{-1}(\partial_i K)K^{-1}\boldsymbol{z} - \frac{1}{2}\operatorname{tr}[K^{-1}(\partial_i K)]$$

$$\downarrow$$

$$h_i^N = \frac{1}{2}\boldsymbol{z}^T K^{-1}(\partial_i K)K^{-1}\boldsymbol{z} - \frac{1}{2N}\sum_{j=1}^{N} \boldsymbol{u}_j^T K^{-1}(\partial_i K)\boldsymbol{u}_j$$

  where the $\boldsymbol{u}_j$'s are independent symmetric Bernoulli vectors (taking ± 1 with equal probability)

Theorem

$$\mathcal{G}\{\boldsymbol{h}^N\} \succeq \left\{1 + \frac{(\kappa+1)^2}{4\kappa N}\right\}^{-1} \mathcal{G}\{\boldsymbol{h}\}, \quad \text{where } \kappa \text{ is the condition number of } K.$$

Preconditioning overview

Overview

- In all the techniques that follow, preconditioning is in the form

$$\widetilde{K} = LKL^T,$$

  where L is a discrete analog of differential operators.
- Because of domain boundary, L may have fewer rows than columns.
- Such techniques correspond to *whitening* a process:

$$\mathrm{Var}\{L\boldsymbol{z}\} = LKL^T \longrightarrow \text{well conditioned}$$

- $L \in \mathbb{R}^{m\times n}$. As long as $m \approx n$, estimation asymptotics is preserved.
- For example, the quadratic term in the likelihood function

$$\boldsymbol{z}^T K^{-1} \boldsymbol{z} \quad \xrightarrow{\text{new problem}} \quad \boldsymbol{z}^T L^T (\underbrace{LKL^T}_{\widetilde{K}})^{-1} \underbrace{L\boldsymbol{z}}_{\widetilde{z}}$$

- In some scenarios, we may get an $O(1)$ condition number.

Part 1: 1D, irregular grid

f's tail behaves like ω^{-2}

- Let $0 \leq x_0 < x_1 < \cdots < x_n \leq T$. Define $d_j = x_j - x_{j-1}$.

Theorem A

Assume that $f(\omega)\omega^2$ is bounded away from 0 and ∞ as $\omega \to \infty$. Define filtered random variables

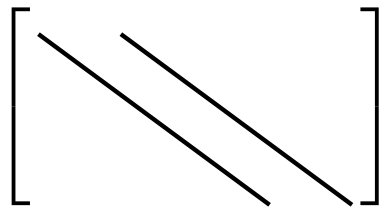
$$Y_j^{(1)} = [Z(x_j) - Z(x_{j-1})]/\sqrt{d_j}$$

and denote by $K^{(1)}$ their covariance matrix. Then, there exists a constant depending only on T and f that bounds the condition number of $K^{(1)}$ for all n.

That is,

$$K^{(1)} = L^{(1)} K {L^{(1)}}^T,$$

where $L^{(1)}_{j,j-1} = -1/\sqrt{d_j}$ and $L^{(1)}_{j,j} = 1/\sqrt{d_j}$.

$L^{(1)} =$ 

f's tail behaves like ω^{-2}

- Note that $L^{(1)}\mathbf{1} = \mathbf{0}$
- For any $\boldsymbol{a}$, the quadratic form of $K^{(1)}$ becomes

$$\boldsymbol{a}^T K^{(1)} \boldsymbol{a} = \boldsymbol{a}^T L^{(1)} K \underbrace{{L^{(1)}}^T \boldsymbol{a}}_{\boldsymbol{b}} = \boldsymbol{b}^T K \boldsymbol{b},$$

  where $\boldsymbol{b}^T \mathbf{1} = 0$.
- In other words, we only need to consider the quadratic form of K in the orthogonal complement of $\mathbf{1}$.

f's tail behaves like ω^{-2}

Proof sketch.

- Construct f_R where $f_R = f$ near the origin and $f_R \propto (1+\omega^2)^{-1}$ at the tail. Then, because $f(\omega)\omega^2$ is bounded away from 0 and ∞ at the tail, there exists C_0 and C_1 such that $C_0 f_R \leq f \leq C_1 f_R$ for all ω. Then, based on Fourier integral,

$$C_0 \boldsymbol{a}^T K_{f_R}^{(1)} \boldsymbol{a} \leq \boldsymbol{a}^T K_f^{(1)} \boldsymbol{a} \leq C_1 \boldsymbol{a}^T K_{f_R}^{(1)} \boldsymbol{a}, \qquad \forall \boldsymbol{a}. \tag{1}$$

- Brownian motion is not stationary, but it also admits a Fourier integral representation:

$$\mathrm{Var}\left\{\sum_j b_j Z(x_j)\right\} = \sum_{i,j} b_i b_j G(x_i - x_j) = \int g(\omega) \left| \sum_j b_j e^{\mathbf{i}\omega x_j} \right|^2 d\omega,$$

for all $\boldsymbol{b}$ satisfying $\boldsymbol{b}^T \mathbf{1} = 0$, where $g = \omega^{-2}$ and $G \propto |x|$.

f's tail behaves like ω^{-2}

Proof sketch (*continued*).

- We leverage some results on the equivalence of Gaussian measures:

$$P_{T,0}(f_R) \equiv P_{T,0}((1+\omega^2)^{-1}) \equiv P_{T,0}(g).$$

That is, there exists C_2 and C_3 such that

$$C_2 \operatorname{Var}_g \left\{ \sum_j b_j Z(x_j) \right\} \leq \operatorname{Var}_{f_R} \left\{ \sum_j b_j Z(x_j) \right\} \leq C_3 \operatorname{Var}_g \left\{ \sum_j b_j Z(x_j) \right\}.$$

- For any $\boldsymbol{a}$, we have $\boldsymbol{b} = {L^{(1)}}^T \boldsymbol{a}$ satisfying $\boldsymbol{b}^T \mathbf{1} = 0$. Therefore, relating the variance to the quadratic form, we have

$$C_2 \boldsymbol{a}^T K_g^{(1)} \boldsymbol{a} \leq \boldsymbol{a}^T K_{f_R}^{(1)} \boldsymbol{a} \leq C_3 \boldsymbol{a}^T K_g^{(1)} \boldsymbol{a}, \qquad \forall \boldsymbol{a}. \tag{2}$$

f's tail behaves like ω^{-2}

Proof sketch (*continued*).

- Combining (1) and (2) leads to

$$C_0 C_2 \boldsymbol{a}^T K_g^{(1)} \boldsymbol{a} \leq \boldsymbol{a}^T K_f^{(1)} \boldsymbol{a} \leq C_1 C_3 \boldsymbol{a}^T K_g^{(1)} \boldsymbol{a}, \qquad \forall \boldsymbol{a}.$$

- For Brownian motion, $K_g^{(1)}$ is a multiple of the identity. Therefore, the condition number of $K_f^{(1)}$ is bounded by $(C_1 C_3)/(C_0 C_2)$. □

f's tail behaves like ω^{-2}

In Theorem A, $L^{(1)}$ is rectangular. We now make it square.

Corollary A

Based on Theorem A, we additionally define

$$Y_0^{(1)} = Z(x_0)$$

and denote by $\widetilde{K}^{(1)}$ the covariance matrix of the $Y_j^{(1)}$'s. Then, there exists a constant depending only on T and f that bounds the condition number of $\widetilde{K}^{(1)}$ for all n.

$$\widetilde{K}^{(1)} = \widetilde{L}^{(1)} K \widetilde{L}^{(1)^T}, \qquad \widetilde{L}^{(1)} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

f's tail behaves like ω^{-4}

In Theorem A, the tail of f behaves like ω^{-2}. We now assume a different tail.

Theorem B

Assume that $f(\omega)\omega^4$ is bounded away from 0 and ∞ as $\omega \to \infty$. Define filtered random variables

$$Y_j^{(2)} = \frac{[Z(x_{j+1}) - Z(x_j)]/d_{j+1} - [Z(x_j) - Z(x_{j-1})]/d_j}{2\sqrt{d_{j+1} + d_j}}$$

and denote by $K^{(2)}$ their covariance matrix. Then, there exists a constant depending only on T and f that bounds the condition number of $K^{(2)}$ for all n.

$$K^{(2)} = L^{(2)} K {L^{(2)}}^T,$$

$$\begin{cases} L^{(2)}_{j,j-1} = (2d_j\sqrt{d_{j+1} + d_j})^{-1} \\ L^{(2)}_{j,j+1} = (2d_{j+1}\sqrt{d_{j+1} + d_j})^{-1} \\ L^{(2)}_{j,j} = -L^{(2)}_{j,j-1} - L^{(2)}_{j,j+1} \end{cases}$$

$$L^{(2)} = \begin{bmatrix} & & & & \\ & & & & \\ & & & & \end{bmatrix}$$

f's tail behaves like ω^{-4}

In Theorem B, $L^{(2)}$ is rectangular. We now make it square.

Corollary B

Based on Theorem B, we additionally define

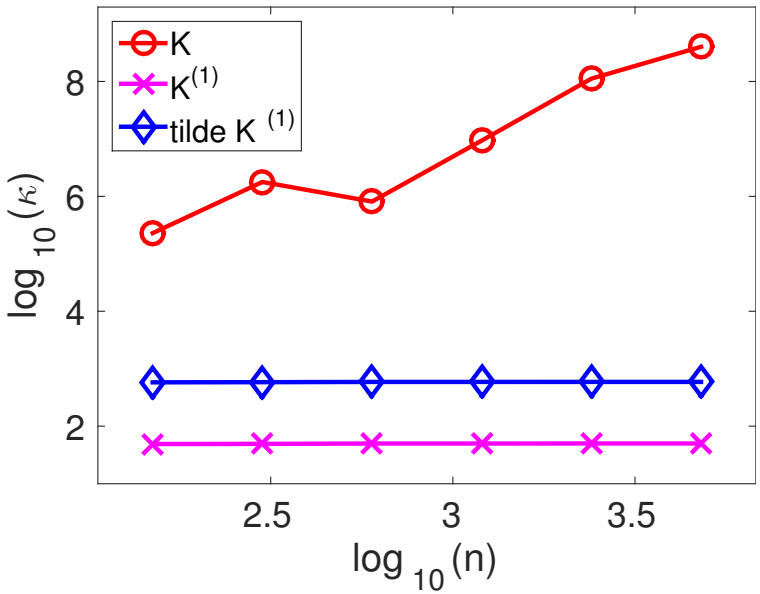
$$Y_0^{(2)} = Z(x_0) + Z(x_n) \quad \text{and} \quad Y_n^{(2)} = [Z(x_n) - Z(x_0)]/(x_n - x_0)$$

and denote by $\widetilde{K}^{(2)}$ the covariance matrix of the $Y_j^{(2)}$'s. Then, there exists a constant depending only on T and f that bounds the condition number of $\widetilde{K}^{(2)}$ for all n.

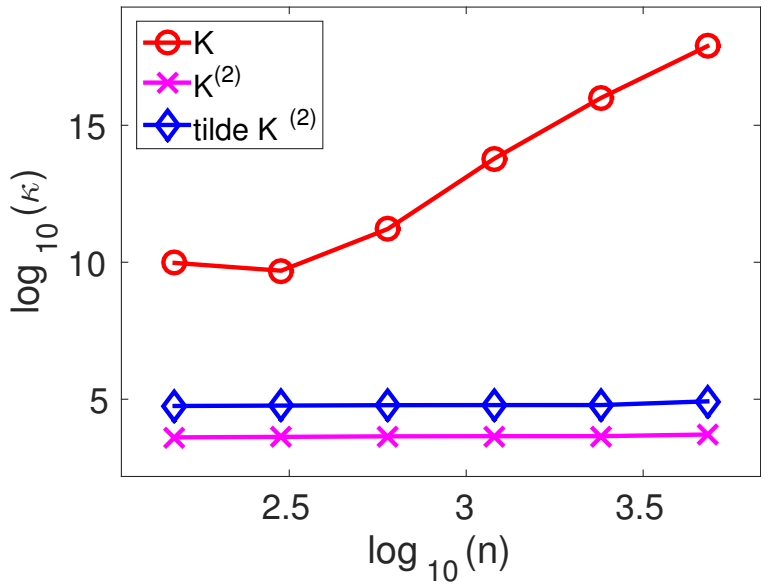
$$\widetilde{K}^{(2)} = \widetilde{L}^{(2)} K \widetilde{L}^{(2)^T}, \qquad \widetilde{L}^{(2)} = \begin{bmatrix} 1 & & & 1 \\ & \ddots & \iddots & \\ -\delta & & & \delta \end{bmatrix}, \quad \delta = 1/(x_n - x_0)$$

Numerical examples

- Uniformly random points in $[0, 1]$
- Length scale $\ell = 0.05$



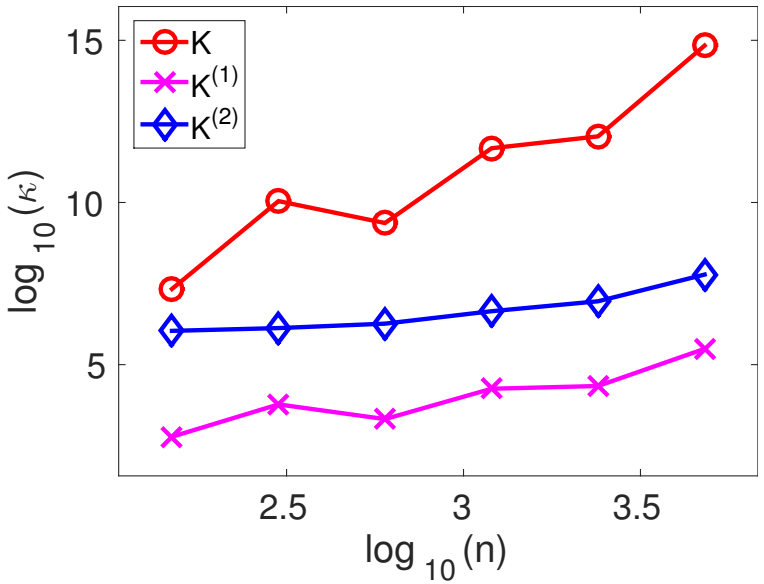
(a) $d = 1$, $\nu = 0.5$. (f tail ω^{-2})



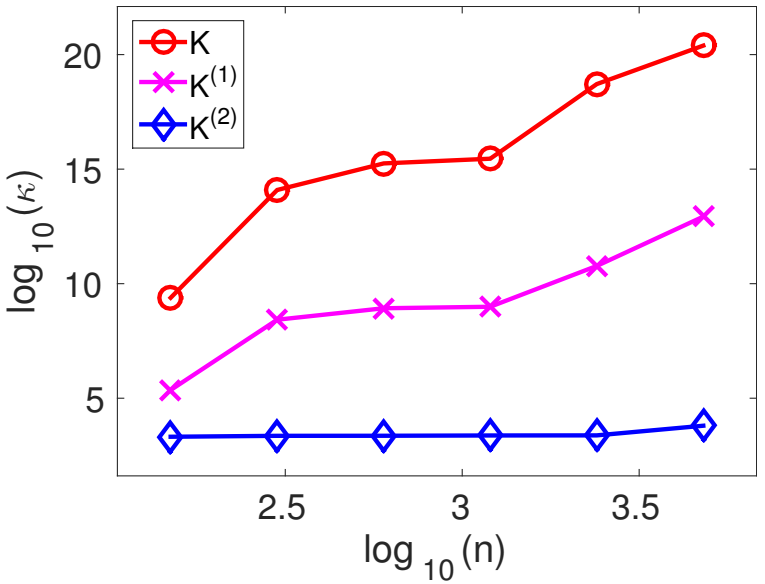
(b) $d = 1$, $\nu = 1.5$. (f tail ω^{-4})

Numerical examples

- What if the tail of f is similar to neither ω^{-2} nor ω^{-4}?
- Preconditioning is still useful (more on this in Parts 4 and 5).

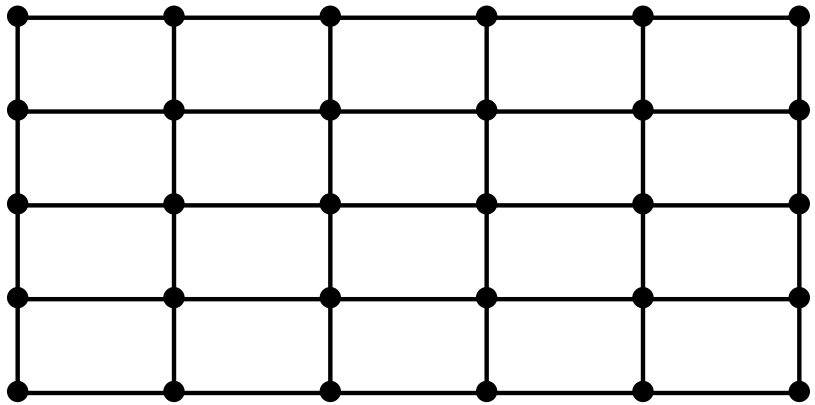


(a) $d = 1$, $\nu = 1$. (f tail $|\omega|^{-3}$)



(b) $d = 1$, $\nu = 2$. (f tail $|\omega|^{-5}$)

Part 2: d dimensions, regular grid



f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-4\tau}$

- WLOG, assume equal spacing δ along each dimension. Different spacings may be absorbed by the anisotropy of the kernel.
- Using vector index, a grid point is $\delta\boldsymbol{j}$, $\mathbf{0} \le \boldsymbol{j} \le \boldsymbol{n}$
- Define discrete Laplace operator D

$$\mathrm{D}\, Z(\boldsymbol{x}) = \sum_{p=1}^{d} Z(\boldsymbol{x} - \delta\boldsymbol{e}_p) - 2Z(\boldsymbol{x}) + Z(\boldsymbol{x} + \delta\boldsymbol{e}_p), \quad \boldsymbol{x} \in \text{grid}.$$

- For any positive integer τ, define filtered random variables

$$Y_{\boldsymbol{j}}^{[\tau]} = \mathrm{D}^{\tau}\, Z(\delta\boldsymbol{j}).$$

Theorem

Assume that $f(\boldsymbol{\omega}) \asymp (1 + \|\boldsymbol{\omega}\|)^{4\tau}$. Denote by $K^{[\tau]}$ the covariance matrix of the $Y_{\boldsymbol{j}}^{[\tau]}$'s. Then, there exists a constant depending only on the domain size and on f that bounds the condition number of $K^{[\tau]}$ for all n.

f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-4\tau}$

The preconditioned matrix

$$K^{[\tau]} = L^{[\tau]} K {L^{[\tau]}}^T$$

where

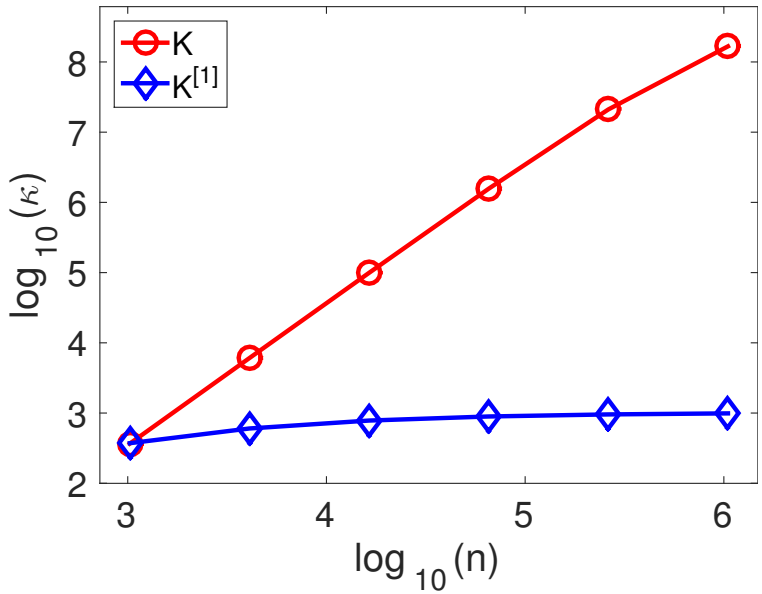
- $L^{[\tau]} = L_{n-\tau+1} \cdots L_{n-1} L_n$
- size of L_s is $(s-1)^d \times (s+1)^d$ with

$$(\boldsymbol{i}, \boldsymbol{j})\text{-element} = \begin{cases} -2d, & \boldsymbol{i} = \boldsymbol{j}, \\ 1, & \boldsymbol{i} = \boldsymbol{j} \pm \boldsymbol{e}_p,\ p = 1, \ldots, d, \\ 0, & \text{otherwise.} \end{cases}$$

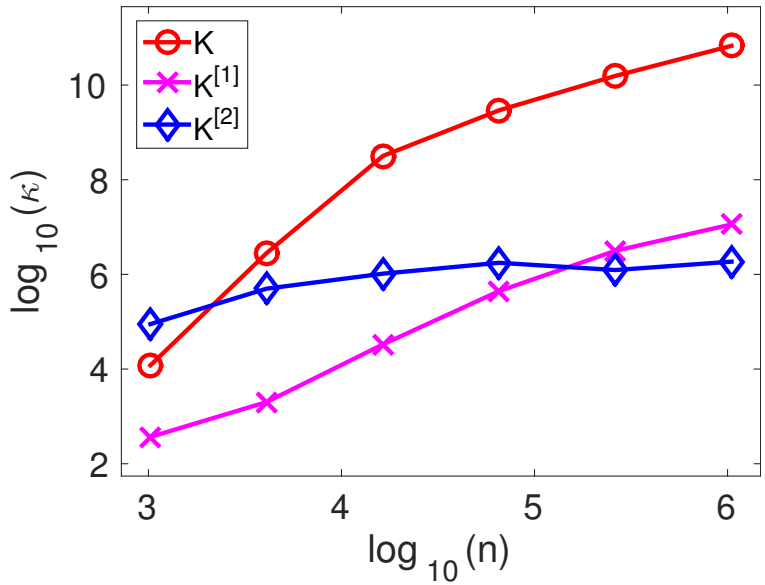
Note that the proof relies on the assumption $f(\boldsymbol{\omega}) \asymp (1 + \|\boldsymbol{\omega}\|)^{4\tau}$ *for all* $\boldsymbol{\omega}$, not just the tail.

Numerical examples

- Domain $[0,1]^2$
- Length scale $\ell_1 = 0.05$, $\ell_2 = 0.08$
- Extreme eigenvalues are estimated by using Lanczos[1]



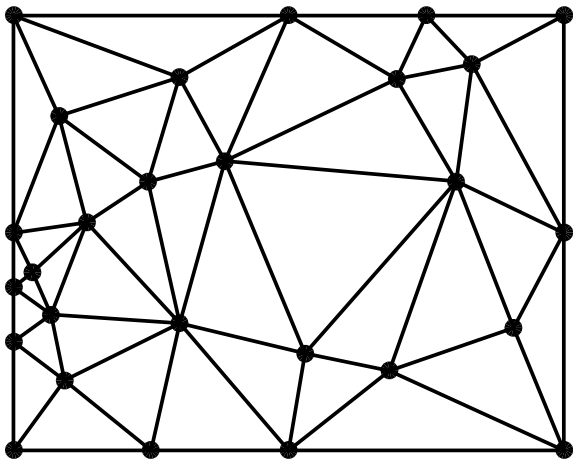
(a) $d = 2$, $\nu = 1$. $(f \asymp (1 + \|\boldsymbol{\omega}\|)^{-4})$



(b) $d = 2$, $\nu = 3$. $(f \asymp (1 + \|\boldsymbol{\omega}\|)^{-8})$

[1] Not quite accurate for the smallest eigenvalue if condition number is high

Part 3: d dimensions, irregular grid



f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-4\tau}$

- What is the discrete Laplace operator D on an irregular grid?
- For this, we borrow ideas from finite elements.
- For $u \in C^2(\Omega)$, discretize the Green's identity to obtain discrete Δu

$$\int_\Omega (v\Delta u + \nabla v \cdot \nabla u) = \oint_{\partial\Omega} v(\nabla u \cdot \boldsymbol{n}).$$

  - Use d-simplex elements and linear basis functions for simplicity
  - Let $v_i(\boldsymbol{x})$ denotes the basis at node $\boldsymbol{x}_i$
  - Result:

$$M \begin{bmatrix} \vdots \\ \Delta u(\boldsymbol{x}_i) \\ \vdots \end{bmatrix} = (S + B) \begin{bmatrix} \vdots \\ u(\boldsymbol{x}_i) \\ \vdots \end{bmatrix},$$

  where

$$\underbrace{M_{ki} = \int_\Omega v_k v_i,}_{\text{mass matrix}} \quad \underbrace{S_{ki} = -\int_\Omega \nabla v_k \cdot \nabla v_i,}_{\text{stiffness matrix}} \quad \underbrace{B_{ki} = \oint_{\partial\Omega} v_k(\nabla v_i \cdot \boldsymbol{n})}_{\text{boundary}}.$$

f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-4\tau}$

- $\mathrm{D} = M^{-1}(S + B)$?
  - Not good enough, because unlike differential equations that come with a boundary condition, we have "unknown" points $\boldsymbol{x}_i$ on the boundary.
- To make a proper definition, we need some properties of M, S, and B.

Proposition

For every k,

- $M_{kk} = 2\sum_{i \neq k} M_{ki}$
- $\sum_i S_{ki} = 0$. In particular for every $\boldsymbol{x}_k \notin \partial\Omega$, $\sum_i S_{ki}\boldsymbol{x}_i = \boldsymbol{0}$
- $\sum_i B_{ki} = 0$. In particular for every $\boldsymbol{x}_k \notin \partial\Omega$, $B_{ki} = 0$ for all i
- $\sum_i (S + B)_{ki}\boldsymbol{x}_i = \boldsymbol{0}$

- $M_{kk} = \dfrac{2}{d(d+1)} \displaystyle\sum_{E \ni \boldsymbol{x}_k} \mathrm{meas}(E)$, so M is well conditioned for "good" meshes

f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-4\tau}$

- We are now ready to deal with boundary
  - Let M' be diagonal with $M'_{kk} = \frac{3}{2} M_{kk}$ (absorbing off-diagonals)
  - For M', remove the rows and columns corresponding to boundary
  - For S and B, remove the rows corresponding to boundary
  - After removing the rows, B becomes empty
- Thus, our version of the discrete Laplace operator is the matrix L with

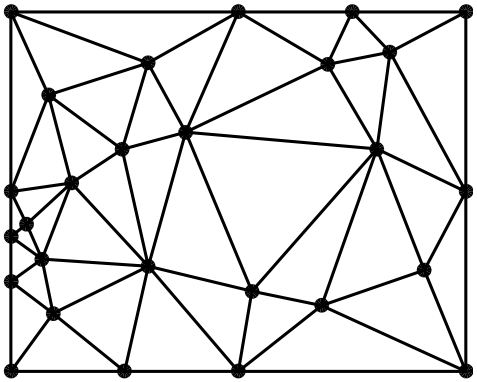
$$L_{ki} = \frac{S_{ki}}{M_{kk}}, \quad \forall \boldsymbol{x}_k \notin \partial\Omega, \ \forall \boldsymbol{x}_i$$

- Similar to the regular grid case, the preconditioned matrix

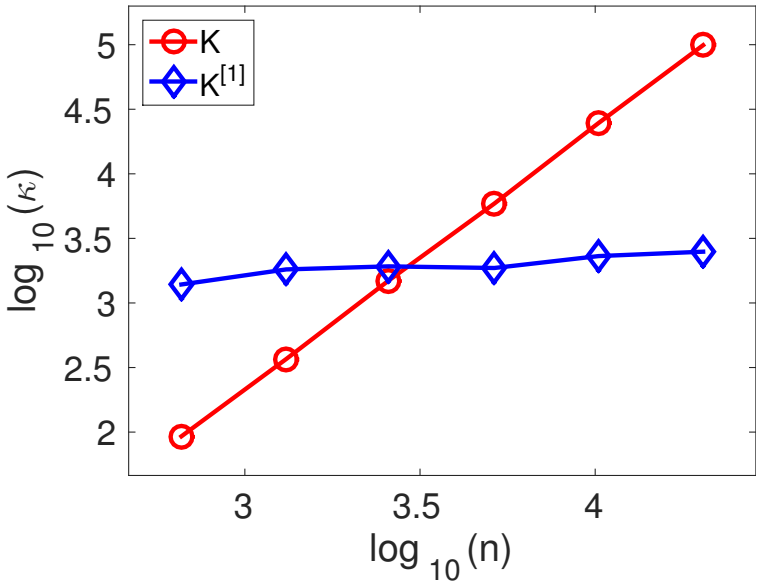
$$K^{[\tau]} = L^{[\tau]} K {L^{[\tau]}}^T$$

where $L^{[\tau]} = L_{n-\tau+1} \cdots L_{n-1} L_n$ and each L_s is a copy of L defined above, with layer(s) of boundary points removed.

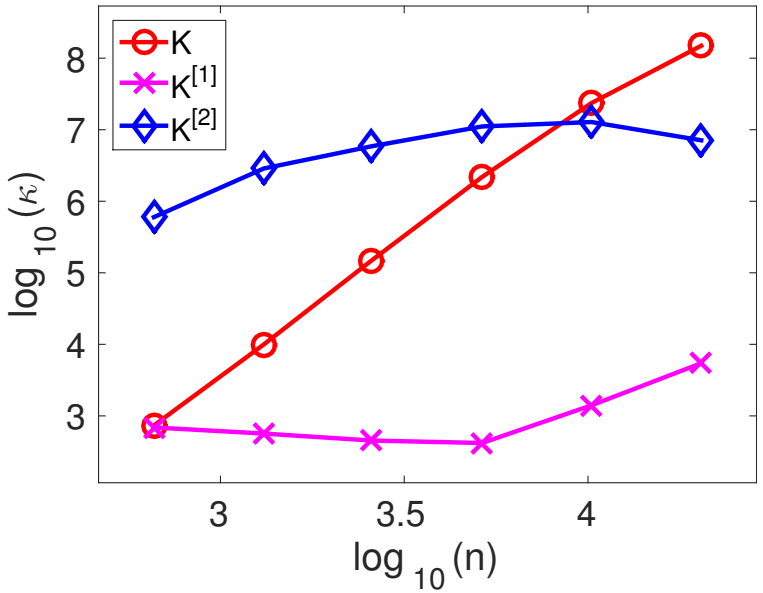
Numerical examples



- Seeded with random points inside $[0, 1] \times [0, 0.8]$
- Use `triangle` software to triangulate and refine the mesh recursively, based on area constraint
- Length scale $\ell = 0.05$



(a) $d = 2$, $\nu = 1$. $(f \asymp (1 + \|\boldsymbol{\omega}\|)^{-4})$



(b) $d = 2$, $\nu = 3$. $(f \asymp (1 + \|\boldsymbol{\omega}\|)^{-8})$

Part 4: f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-\alpha}$ for general α

f behaves like $(1 + \|\boldsymbol{\omega}\|)^{-\alpha}$

Intuitions:

- For regularly grided data, rewrite the Fourier integral as one in $[-\pi, \pi]^d$. Spectral density f becomes a periodic function $f_{\boldsymbol{n}}$. Eigenvalues of K are approximately the equally-spaced samples of $f_{\boldsymbol{n}}$:

$$\boldsymbol{a}^T K \boldsymbol{a} = \int_{[-\pi,\pi]^d} f_{\boldsymbol{n}}(\boldsymbol{\omega}) \left| \sum_{\boldsymbol{j}} a_{\boldsymbol{j}} e^{\mathbf{i}\boldsymbol{\omega}^T \boldsymbol{j}} \right|^2 \approx \frac{(2\pi)^d}{n} \sum_{\boldsymbol{k}} f_{\boldsymbol{n}}(2\pi \boldsymbol{k}/\boldsymbol{n}) \left| \sum_{\boldsymbol{j}} a_{\boldsymbol{j}} e^{\mathbf{i}(2\pi \boldsymbol{k}/\boldsymbol{n})^T \boldsymbol{j}} \right|^2$$

- After applying discrete Laplace operator $2s$ times, K becomes $K^{[s]}$ and $f_{\boldsymbol{n}}$ becomes $f_{\boldsymbol{n}}^{[s]}$, where

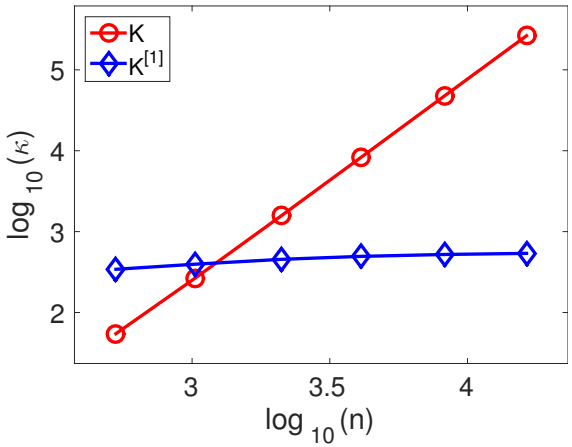
$$f_{\boldsymbol{n}}^{[s]}(\boldsymbol{\omega}) = f_{\boldsymbol{n}}(\boldsymbol{\omega}) \left[\sum_{p=1}^{d} 4 n_p^2 \sin^2 \left(\frac{\omega_p}{2} \right) \right]^{2s}.$$

- In the continuous case, the spectral density is similarly flattened:

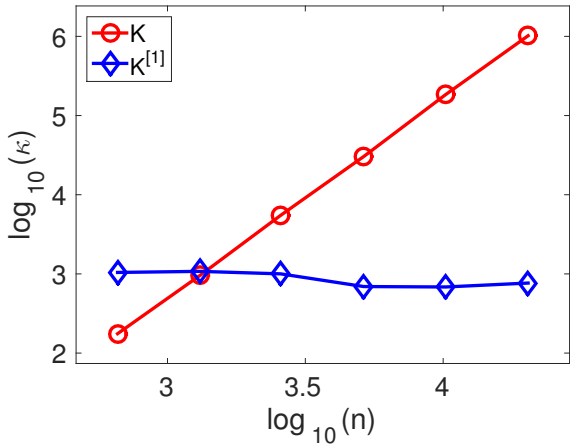
$$\Delta^{2s} k(\boldsymbol{x}) = \int_{\mathbb{R}^d} \underbrace{\|\boldsymbol{\omega}\|^{4s} f(\boldsymbol{\omega})}_{\text{decreases slower than } f} e^{\mathbf{i}\boldsymbol{\omega}^T \boldsymbol{x}} \, d\boldsymbol{\omega}.$$

Numerical examples

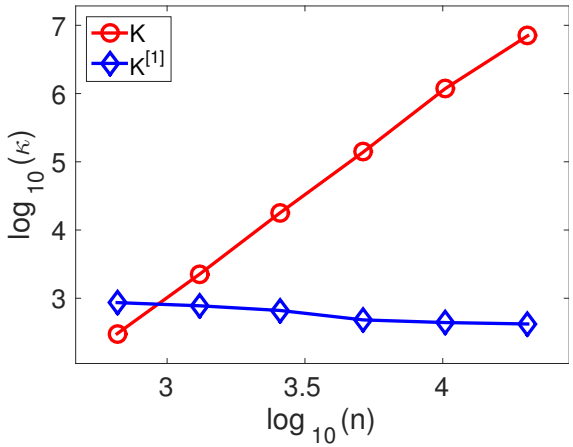
- Rule of thumb: Apply discrete Laplace operator $2s$ times such that $4s$ is closest to α
- Recall, for Matérn, $\alpha = 2\nu + d$
- $d = 2$ for all the following plots



(a) $\nu = 1.5$, regular grid



(b) $\nu = 1.5$, FEM mesh



(c) $\nu = 2$, FEM mesh

Numerical examples

Number of CG iterations to solve $K^{[s]}\boldsymbol{x} = \mathbf{1}$ such that relative residual $<$ 1e-8

	ν	s	$\log_2 n \approx$ 9	10	11	12	13	14
Grid	1	1	33	33	34	34	36	37
Grid	3	1	25	34	52	88	151	287
Grid	3	2	57	77	102	127	157	185
Grid	1.5	1	25	27	30	34	38	42
Mesh	1.5	1	97	104	103	96	95	94
Mesh	2	1	88	91	88	90	100	119

Part 5: Generalized covariance functions

Generalized covariance functions

- We have seen in the proof of Theorem A that some Gaussian processes, despite nonstationary, admit a Fourier integral representation

$$\mathrm{Var}\left\{\sum_j a_j Z(\boldsymbol{x}_j)\right\} = \sum_{i,j} a_i a_j G(\boldsymbol{x}_i - \boldsymbol{x}_j) = \int g(\boldsymbol{\omega}) \left| \sum_j a_j e^{\mathbf{i}\boldsymbol{\omega}^T \boldsymbol{x}_j} \right|^2 d\boldsymbol{\omega},$$

  for all $\boldsymbol{a}$ lying in a subspace.

- For example, the powerlaw kernel

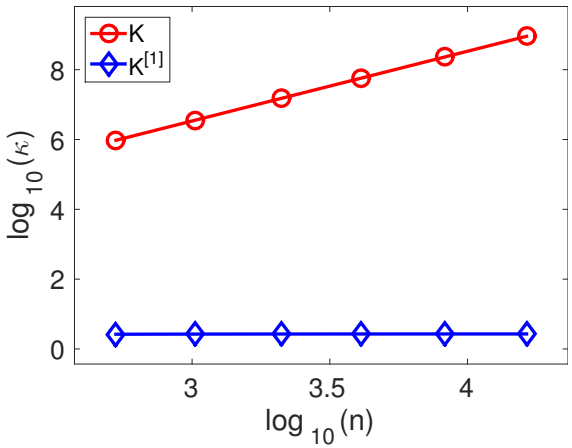
$$G(\boldsymbol{x}) = \begin{cases} \Gamma(-\beta/2)\|\boldsymbol{x}\|^\beta, & \beta/2 \notin \mathbb{N}, \\ \frac{2(-1)^{\beta/2+1}}{(\beta/2)!}\|\boldsymbol{x}\|^\beta \log\|\boldsymbol{x}\|, & \beta/2 \in \mathbb{N}, \end{cases}$$

$$g(\boldsymbol{\omega}) = \frac{2^\beta}{\pi^{d/2}}\Gamma\left(\frac{\beta+d}{2}\right)\|\boldsymbol{\omega}\|^{-\beta-d},$$

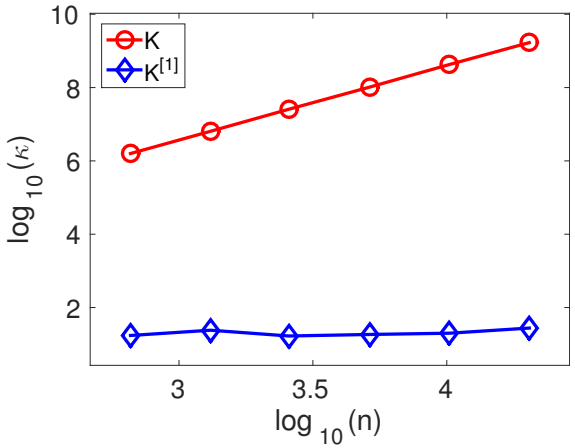
$$\sum_j a_j P(\boldsymbol{x}_j) = 0 \quad \text{for all polynomials } P \text{ of degree up to } \lfloor \beta/2 \rfloor.$$

Numerical examples

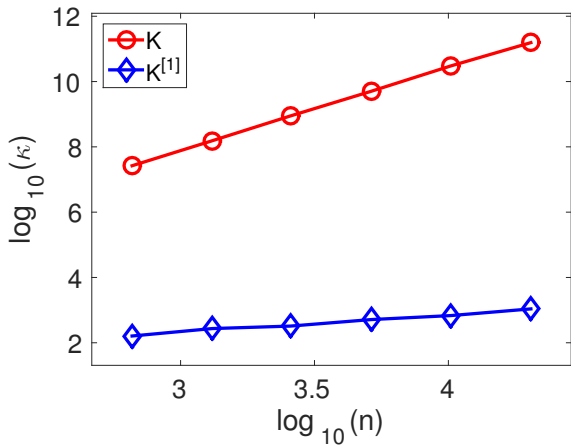
- Rule of thumb: Apply discrete Laplace operator $2s$ times such that $4s$ is closest to $\beta + d$
- $d = 2$ for all the following plots



(a) $\beta = 2$, regular grid



(b) $\beta = 2$, FEM mesh



(c) $\beta = 3$, FEM mesh

Numerical examples

Number of CG iterations to solve $K^{[s]}\boldsymbol{x} = \mathbf{1}$ such that relative residual $<$ 1e−8

	β	s	$\log_2 n \approx$ 9	10	11	12	13	14
Grid	2	1	13	13	13	13	13	13
Mesh	2	1	32	36	33	35	37	40
Mesh	3	1	60	78	85	108	128	160

Concluding remarks

- Gaussian processes pose substantial challenges for linear algebra
- We initially thought of doing things in a matrix-free way [1, 2, 3]: Turn everything (square root [4], determinant [5, 6], linear solves [7, 8], etc) into fast matvec [9, 10] + preconditioning [11, 12]
- There is a lot to exploit from the covariance function k
- For preconditioning, look at the decay rate of the Fourier transform f and differentiate it a number of times (to flatten the spectrum)
- The proposed method is mathematically interesting and it empirically works well, but is it the best approach?
- See my talk tomorrow at Minisymposium 5: Preconditioning in the Context of Radial Basis Functions, Part I. 09:45am–11:45am. FSC 1005

References

[1] Anitescu, Chen, and Wang. A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. *SISC*, 2012.

[2] Stein, Chen, and Anitescu. Stochastic approximation of score functions for Gaussian processes. *AOAS*, 2013.

[3] Anitescu, Chen, and Stein. An inversion-free estimating equations approach for Gaussian process models. *JCGS*, 2017.

[4] Chen, Anitescu, and Saad. Computing $f(A)b$ via least squares polynomial approximations. *SISC*, 2011.

[5] Chen. How accurately should I compute implicit matrix-vector products when applying the Hutchinson trace estimator? *SISC*, 2016.

[6] Chen and Saad. A posteriori error estimate for computing $\mathrm{tr}(f(A))$ by using the lanczos method. Technical report, 2017.

References

[7] Chen. A deflated version of the block conjugate gradient algorithm with an application to Gaussian process maximum likelihood estimation. Technical Report, 2011.

[8] Chen, Li, and Anitescu. A parallel linear solver for multilevel Toeplitz systems with possibly several right-hand sides. *PARCO*, 2014.

[9] Chen, Wang, and Anitescu. A fast summation tree code for Matérn kernel. *SISC*, 2014.

[10] Chen, Wang, and Anitescu. A parallel tree code for computing matrix-vector products with the Matérn kernel. Technical Report, 2013.

[11] Stein, Chen, and Anitescu. Difference filter preconditioning for large covariance matrices. *SIMAX*, 2012.

[12] Chen. On the use of discrete Laplace operator for preconditioning kernel matrices. *SISC*, 2013.