Kernel Packet:

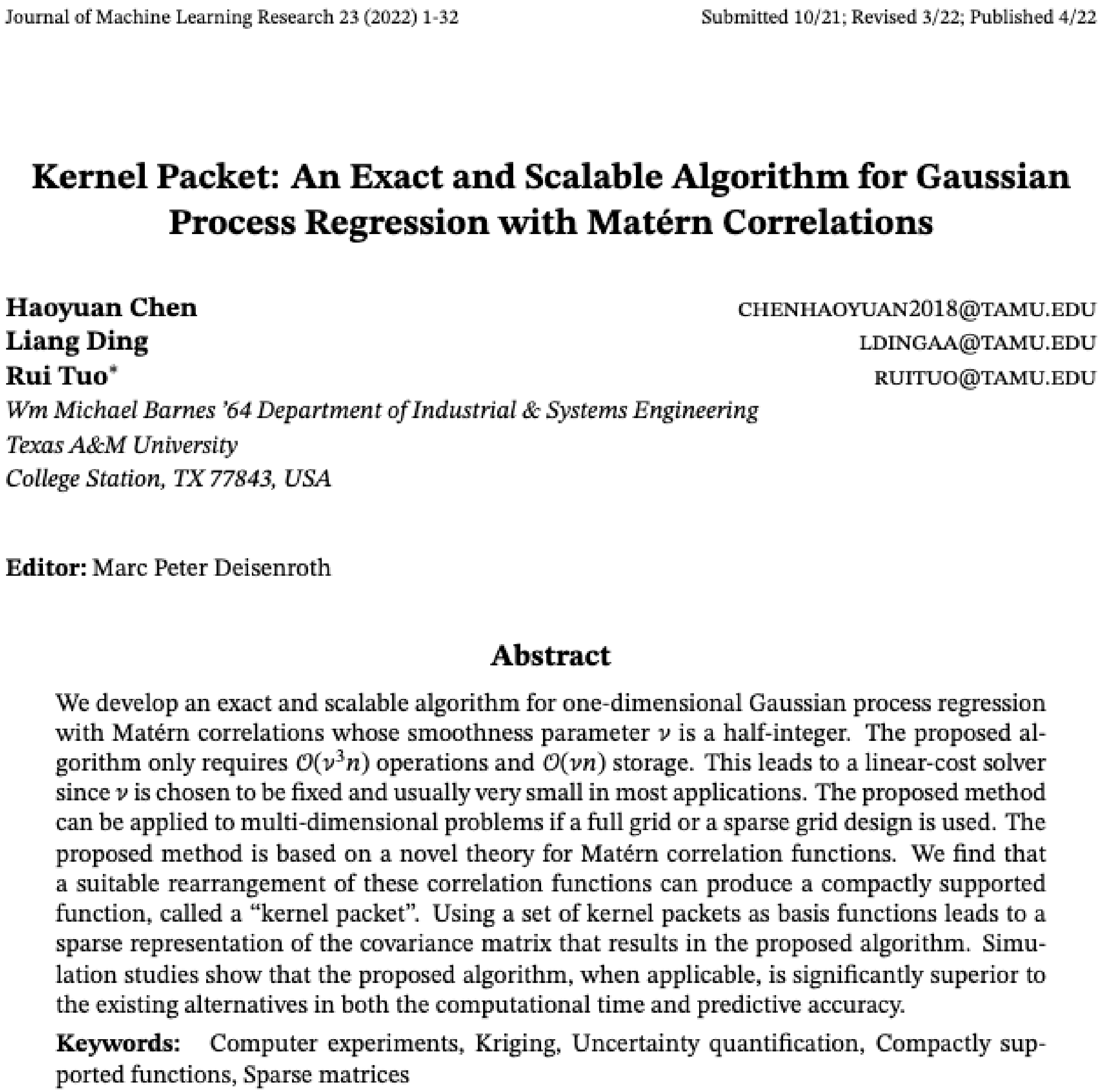
An Exact and Scalable Algorithm for Gaussian Process Regression with Matérn Correlations

Haoyuan Chen

Department of Industrial and Systems Engineering

Texas A&M University

Kernel Packet



Journal of Machine Learning Research 23 (2022) 1-32 Submitted 10/21; Revised 3/22; Published 4/22

Kernel Packet: An Exact and Scalable Algorithm for Gaussian Process Regression with Matérn Correlations

Haoyuan Chen CHENHAOYUAN2018@TAMU.EDU
Liang Ding LDINGAA@TAMU.EDU
Rui Tuo[*] RUITUO@TAMU.EDU
Wm Michael Barnes ’64 Department of Industrial & Systems Engineering
Texas A&M University
College Station, TX 77843, USA

Editor: Marc Peter Deisenroth

Abstract

We develop an exact and scalable algorithm for one-dimensional Gaussian process regression with Matérn correlations whose smoothness parameter ν is a half-integer. The proposed algorithm only requires $\mathcal{O}(\nu^3 n)$ operations and $\mathcal{O}(\nu n)$ storage. This leads to a linear-cost solver since ν is chosen to be fixed and usually very small in most applications. The proposed method can be applied to multi-dimensional problems if a full grid or a sparse grid design is used. The proposed method is based on a novel theory for Matérn correlation functions. We find that a suitable rearrangement of these correlation functions can produce a compactly supported function, called a “kernel packet”. Using a set of kernel packets as basis functions leads to a sparse representation of the covariance matrix that results in the proposed algorithm. Simulation studies show that the proposed algorithm, when applicable, is significantly superior to the existing alternatives in both the computational time and predictive accuracy.

Keywords: Computer experiments, Kriging, Uncertainty quantification, Compactly supported functions, Sparse matrices

Haoyuan Chen, Liang Ding, and Rui Tuo (2022).
Kernel Packet: An Exact and Scalable Algorithm for Gaussian Process Regression with Matérn Correlations.
Journal of Machine Learning Research (JMLR), 23(127), 1-32.

PDF: https://jmlr.org/papers/volume23/21-1232/21-1232.pdf
Code: https://github.com/hchen19/kernel_packet

Outline

- Introduction
  - Gaussian processes (GPs)
  - Main challenge
- Kernel Packets (KPs)
  - Main idea
  - Definition of KPs
  - Construct KPs
  - Existence of KPs
  - Matrix factorization for KPs
  - KPs for GP regression
  - Experiments
  - Conclusions

- Introduction
  - Gaussian processes (GPs)
  - Main challenge

Gaussian Processes

- A *Gaussian process* (GP) is a set of random variables such that every finite collection of those random variables has a multivariate normal distribution. A GP is defined by a mean function $\mu(\cdot)$ and a covariance function (kernel function) $K(\cdot,\cdot)$, denoted by $\mathcal{GP}(\mu(\cdot), K(\cdot,\cdot))$

- **Why are the GPs important?**
  - Probabilistic framework for *Uncertainty Quantification*
  - Robust solutions for small datasets
  - Encode prior knowledge, flexible and interpretable

GP Regression

- *Regression* is used to find a function for estimating the relationships between a response and the one or more predictors.

- **Suppose** that $f(\cdot) \sim \mathcal{GP}(\mu(\cdot), K(\cdot,\cdot))$ and we have observed $(y_1, \ldots, y_n)$ on n distinct points $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$, where $y_i = f(\boldsymbol{x}_i) + \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\epsilon^2), i = 1, \ldots, n.$

- The **aim** of the *GP regression* is to predict the output at an untried input $\boldsymbol{x}^*$ by computing the distribution of $f(\boldsymbol{x}^*)$ conditional on $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, which is a normal distribution with the following conditional mean and variance:

$$\mathbb{E}\left[f(\boldsymbol{x}^*)|\boldsymbol{y}\right] = \mu(\boldsymbol{x}^*) + K(\boldsymbol{x}^*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I}_n]^{-1}(\boldsymbol{y} - \mu(\boldsymbol{X})) \qquad (1)$$

$$\mathrm{Var}\left[f(\boldsymbol{x}^*)|\boldsymbol{y}\right] = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - K(\boldsymbol{x}^*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I}_n]^{-1} K(\boldsymbol{X}, \boldsymbol{x}^*) \qquad (2)$$

- Introduction
  - Gaussian processes (GPs)
  - Main challenge

Main Challenge

- Computation:
  - GP regression:
    - Noise-free: inverse of the covariance matrix $[K(\boldsymbol{X}, \boldsymbol{X})]^{-1}$
    - Noisy: inverse of the covariance matrix with noise $[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I}_n]^{-1}$

- Issues:
  - Inversion requires $\mathcal{O}(n^3)$ time, which limits the scalability of GPs when n is large.

- Introduction
- Kernel Packets (KPs)

Main Idea

- Assumption:
  - Consider noise-free and one-dimensional GPs
- Motivation:
  - Note that each entry of $K(\boldsymbol{X}, \boldsymbol{X})$ is an evaluation of function $K(\cdot, \boldsymbol{x}_j)$ for some j. The matrix $K(\boldsymbol{X}, \boldsymbol{X})$ is not sparse because the support of K is the entire real line.
  - The main idea of this work is to find an *exact* representation of K in terms of **sparse matrices**.
  - This exact representation is built in terms of a *change-of-basis* transformation.

Definition of KPs

- Preliminary

Definition. Given a correlation function $K(\cdot,\cdot)$ and input points $x_1 < \cdots < x_k$, a non-zero function $\phi(\cdot)$ is called a Kernel Packet (KP) of degree k, if it admits the representation

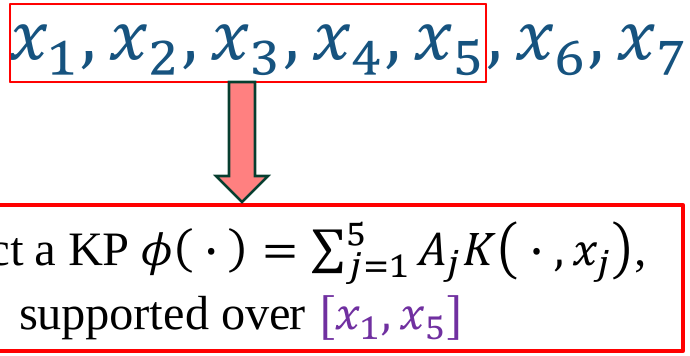
$$\phi(\cdot) = \sum_{j=1}^{k} A_j K(\cdot, x_j),$$

and the support of ϕ is $[x_1, x_k]$.

- Questions:
  - How to construct KPs?
  - Do KPs exist?

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

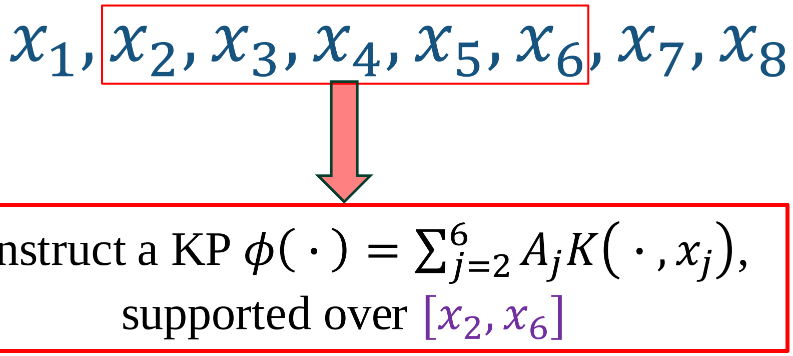


$\boxed{x_1, x_2, x_3, x_4, x_5}, x_6, x_7, x_8, x_9, x_{10}$

Construct a KP $\phi(\cdot) = \sum_{j=1}^{5} A_j K(\cdot, x_j)$,
supported over $[x_1, x_5]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

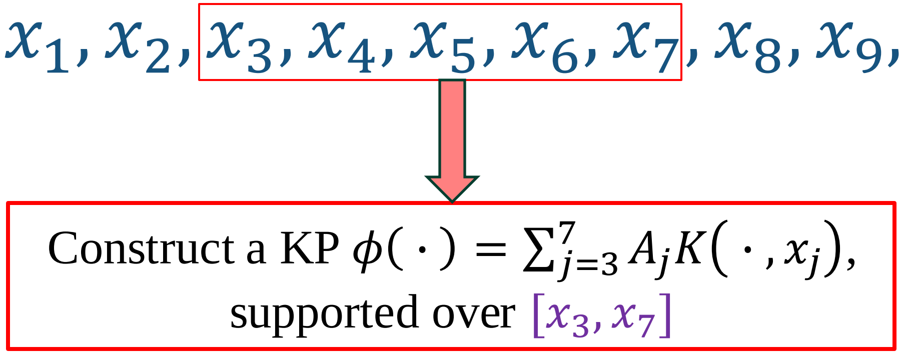


$$x_1, \boxed{x_2, x_3, x_4, x_5, x_6}, x_7, x_8, x_9, x_{10}$$

Construct a KP $\phi(\cdot) = \sum_{j=2}^{6} A_j K(\cdot, x_j)$,
supported over $[x_2, x_6]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

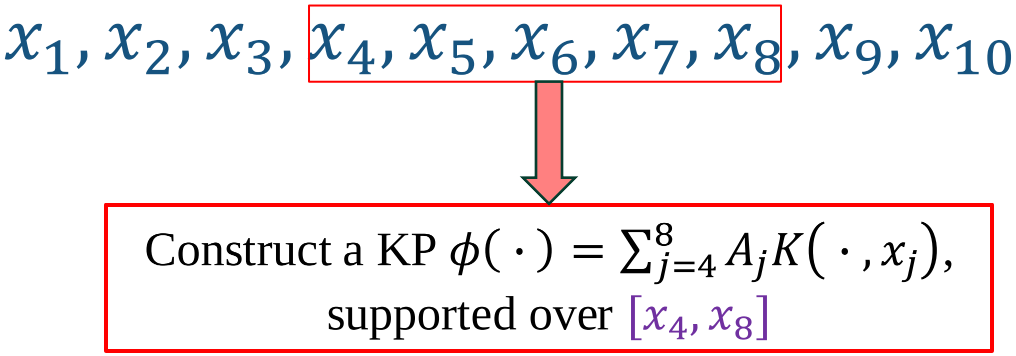


$$x_1, x_2, \boxed{x_3, x_4, x_5, x_6, x_7}, x_8, x_9, x_{10}$$

Construct a KP $\phi(\cdot) = \sum_{j=3}^{7} A_j K(\cdot, x_j)$,
supported over $[x_3, x_7]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

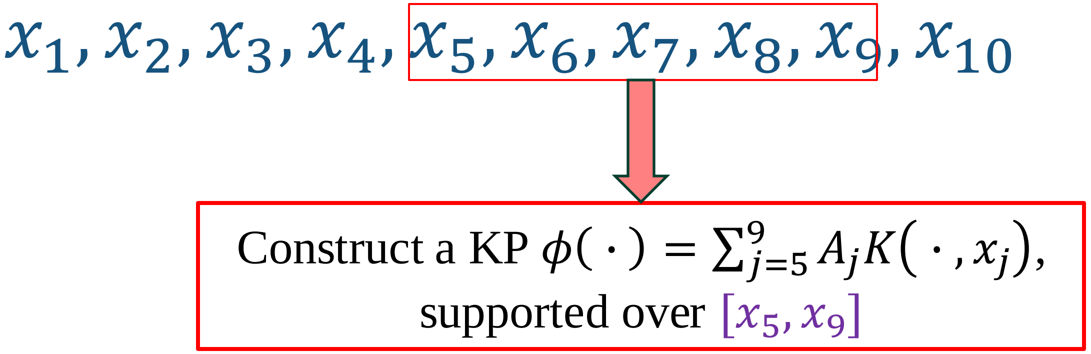


$$x_1, x_2, x_3, \boxed{x_4, x_5, x_6, x_7, x_8}, x_9, x_{10}$$

Construct a KP $\phi(\,\cdot\,) = \sum_{j=4}^{8} A_j K(\,\cdot\,, x_j)$,
supported over $[x_4, x_8]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

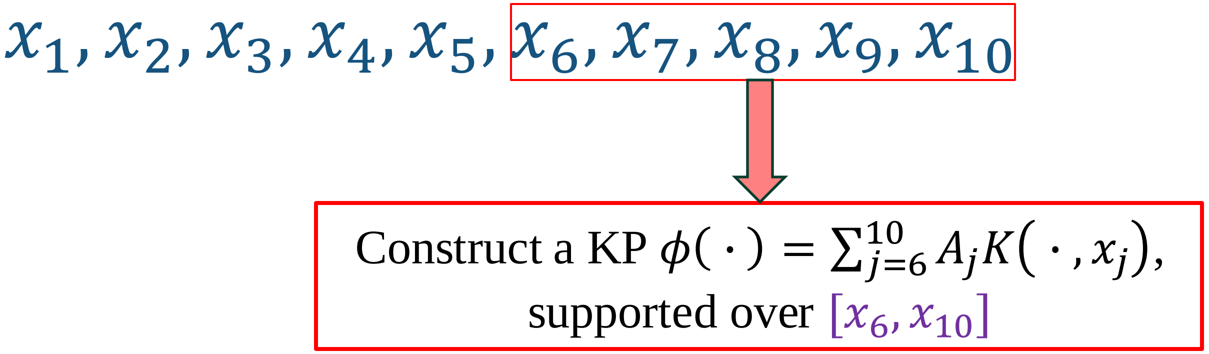


$$x_1, x_2, x_3, x_4, \boxed{x_5, x_6, x_7, x_8, x_9}, x_{10}$$

Construct a KP $\phi(\cdot) = \sum_{j=5}^{9} A_j K(\cdot, x_j)$, supported over $[x_5, x_9]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$



$$x_1, x_2, x_3, x_4, x_5, \boxed{x_6, x_7, x_8, x_9, x_{10}}$$

Construct a KP $\phi(\cdot) = \sum_{j=6}^{10} A_j K(\cdot, x_j)$,
supported over $[x_6, x_{10}]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$$

❖ Issue:

- We now have 6 KPs of degree $k = 5$, however, to construct a basis of $\text{span}\{K(\cdot, x_i)\}_{i=1}^{10}$, we should have 10 KPs.

❖ Question:

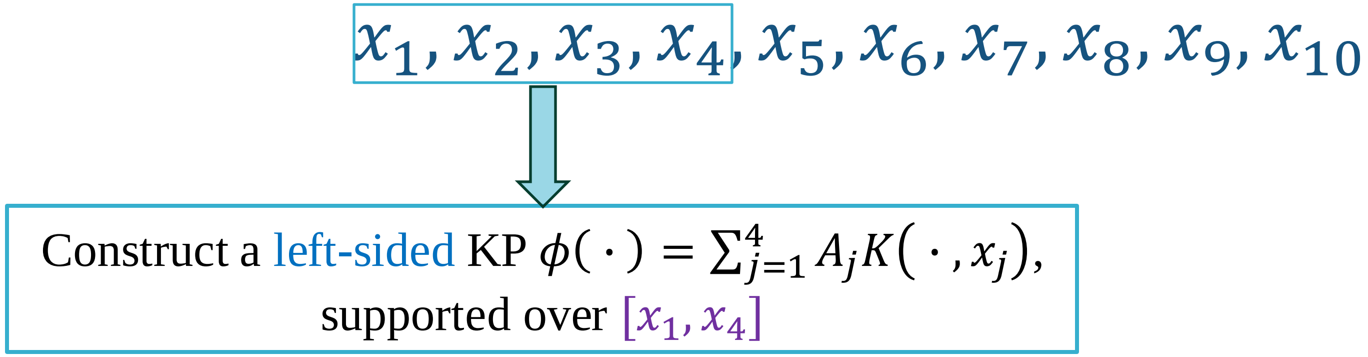
- How to construct another 4 KPs, which are linear independent of the existing 6 KPs.

❖ Solution:

- Construct 4 boundary KPs of degree less than $k = 5$:
  2 left-sided KPs, 2 right-sided KPs

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

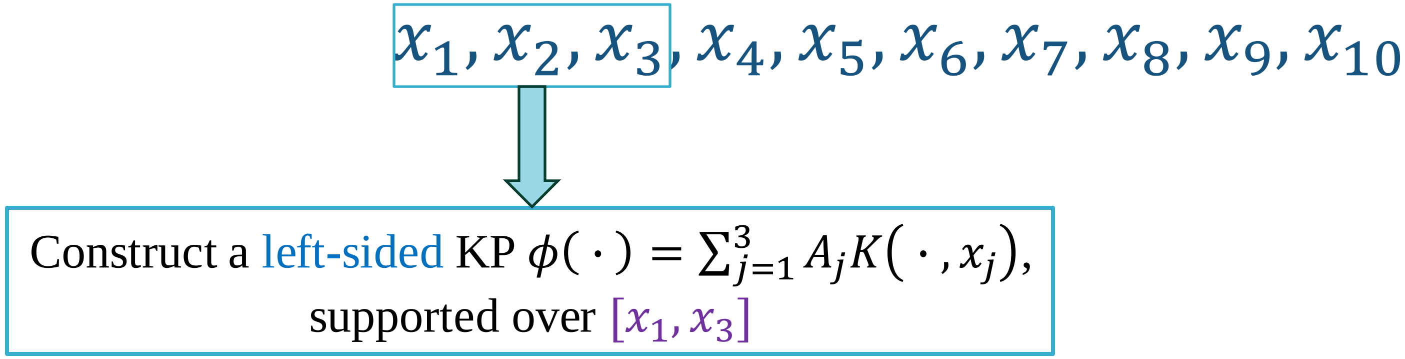


$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$

Construct a left-sided KP $\phi(\cdot) = \sum_{j=1}^{4} A_j K(\cdot, x_j)$,
supported over $[x_1, x_4]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

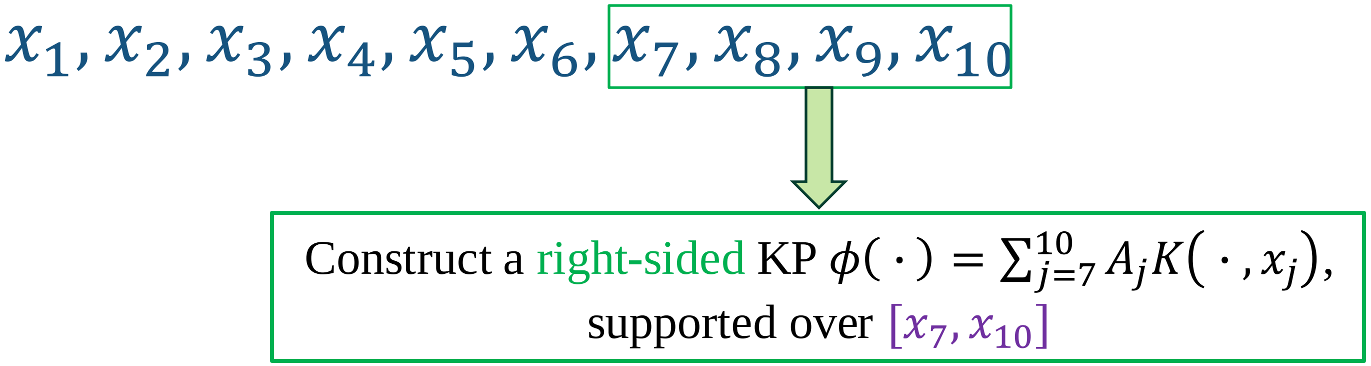


$\boxed{x_1, x_2, x_3}, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$

Construct a left-sided KP $\phi(\cdot) = \sum_{j=1}^{3} A_j K(\cdot, x_j)$, supported over $[x_1, x_3]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

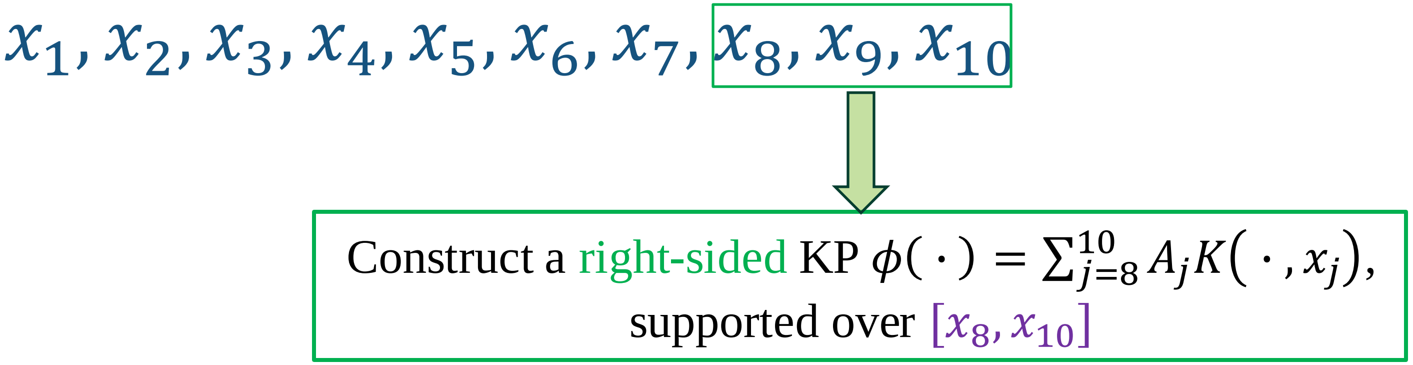


$$x_1, x_2, x_3, x_4, x_5, x_6, \boxed{x_7, x_8, x_9, x_{10}}$$

Construct a right-sided KP $\phi(\cdot) = \sum_{j=7}^{10} A_j K(\cdot, x_j)$,
supported over $[x_7, x_{10}]$

Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$



$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, \boxed{x_8, x_9, x_{10}}$$

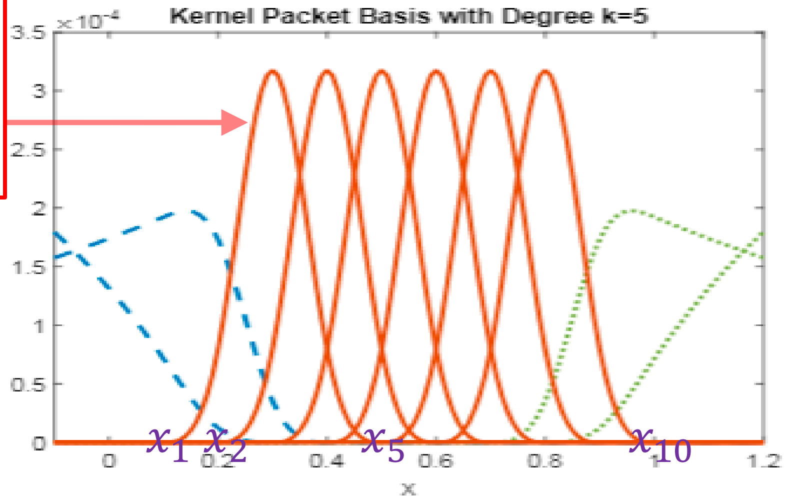
Construct a right-sided KP $\phi(\cdot) = \sum_{j=8}^{10} A_j K(\cdot, x_j)$, supported over $[x_8, x_{10}]$

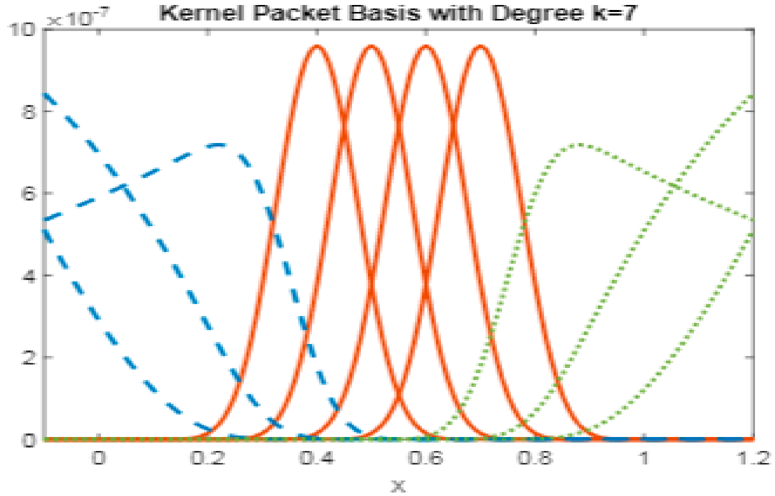
Construct KPs

- How can KPs help with the computation?
- Toy Example: Construct KPs from a moving window of the data with $k = 5$

Construct a KP $\phi(\cdot) = \sum_{j=1}^{5} A_j K(\cdot, x_j)$, supported over $[x_1, x_5]$

$\boxed{x_1, x_2, x_3, x_4, x_5}, x_6, x_7, x_8, x_9, x_{10}$

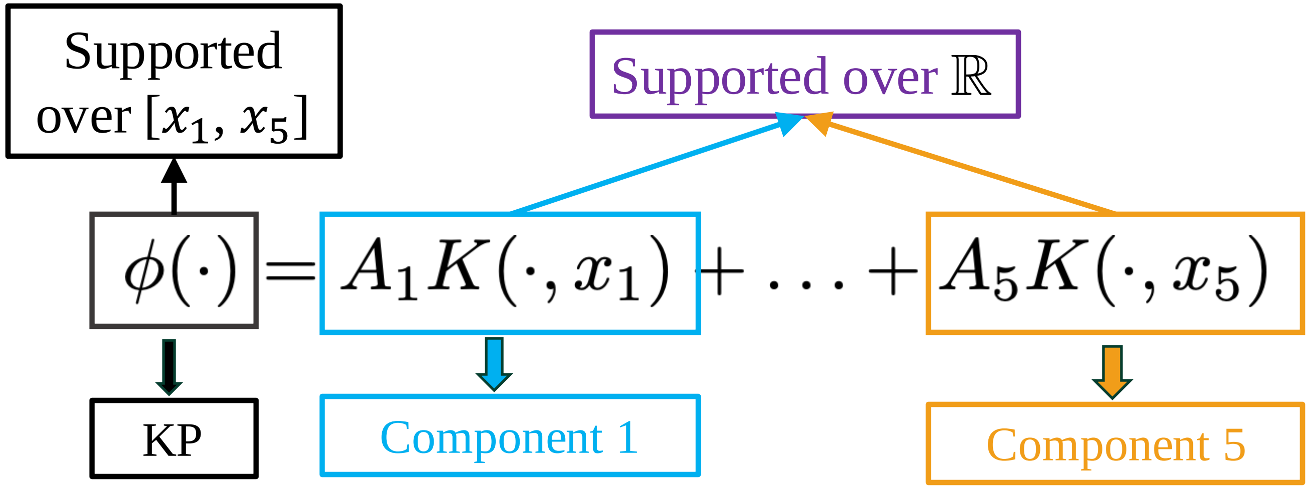




- Given 1D input data $x_1 < \cdots < x_n$, we can find another basis of $\text{span}\{K(\cdot, x_i)\}$, called the **KP basis**, which consists of n KPs for n inputs
- KP basis consists of $(n - k + 1)$ Intermediate KPs (in red), $(k - 1)/2$ Left-sided KPs (in blue), and $(k - 1)/2$ Right-sided KPs (in green)

Existence of KPs

- KPs do exist for Matérn correlations with half-integer smoothness
  - Matérn-ν correlation has a KP of degree $2\nu + 2$.
  - $2\nu + 2$ is proved to be the lowest possible degree.



Supported over $[x_1, x_5]$

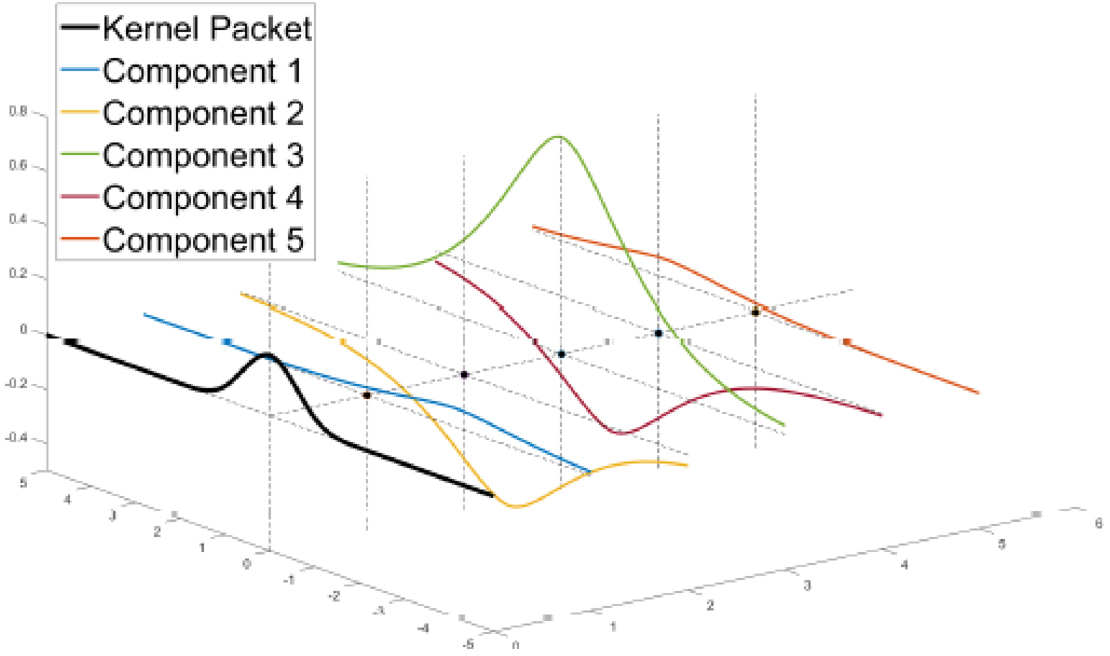
Supported over $\mathbb{R}$

$$\phi(\cdot) = A_1 K(\cdot, x_1) + \ldots + A_5 K(\cdot, x_5)$$

KP

Component 1

Component 5



Note: We shall denote $k := 2\nu + 2$ and use k to parametrize the Matérn correlation.

A Matérn 3/2 correlation can generate KP of degree 5.

Existence of KPs

Paley-Wiener Theorem

If f is continuous and of moderate decrease, $\hat{f}$ is the Fourier transform of f. Then,

f is an entire function of exponential type M, i.e. f is holomorphic on the whole complex plane and there exists some constant $C > 0$ such that $|f(z)| \leq Ce^{M|z|}$

if and only if

$\hat{f}$ is supported in $[-M, M]$.

Consider a small subset of one-dimensional training inputs $\boldsymbol{x}_{1:k} := [x_1, \ldots, x_k]^\top$ and a Matérn kernel $K(\cdot,\cdot)$ of smoothness υ and lengthscale ω

$$\phi_{\boldsymbol{x}_{1:k}}(\cdot) = \sum_{j=1}^{k} A_j K(\cdot, x_j)$$

$\Downarrow$ Inverse Fourier Transform

$$\tilde{\phi}_{\boldsymbol{x}_{1:k}}(z) \propto \left[\sum_{j=1}^{k} A_j \exp\{ix_j z\}\right](2\nu/\omega^2 + z^2)^{-(k-1)/2} := \gamma(z)(c^2 + z^2)^{-(k-1)/2}$$

Existence of KPs

Note: $(c^2+z^2)^{-(k-1)/2}$ has poles at $z=\pm ci$, each with multiplicity $(k-1)/2$

$\gamma(z)(c^2+z^2)^{-(k-1)/2}$ is an entire function

$$\begin{aligned} \gamma(ci)&=0, \quad \gamma(-ci)=0\\ \gamma'(ci)&=0, \quad \gamma'(-ci)=0\\ &\cdots\\ \gamma^{\left(\frac{k-3}{2}\right)}(ci)&=0, \quad \gamma^{\left(\frac{k-3}{2}\right)}(-ci)=0 \end{aligned}$$

$$\sum_{j=1}^{k} A_j x_j^l \exp\{\delta c x_j\}=0, \quad l=0,\ldots,(k-3)/2,\ \delta=\pm 1$$

The null space of this $(k-1)\times k$ linear system is one-dimensional if $x_1,\ldots,x_k$ are distinct

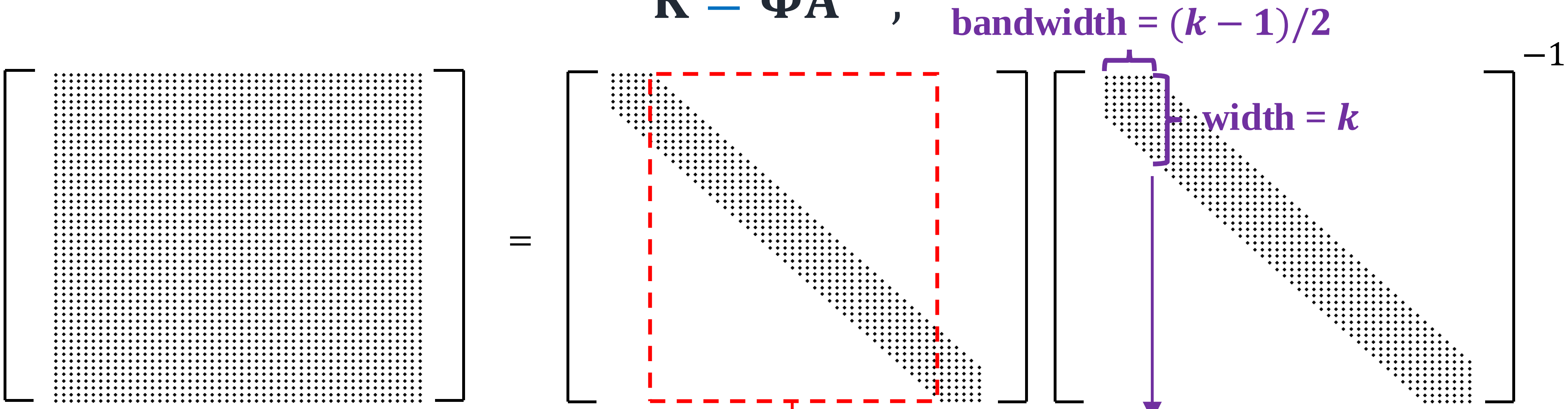
$\Downarrow$

For one small subset of training inputs $\boldsymbol{x}_{1:k}$, we can obtain $\{A_j\}_{j=1}^{k}$ by solving the corresponding linear system

Matrix Factorization for KPs

- In a matrix form, the correlation matrix admit the following factorization according to KP definition:

$$\mathbf{K} = \mathbf{\Phi}\mathbf{A}^{-1},$$



Intermediate KPs
Each column corresponds to a KP $\phi(\cdot)$ evaluated over the inputs $\{x_i\}_{i=1}^{n}$

A column consists of the coefficients $\{A_i\}_{i=1}^{k}$ for a KP $\phi(\cdot) = \sum_{i=1}^{k} A_i K(\cdot, x_i)$

KPs for GP Regression

- Suppose K is a 1D Matérn correlation with a half-integer smoothness, e.g., $\nu = 1.5, 2.5, 3.5, \ldots$
- We prove that for any input points in 1D sorted in an increasing order, the correlation matrix admit the factorization:

$$\mathbf{K} = \mathbf{\Phi}\mathbf{A}^{-1},$$

where

  - $\mathbf{\Phi}$ is a banded matrix with bandwidth $\nu - 1/2$.
  - $\mathbf{A}$ is a banded matrix with bandwidth $\nu + 1/2$.

The $n \times n$ matrix $\mathbf{A}$ is computed by solving n $(k-1) \times k$ linear systems, each column of $\mathbf{A}$ corresponds to the coefficients solved by one linear system

The $n \times n$ matrix $\mathbf{\Phi}$ is computed by $\mathbf{\Phi} = \mathbf{K}\mathbf{A}$

- Matrix inversion computed as

$$\mathbf{K}^{-1}\mathbf{y} = \mathbf{A}(\mathbf{\Phi}^{-1}\mathbf{y}),$$

which takes only $O(n)$ time and $O(n)$ storage (assuming ν is fixed)!

Experiments for KPs

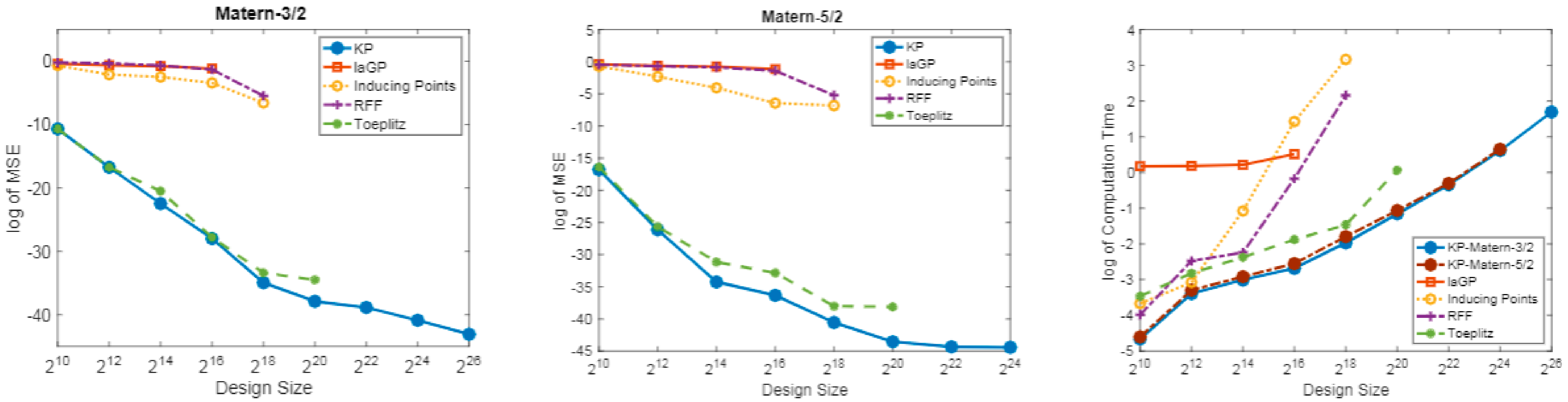


Figure 3: Logarithm of MSE for predictions with Matérn-3/2 correlation function (left) and Matérn-5/2 correlation function (middle) and logarithm of averaged computational time (right). The laGP uses the Gaussian covariance family in both the left and the middle figure. No results are shown for the cases when a runtime error occurs or the prediction error ceases to improve.

Conclusions for KPs

- propose a *rapid* and *exact* algorithm for one-dimensional Gaussian process regression under Matérn correlations with half-integer smoothness.
- The proposed algorithm only requires $O(n)$ time and $O(n)$ storage
- The proposed method can be applied to some multi-dimensional problems by using tensor product techniques, including grid and sparse grid designs, and their generalizations.

- Q & A

Thank you!

Any Questions?