Randomized and Sparse Cholesky

for learning with Gaussian processes

Yifan Chen

Courant Institute, New York University

SIAM Mathematics of Data Science, Oct 2024

Gaussian Processes and Kernel Methods

Gaussian processes: $\xi \sim \mathcal{GP}(0, k)$

- Here $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a PSD kernel function
  For $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^d$, it holds

  $$(\xi(\mathbf{x}_1), ..., \xi(\mathbf{x}_N)) \sim \mathcal{N}(0, \Theta) \text{ where } \Theta = k(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{N \times N}$$

- Widely used in scientific computing and machine learning

Computational challenge: dense kernel matrices Θ

- In PDE applications, e.g., [Chen, Hosseni, Owhadi, Stuart 2021]

  Derivatives may rise $\Theta = \begin{pmatrix} k(\mathbf{X}, \mathbf{X}) & \Delta_{\mathbf{y}} k(\mathbf{X}, \mathbf{X}) \\ \Delta_{\mathbf{x}} k(\mathbf{X}, \mathbf{X}) & \Delta_{\mathbf{x}} \Delta_{\mathbf{y}} k(\mathbf{X}, \mathbf{X}) \end{pmatrix}$

Cubic bottleneck $O(N^3)$

Fast, Scalable Algorithms for Dense Kernel Matrices

Many approximate methods:

- Nyström approximation, inducing points, sparse GPs, random features, covariance tapering, divide-and-conquer, structured kernel interpolation, hierarchical matrices, wavelets based methods, sparse Cholesky factorization ...
- Based on low-rank/sparse ideas and their multiscale variants

Fast, Scalable Algorithms for Dense Kernel Matrices

Many approximate methods:

- Nyström approximation, inducing points, sparse GPs, random features, covariance tapering, divide-and-conquer, structured kernel interpolation, hierarchical matrices, wavelets based methods, sparse Cholesky factorization ...
- Based on low-rank/sparse ideas and their multiscale variants

Goal: This talk will discuss two basic yet efficient algorithms

- Low rank approximation: randomly pivoted Cholesky [Musco, Woodruff 2017], [Chen, Epperly, Tropp, Webber 2022], [Díaz, Epperly, Frangella, Tropp, Webber 2023], [Epperly, Tropp, Webber 2024], etc.
- Full rank approximation: sparse Cholesky [Schäfer, Sullivan, Owhadi 2021], [Schäfer, Katzfuss, Owhadi 2021], [Chen, Owhadi, Schäfer 2023], etc.

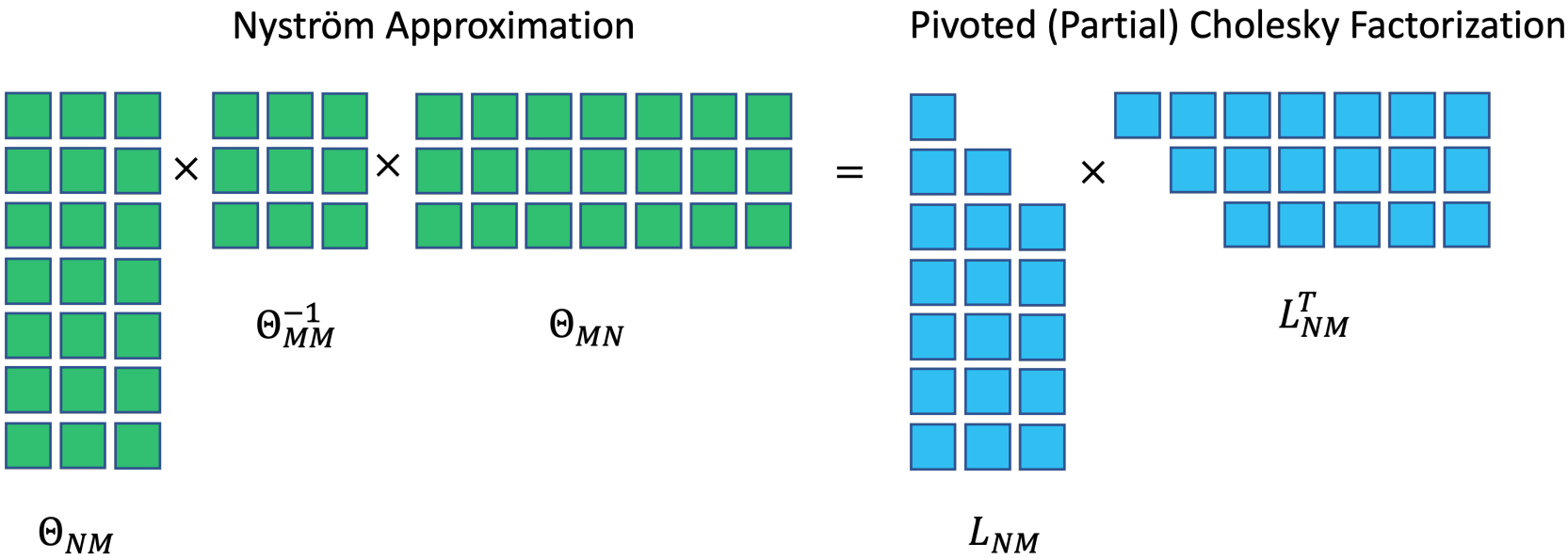
We focus on a Gaussian process interpretation

Outline

Part I: Low Rank Approximation

With M column indices $I_M = \{i_1, i_2, ..., i_M\}$, we set

$$\Theta_{MM} = \Theta[I_M, I_M] \quad \text{and} \quad \Theta_{NM} = \Theta[:, I_M]$$



- GP interpretation: Let $\Theta \in \mathbb{R}^{N\times N}$, $X \sim \mathcal{N}(0, \Theta)$
  Residue matrix $\Theta - L_{NM}L_{NM}^T = \mathrm{Cov}[X|X_{i_1}, ..., X_{i_M}]$

$$(L_{NM})_{kk} = \sqrt{\mathrm{Cov}[X_{i_k}|X_{i_1}, ..., X_{i_{k-1}}]}$$

- Small conditional variance $\rightsquigarrow$ near low rank

Experimental design, active learning, column selections, etc.

Let $\Theta \in \mathbb{R}^{N \times N}$, and $X \sim \mathcal{N}(0, \Theta)$

Uniform selection

- Choose $i_1, i_2, ..., i_M$ uniformly random

Greedy selection

- Choose $i_1 = \text{argmax}_{1 \leq i \leq N} \text{Cov}[X_i]$
- For $2 \leq k \leq M$:

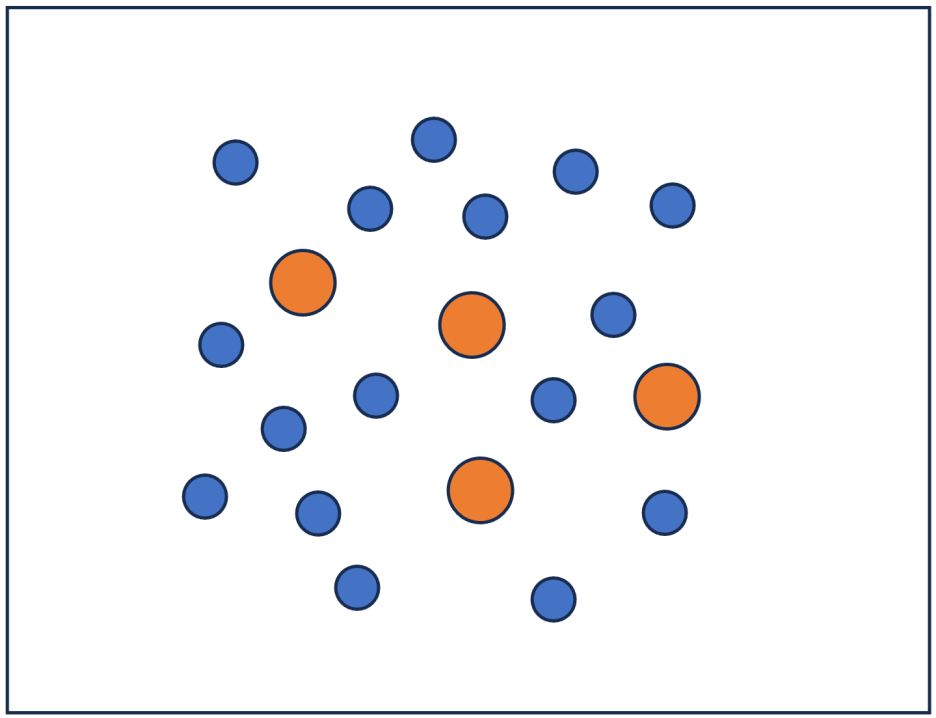
$$\text{Choose } i_k = \text{argmax}_{1 \leq i \leq N} \text{Cov}[X_i | X_{i_1}, ..., X_{i_{k-1}}]$$

Use variance uncertainties to guide the selection

Both selections can be made in arithmetic complexity $O(NM^2)$

Uniform: [Williams, Seeger 2000], [Drineas and Mahoney 2005]
Greedy: Cholesky with complete pivoting [Higham 1990]

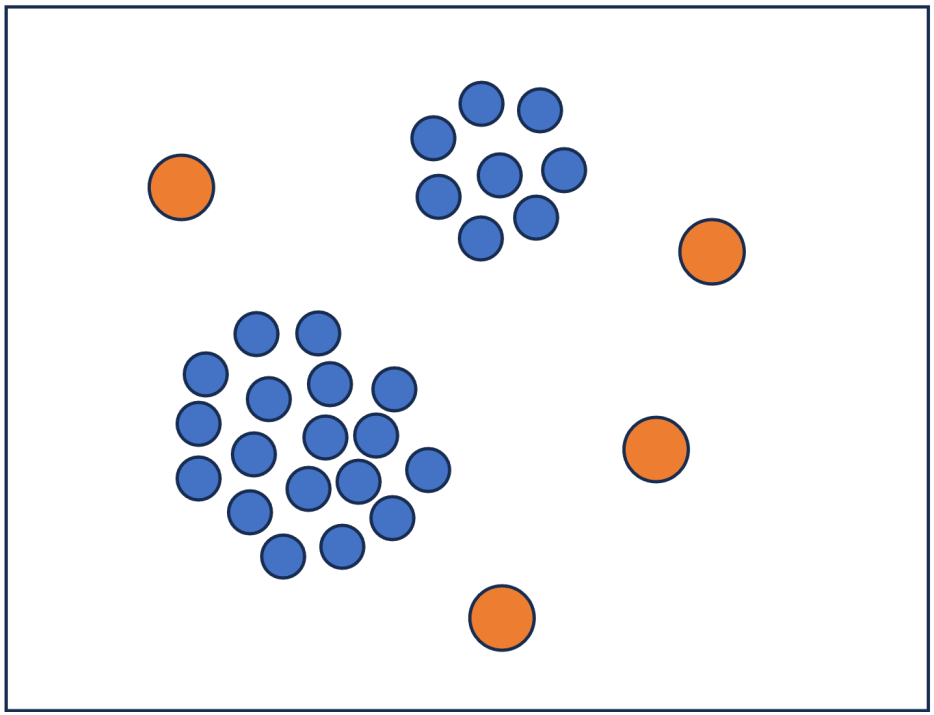
Exploration and Exploitation – Failure Mode 1



- large uncertainties
- small uncertainties

- Uniform cannot exploit large uncertainties

Exploration and Exploitation – Failure Mode 2



- large uncertainties (isolated)
- small uncertainties (clustered)

- Greedy may prioritize outliers, overlook predominant patterns
- Exploit large uncertainties too much without exploration of small uncertainties

Random Pivoting Balances Exploration and Exploitation

Random pivoting selection (RPCholesky)

- Sample i_1 according to

$$p_i \propto \mathrm{Cov}[X_i], 1 \leq i \leq N$$

- For $2 \leq k \leq M$: sample i_k according to

$$p_i \propto \mathrm{Cov}[X_i | X_{i_1}, ..., X_{i_{k-1}}], 1 \leq i \leq N$$

- Overcome the previous two failure modes
- **Theorem:** nearly optimal approximation guarantee

$$\mathbb{E}\,\mathrm{tr}\left(\Theta - L_{NM} L_{NM}^T\right) \leq (1+\varepsilon)\,\mathrm{tr}(\Theta - [\Theta]_r)$$

provided $M \geq \frac{r}{\varepsilon} + r \log\left(\frac{1}{\varepsilon\eta}\right)$ where $\eta = \frac{\mathrm{tr}(\Theta - [\Theta]_r)}{\mathrm{tr}(\Theta)}$ and $[\Theta]_r$ is the best rank-r approximation of Θ

[Chen, Epperly, Tropp, Webber 2022]

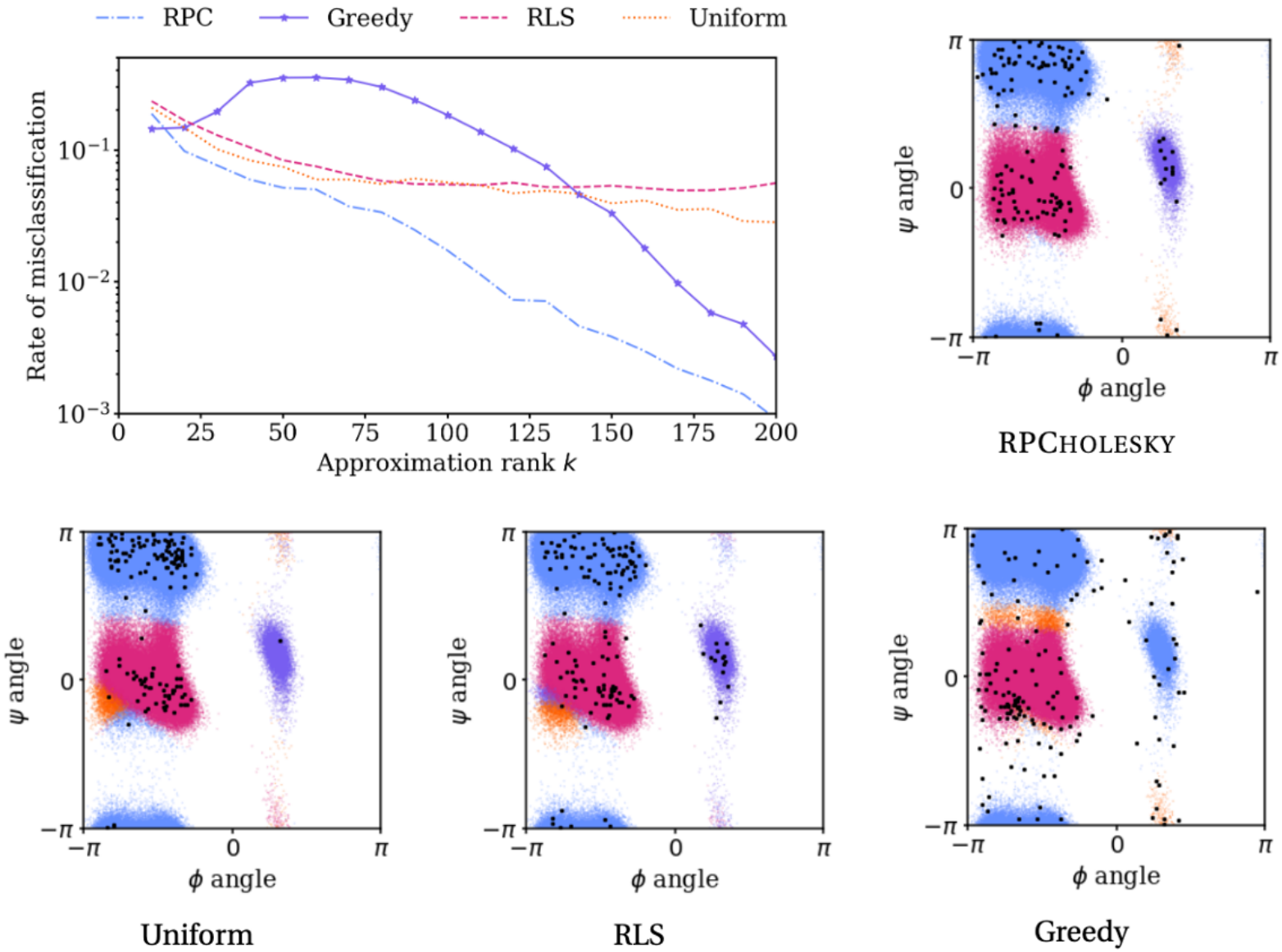


Figure: Clustering Alanine dipeptide trajectories: $N = 2.5 \times 10^5$ data points in $\mathbb{R}^{30}$ using kernel spectral clustering with low rank approximation of kernel matrices. For other numerical examples check the paper

Outline

Part II: Full Rank Approximation

When low rank approximation is not accurate enough

- In PDE applications, e.g., [Chen, Hosseni, Owhadi, Stuart 2021]

$$\Theta = \begin{pmatrix} k(\mathbf{X},\mathbf{X}) & \Delta_{\mathbf{y}} k(\mathbf{X},\mathbf{X}) \\ \Delta_{\mathbf{x}} k(\mathbf{X},\mathbf{X}) & \Delta_{\mathbf{x}}\Delta_{\mathbf{y}} k(\mathbf{X},\mathbf{X}) \end{pmatrix} \in \mathbb{R}^N$$

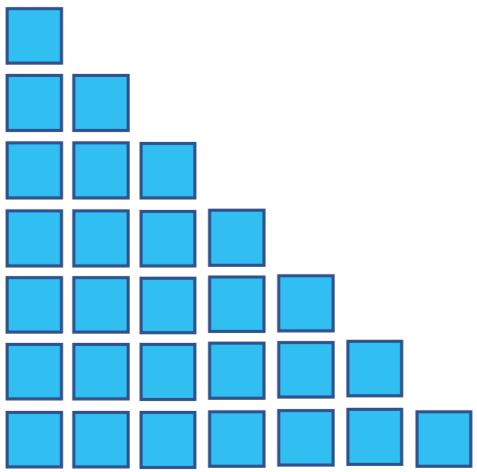
  arises when using GPs to solve $-\Delta u + u^3 = f$, etc.

- For $k(\mathbf{X},\mathbf{X})$, one can order the columns from coarse to fine scales for nearly optimal low rank approximations

- Spectrum of Θ can decay slowly for some Matérn kernels
  $\rightsquigarrow$ full rank approximation needed

Full Rank but Sparse Cholesky Factors?

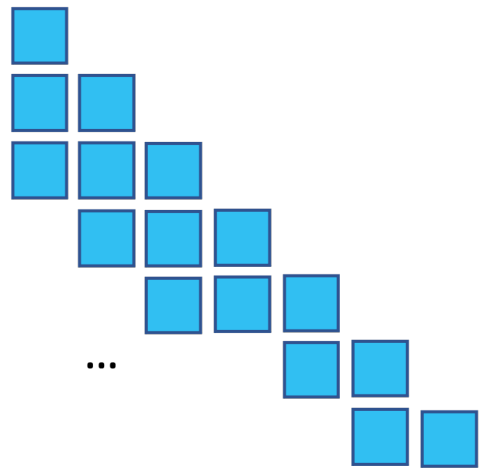
Pivoted Full Cholesky Factorization



L_{NN}

$\overset{?}{\approx}$

Pivoted Sparse Cholesky Factorization



$\hat{L}_{NN}$

Full Cholesky is not affordable. Is the factor approximately sparse?

Cholesky Factors: GP Interpretation+

GP interpretation for off-diagonals

Let $\Theta \in \mathbb{R}^{N\times N}$, and $X \sim \mathcal{N}(0, \Theta)$

- Lower-triangular Cholesky factor of $\Theta = LL^T$ satisfies

$$\frac{L_{ij}}{L_{jj}} = \frac{\mathrm{Cov}[X_i, X_j | X_1, ..., X_{j-1}]}{\mathrm{Var}[X_j | X_1, ..., X_{j-1}]} \qquad (i \geq j)$$

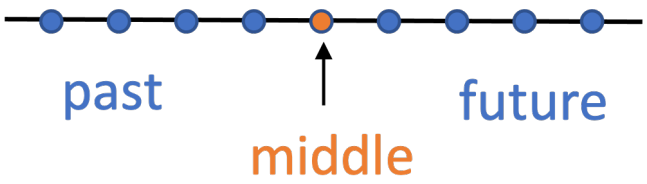
- Upper-triangular Cholesky factor of $\Theta^{-1} = UU^T$

$$\frac{U_{ij}}{U_{jj}} = (-1)^{i\neq j} \frac{\mathrm{Cov}[X_i, X_j | X_{1:j-1\setminus\{i\}}]}{\mathrm{Var}[X_j | X_{1:j-1\setminus\{i\}}]} \qquad (i \leq j)$$

- Near conditional independence $\rightsquigarrow$ near sparsity

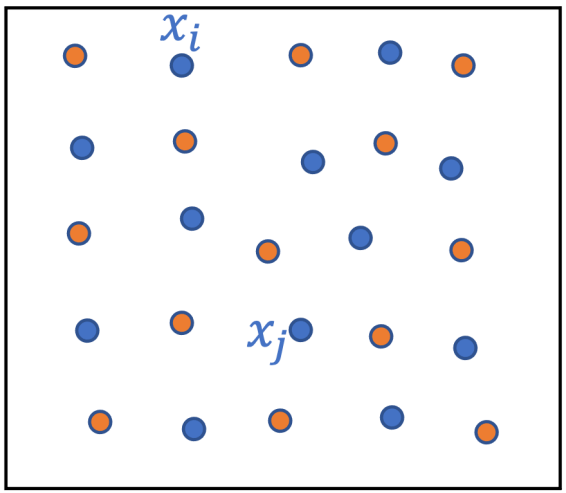
Screening Effects and Sparsity

"The screening effect is the geostatistical term for the phenomenon of nearby observations tending to reduce the influence of more distant observations when using kriging (optimal linear prediction) for spatial interpolation" [Stein 2002]



$$k(x, y) = \exp(-|x - y|)$$

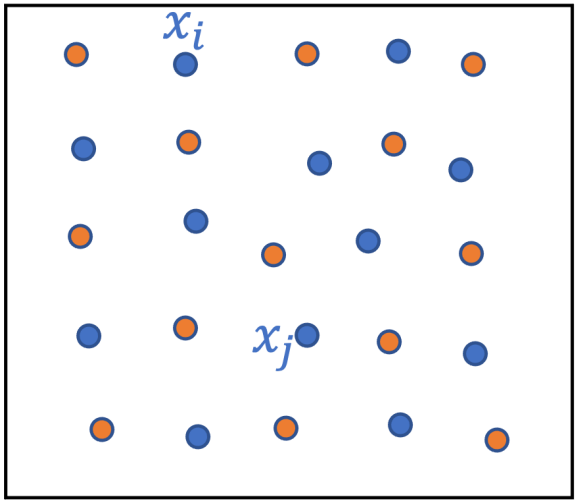
Cov [past, future | middle] = 0



Matérn's kernel

Cov [fine $\xi(x_i)$, fine $\xi(x_j)$ | ξ(coarse)] << 1

if x_i and x_j are well separated
$|x_i - x_j| \gtrsim$ coarse length scale



Matérn's kernel

Cov [fine $\xi(x_i)$, fine $\xi(x_j)$ | ξ(coarse)] << 1

if x_i and x_j are well separated
$|x_i - x_j| \gtrsim$ coarse length scale

- **Max-min ordering**: coarse-to-fine through

$$\mathbf{x}_k = \text{argmax}_{\mathbf{x}_i} \text{dist}(\mathbf{x}_i, \{\mathbf{x}_j, 1 \leq j < k\})$$

with its length scale $l_k := \text{dist}(\mathbf{x}_k, \{\mathbf{x}_j, 1 \leq j < k\})$

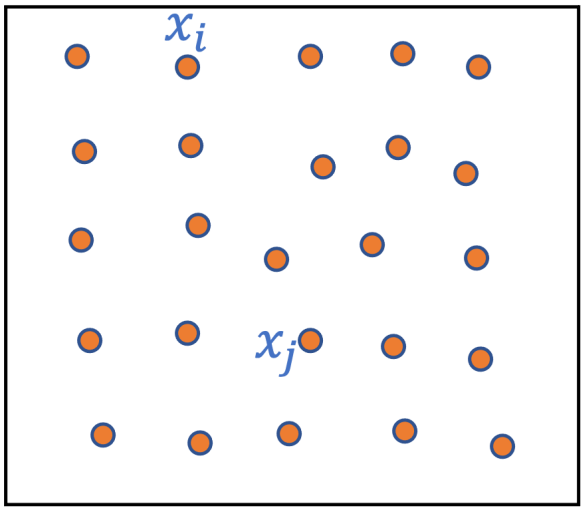
- Under the assumption that k is the Green function of certain differential operators $\mathcal{L} : H_0^s(\Omega) \to H^{-s}(\Omega)$, it holds

$$\text{Cov}[\xi(\mathbf{x}_i), \xi(\mathbf{x}_j) | \xi(\mathbf{x}_{1:k})] \leq C l_k^\alpha \exp\left(-\frac{\text{dist}(\mathbf{x}_i, \mathbf{x}_j)}{C l_k}\right)$$

[Schäfer, Sullivan, Owhadi 2021], [Schäfer, Katzfuss, Owhadi 2021]

PDE tools: [Målqvist, Peterseim 2014], [Owhadi 2015], [Owhadi, Scovel 2017]

For Kernel Matrices Arising from PDEs



Matérn's kernel

Cov $[\Delta\xi(x_i), \Delta\xi(x_j) \mid \xi(\text{points})] \ll 1$

Cov $[\Delta\xi(x_i), \xi(x_j) \mid \xi(\text{points})] \ll 1$

if $|x_i - x_j| \gtrsim$ length scale

Theorem: Let the permutation matrix P order pointwise entries first using max-min ordering, followed by an arbitrary order of derivative entries. Let $P^T\Theta^{-1}P = U^\star U^{\star T}$. Then under certain regularity assumption on k, it holds that

$$|U^\star_{ij}| \leq C l_j^\beta \exp\left(-\frac{\mathrm{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)})}{C l_j}\right), \ 1 \leq i \leq j \leq N$$

Here $\mathbf{x}_{P(i)}$ is the point corresponding to the ith ordered entry

[Chen, Owhadi, Schäfer 2023]

Computing Sparse Factors with Near Linear Complexity

Sparsity pattern: entries are exponentially small outside

$$\begin{aligned} S_{l,\rho} =& \{1 \le i \le j \le N : \operatorname{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)}) \le \rho l_j\} \\ =& \{1 \le i \le j \le N : i \in s_j\} \quad \#s_j = O(\rho^d) \end{aligned}$$

Computing Sparse Factors with Near Linear Complexity

Sparsity pattern: entries are exponentially small outside

$$\begin{aligned} S_{l,\rho} =& \{1 \le i \le j \le N : \operatorname{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)}) \le \rho l_j\} \\ =& \{1 \le i \le j \le N : i \in s_j\} \quad \#s_j = O(\rho^d) \end{aligned}$$

Algorithm: Given the sparsity pattern, using KL optimization to extract an optimal sparse factor U^ρ

- Let $\mathcal{S}_{l,\rho} = \{A \in \mathbb{R}^{N\times N} : A_{ij} \neq 0 \Rightarrow (i,j) \in S_{l,\rho}\}$

$$U^\rho = \operatorname{argmin}_{U\in\mathcal{S}_{l,\rho}} \operatorname{KL}\left(\mathcal{N}(0, P\Theta P^T) \parallel \mathcal{N}(0, (UU^T)^{-1})\right)$$

[Vecchia 1988], [Marzouk et al. 2016], [Schäfer, Katzfuss, Owhadi 2021]

Computing Sparse Factors with Near Linear Complexity

Sparsity pattern: entries are exponentially small outside

$$\begin{aligned} S_{l,\rho} =& \{1 \le i \le j \le N : \operatorname{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)}) \le \rho l_j\} \\ =& \{1 \le i \le j \le N : i \in s_j\} \quad \#s_j = O(\rho^d) \end{aligned}$$

Algorithm: Given the sparsity pattern, using KL optimization to extract an optimal sparse factor U^ρ

- Let $\mathcal{S}_{l,\rho} = \{A \in \mathbb{R}^{N\times N} : A_{ij} \neq 0 \Rightarrow (i,j) \in S_{l,\rho}\}$

$$U^\rho = \operatorname{argmin}_{U\in\mathcal{S}_{l,\rho}} \operatorname{KL}\left(\mathcal{N}(0, P\Theta P^T) \parallel \mathcal{N}(0, (UU^T)^{-1})\right)$$

- Explicit solution: $U_{s_j,j} = \dfrac{\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}{\sqrt{\mathbf{e}^T_{\#s_j}\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}}$

[Vecchia 1988], [Marzouk et al. 2016], [Schäfer, Katzfuss, Owhadi 2021]

Computing Sparse Factors with Near Linear Complexity

Sparsity pattern: entries are exponentially small outside

$$\begin{aligned} S_{l,\rho} =& \{1 \le i \le j \le N : \operatorname{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)}) \le \rho l_j\} \\ =& \{1 \le i \le j \le N : i \in s_j\} \quad \#s_j = O(\rho^d) \end{aligned}$$

Algorithm: Given the sparsity pattern, using KL optimization to extract an optimal sparse factor U^ρ

- Let $\mathcal{S}_{l,\rho} = \{A \in \mathbb{R}^{N\times N} : A_{ij} \neq 0 \Rightarrow (i,j) \in S_{l,\rho}\}$

$$U^\rho = \operatorname{argmin}_{U\in\mathcal{S}_{l,\rho}} \operatorname{KL}\left(\mathcal{N}(0, P\Theta P^T) \parallel \mathcal{N}(0, (UU^T)^{-1})\right)$$

- Explicit solution: $U_{s_j,j} = \dfrac{\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}{\sqrt{\mathbf{e}^T_{\#s_j}\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}}$
- Can be implemented $O(N\rho^d)$ in space and $O(N\rho^{2d})$ time

[Vecchia 1988], [Marzouk et al. 2016], [Schäfer, Katzfuss, Owhadi 2021]

Computing Sparse Factors with Near Linear Complexity

Sparsity pattern: entries are exponentially small outside

$$\begin{aligned} S_{l,\rho} =& \{1 \le i \le j \le N : \mathrm{dist}(\mathbf{x}_{P(i)}, \mathbf{x}_{P(j)}) \le \rho l_j\} \\ =& \{1 \le i \le j \le N : i \in s_j\} \quad \#s_j = O(\rho^d) \end{aligned}$$

Algorithm: Given the sparsity pattern, using KL optimization to extract an optimal sparse factor U^ρ

- Let $\mathcal{S}_{l,\rho} = \{A \in \mathbb{R}^{N\times N} : A_{ij} \neq 0 \Rightarrow (i,j) \in S_{l,\rho}\}$

$$U^\rho = \mathrm{argmin}_{U\in\mathcal{S}_{l,\rho}} \mathrm{KL}\left(\mathcal{N}(0, P\Theta P^T) \parallel \mathcal{N}(0, (UU^T)^{-1})\right)$$

- Explicit solution: $U_{s_j,j} = \dfrac{\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}{\sqrt{\mathbf{e}^T_{\#s_j}\Theta^{-1}_{s_j,s_j}\mathbf{e}_{\#s_j}}}$
- Can be implemented $O(N\rho^d)$ in space and $O(N\rho^{2d})$ time
- Theory: $\rho = O(\log(N/\epsilon)) \Rightarrow \|P^T\Theta^{-1}P - U^\rho(U^\rho)^T\|_{\mathsf{Fro}} \le \epsilon$

[Vecchia 1988], [Marzouk et al. 2016], [Schäfer, Katzfuss, Owhadi 2021]

Experiments using Sparse Chokesky for Solving PDEs

- 2D Example: nonlinear elliptic equation with $\tau(u) = u^3$

$$-\Delta u + \tau(u) = f \quad \text{w/ Dirichlet's boundary condition}$$

- $\Omega = [0,1]^2$. Collocation points uniformly distributed

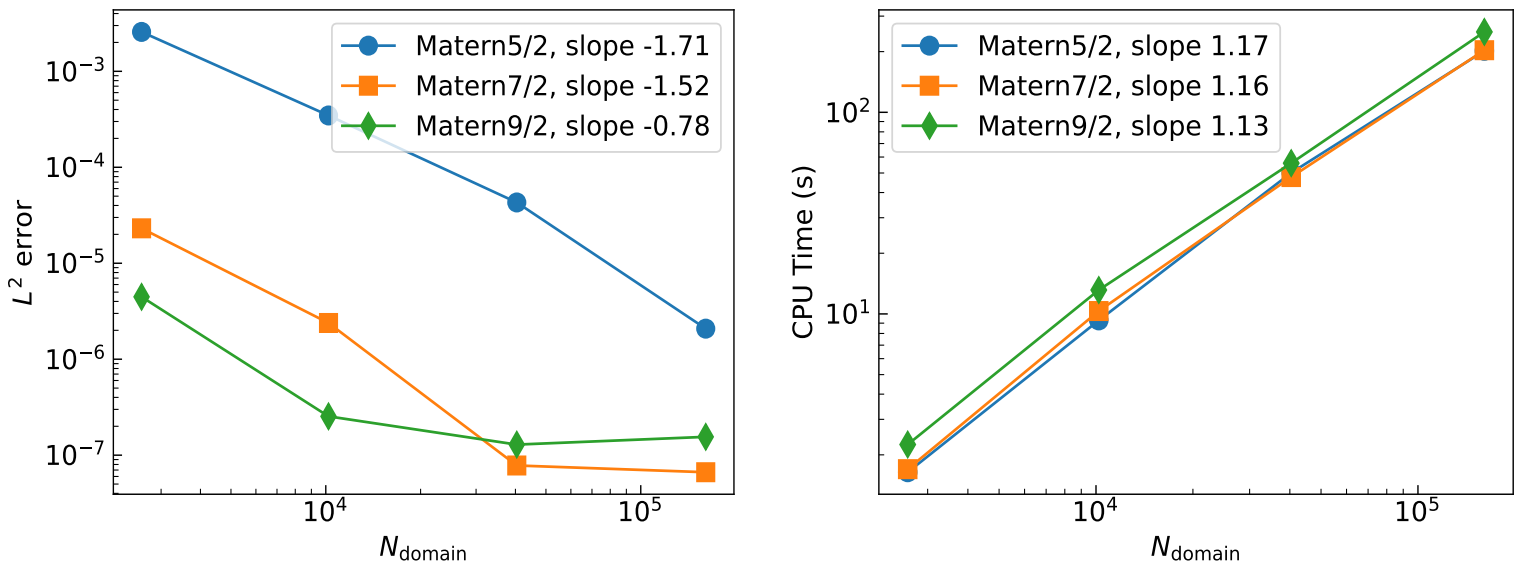


Figure: Run 3 linearization steps with initialization as a zero function. Accuracy floor due to finite $\rho = 4.0$

See other examples of Burgers, Monge-Ampère in the paper

[Chen, Hosseni, Owhadi, Stuart 2021], [Chen, Owhadi, Schäfer 2023]

Summary

Cholesky for GPs in scientific computing and machine learning

- Low rank approximation with random pivoting: balance exploration and exploitation [Musco, Woodruff 2017], [Chen, Epperly, Tropp, Webber 2022], [Díaz, Epperly, Frangella, Tropp, Webber 2023], [Dong, Chen, Martinsson, Pearce 2023], [Steinerberger 2024], [Epperly, Tropp, Webber 2024], etc.
- Full rank approximation with coarse to fine ordering: sparsity due to screening effects applicable also to kernel derivatives [Schäfer, Sullivan, Owhadi 2021], [Schäfer, Katzfuss, Owhadi 2021], [Chen, Owhadi, Schäfer 2023], [Huan et al. 2023], etc.

Interpretation of Cholesky as conditioning of GPs provides insights

- Small conditional variance $\rightsquigarrow$ near low rank
- Near conditional independence $\rightsquigarrow$ near sparsity