Random Fourier Features for Kernel Ridge Regression

Michael Kapralov[1]

[1] EPFL

(Joint work with H. Avron, C. Musco, C. Musco, A. Velingker and A. Zandieh)

Scalable machine learning algorithms with provable guarantees

In this talk: towards scalable numerical linear algebra in kernel spaces with provable guarantees

Linear regression

Input:

- a sequence of d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

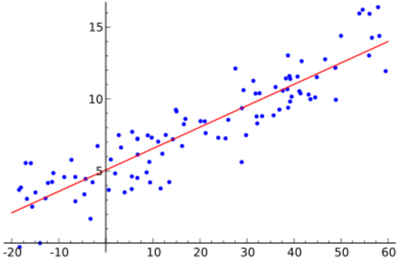
Output: linear approximation to f

Linear regression

Input:

- a sequence of d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

Output: linear approximation to f

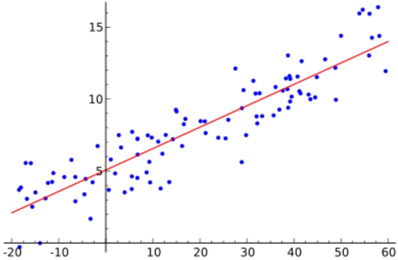


Linear regression

Input:

- a sequence of d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

Output: linear approximation to f



Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^d} \sum_{j=1}^{n} |x_j \alpha - y_j|^2 + \lambda ||\alpha||_2^2$$

Kernel ridge regression

Input:

- a sequence of d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

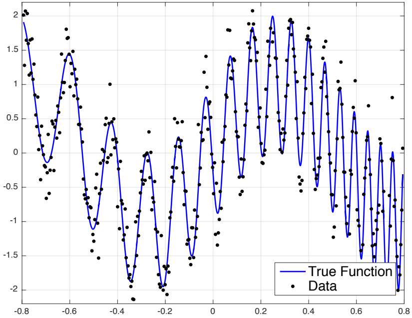
Output: approximation from class of 'smooth' functions on $\mathbb{R}^d$

Kernel ridge regression

Input:

- a sequence of d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

Output: approximation from class of 'smooth' functions on $\mathbb{R}^d$



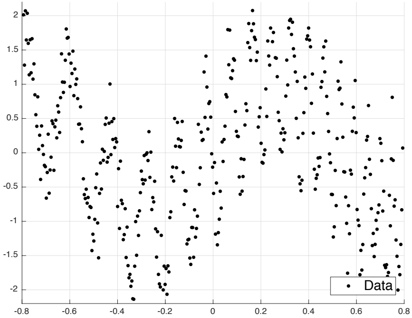
Choose an embedding into a high dimensional feature space

$$\Psi : \mathbb{R} \to \mathbb{R}^D$$

Dimension D may be infinite (e.g. Gaussian kernel).

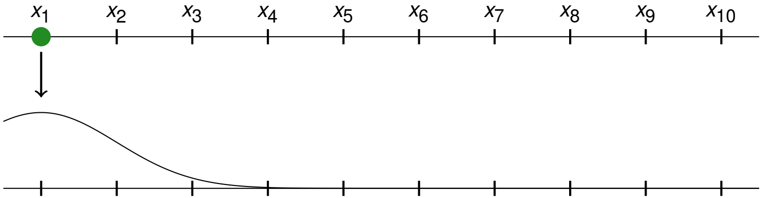
Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^D} \sum_{j=1}^{n} |\Psi(x_j)\alpha - y_j|^2 + \lambda ||\alpha||_2^2$$



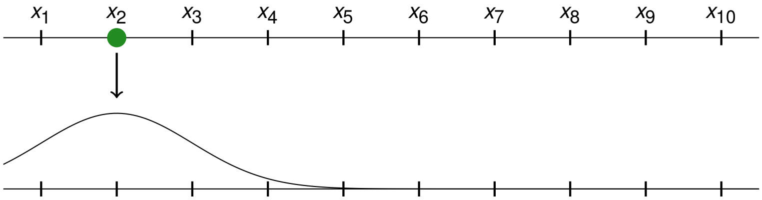
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



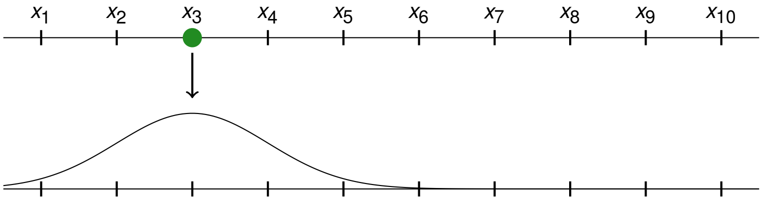
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



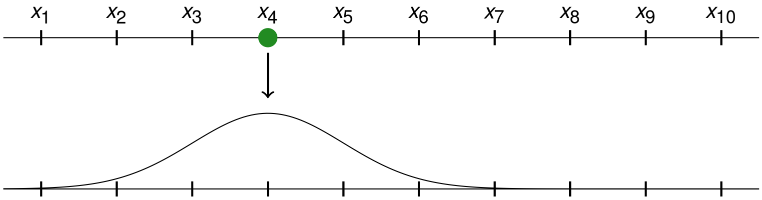
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



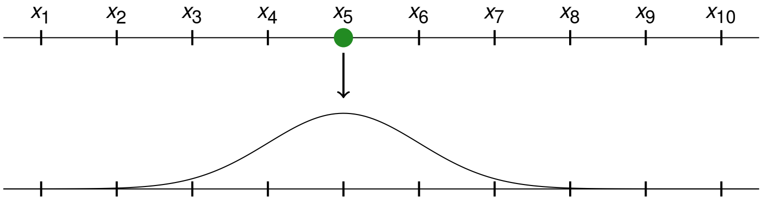
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



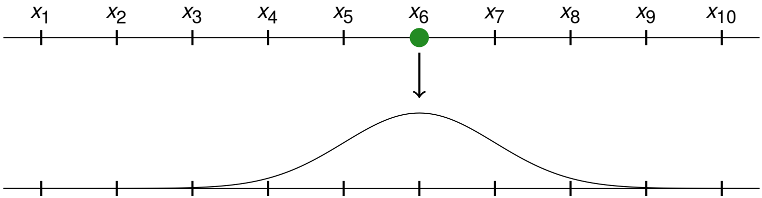
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



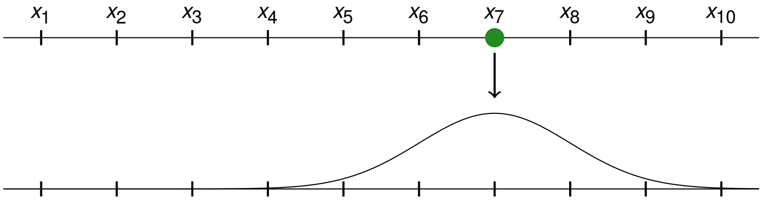
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



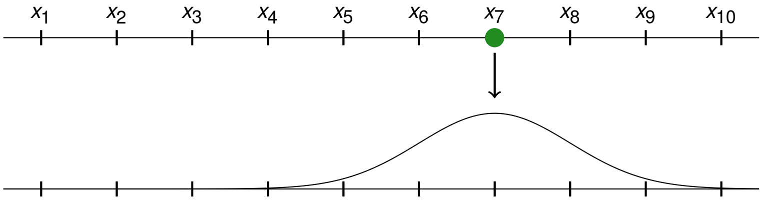
Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$



Choose an embedding into a high dimensional feature space

$$\Psi : x \to \frac{1}{(2\pi)^{1/4}} e^{-(\cdot - x)^2/4}$$

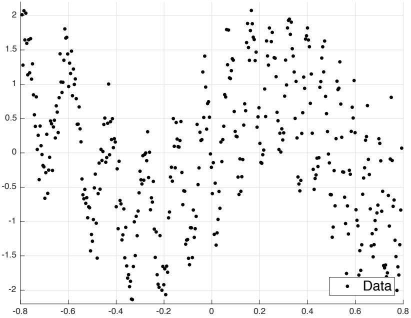


Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^D} \sum_{j=1}^{n} |\Psi(x_j)\alpha - y_j|^2 + \lambda ||\alpha||_2^2$$

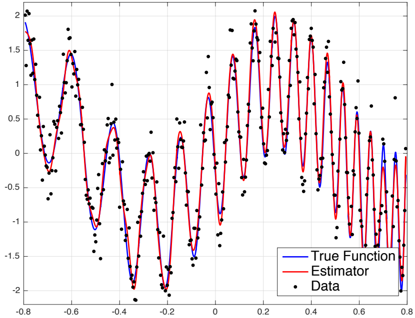
Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^D} \sum_{j=1}^{n} |\Psi(x_j)\alpha - y_j|^2 + \lambda ||\alpha||_2^2$$



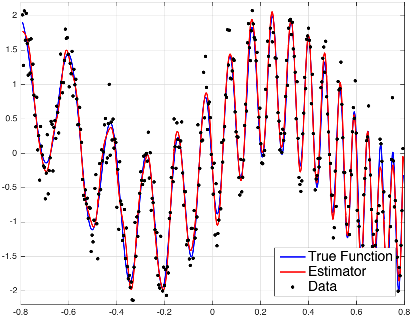
Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^D} \sum_{j=1}^{n} |\Psi(x_j)\alpha - y_j|^2 + \lambda ||\alpha||_2^2$$



Solve least squares problem:

$$\min_{\alpha \in \mathbb{R}^D} \sum_{j=1}^{n} |\Psi(x_j)\alpha - y_j|^2 + \lambda ||\alpha||_2^2$$



After algebraic manipulations

$$\boxed{\alpha^* = \Psi^T (K + \lambda I)^{-1} y}$$

Kernel ridge regression

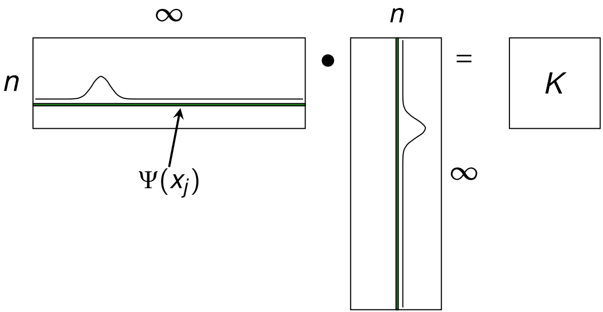
Main computational effort:

$$\boxed{(K + \lambda I)^{-1} y}$$

Kernel ridge regression

Main computational effort:

$$\boxed{(K + \lambda I)^{-1}\, y}$$



The (i,j)-th entry of Gaussian kernel matrix K is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

How quickly can we compute $(K + \lambda I)^{-1} y$?

The (i,j)-th entry of Gaussian kernel matrix K is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

How quickly can we compute $(K+\lambda I)^{-1}y$?

The (i,j)-th entry of Gaussian kernel matrix K is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

n^3 (or n^ω) in full generality...

$\Omega(n^2)$ time needed when $\lambda = 0$ assuming SETH
Backurs-Indyk-Schmidt (NIPS'17)

How quickly can we compute $(K+\lambda I)^{-1}y$?

The (i,j)-th entry of Gaussian kernel matrix K is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

n^3 (or n^ω) in full generality...

$\Omega(n^2)$ time needed when $\lambda = 0$ assuming SETH
Backurs-Indyk-Schmidt (NIPS'17)

In practice: find $Z \in \mathbb{R}^{n\times s}, s \ll n$ such that

$$K \approx ZZ^T$$

and use $ZZ^T + \lambda I$ as a proxy for $K + \lambda I$!

How quickly can we compute $(K+\lambda I)^{-1} y$?

The (i,j)-th entry of Gaussian kernel matrix K is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

n^3 (or n^ω) in full generality...

$\Omega(n^2)$ time needed when $\lambda = 0$ assuming SETH
Backurs-Indyk-Schmidt (NIPS'17)

In practice: find $Z \in \mathbb{R}^{n\times s}, s \ll n$ such that

$$K \approx ZZ^T$$

and use $ZZ^T + \lambda I$ as a proxy for $K + \lambda I$!

$\boxed{\text{Can compute } (ZZ^T + \lambda I)^{-1} y \text{ in } O(ns^2) \text{ time and } O(ns) \text{ space!}}$

Fourier Features

Theorem (Bochner's Theorem)

A normalized continuous function $k : \mathbb{R} \to \mathbb{R}$ *is a shift-invariant kernel if and only if its Fourier transform* $\widehat{k}$ *is a measure.*

Fourier Features

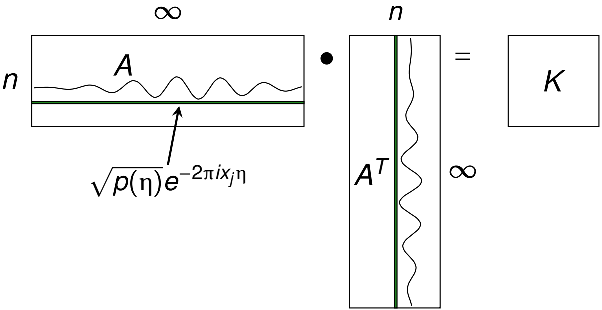
Theorem (Bochner's Theorem)

A normalized continuous function $k : \mathbb{R} \to \mathbb{R}$ is a shift-invariant kernel if and only if its Fourier transform $\widehat{k}$ is a measure.

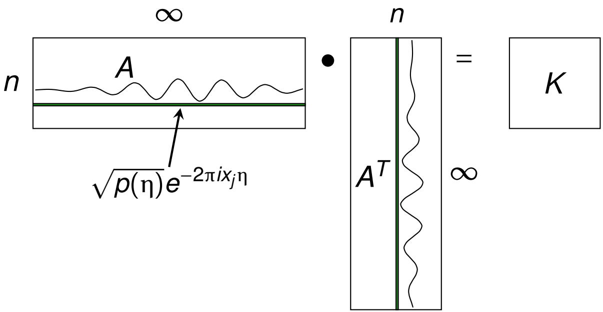
Let $p(\eta) := \widehat{k}(\eta)$. Then for every x_a, x_b

$$
\begin{aligned}
K_{ab} = k(x_a - x_b) &= \int_{\mathbb{R}} \widehat{k}(\eta) e^{-2\pi i (x_a - x_b)\eta} d\eta \\
&= \int_{\mathbb{R}} e^{-2\pi i (x_a - x_b)\eta} p(\eta) d\eta \\
&= \mathbf{E}_{\eta \sim p(\eta)} \left[e^{-2\pi i (x_a - x_b)\eta} \right]
\end{aligned}
$$

Fourier Features



Fourier Features



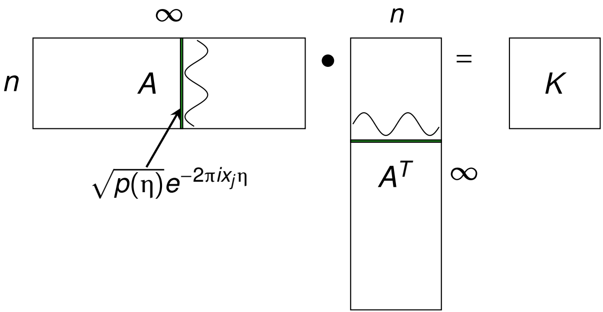
Rahimi-Recht'2007: fix s, sample i.i.d. $\eta_1, \ldots, \eta_s \sim p(\eta)$

Let j-th row of Z be

$$Z_{j,k} := \frac{1}{\sqrt{s}} e^{-2\pi i x_j \eta_k} \quad \text{(samples of pure frequency } x_j\text{)}$$

and use ZZ^T as a proxy for K!

Fourier Features: sampling columns of Fourier factorization of K



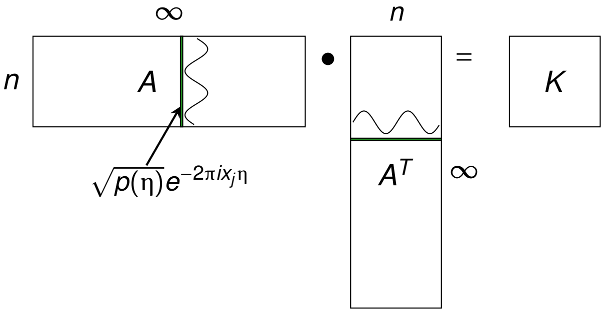
Rahimi-Recht'2007: fix s, sample i.i.d. $\eta_1, \ldots, \eta_s \sim p(\eta)$

Let j-th row of Z be

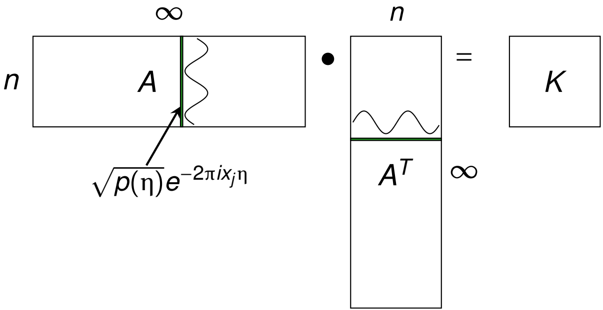
$$Z_{j,k} := \frac{1}{\sqrt{s}} e^{-2\pi i x_j \eta_k} \quad \text{(samples of pure frequency } x_j\text{)}$$

and use ZZ^T as a proxy for K!

Fourier Features: sampling columns of Fourier factorization of K



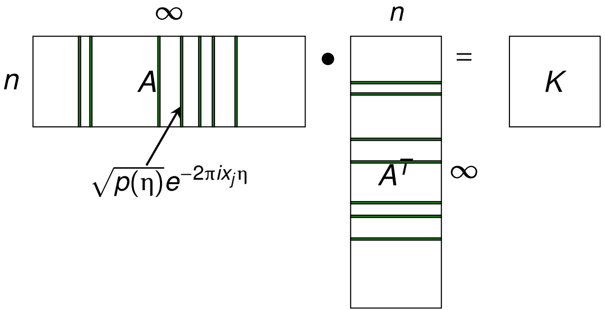
Fourier Features: sampling columns of Fourier factorization of K



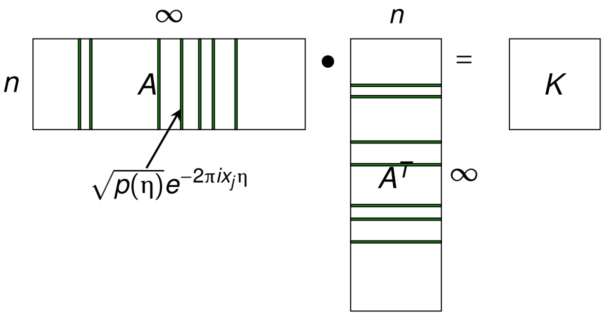
$\boxed{\text{Column } \eta \text{ has } \ell_2^2 \text{ norm } n \cdot p(\eta)!}$

Fourier features = sampling columns of A with probability proportional to column norms squared!

Fourier Features: sampling columns of Fourier factorization of K



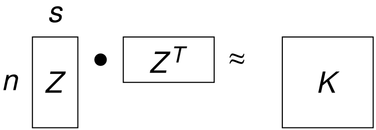
Fourier Features: sampling columns of Fourier factorization of K



$\boxed{\text{Column } \eta \text{ has } \ell_2^2 \text{ norm } n \cdot p(\eta)!}$

Fourier features = sampling columns of A with probability proportional to column norms squared!

Fourier Features: sampling columns of Fourier factorization of K

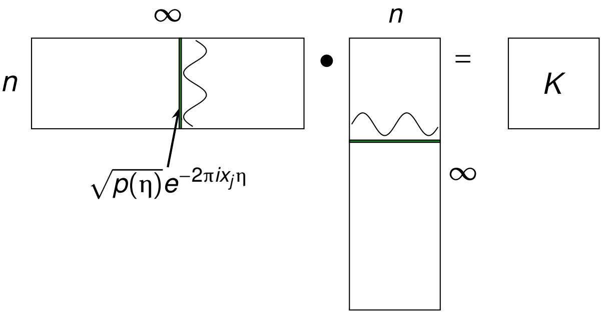


$\boxed{\text{Column } \eta \text{ has } \ell_2^2 \text{ norm } n \cdot p(\eta)!}$

Fourier features = sampling columns of A with probability proportional to column norms squared!

One has $\mathbf{E}[ZZ^T] = K$

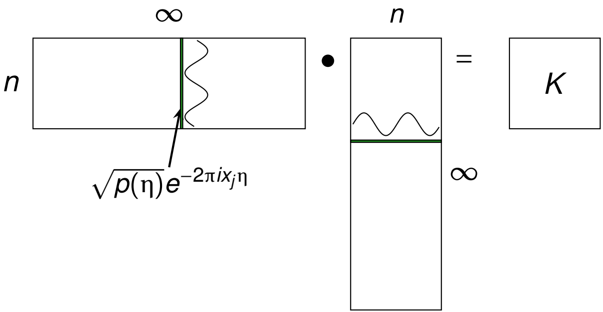
Spectral approximations



Our goal: find $Z \in \mathbb{R}^{n\times s}$, **s ≪ n** such that

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)?$$

Spectral approximations

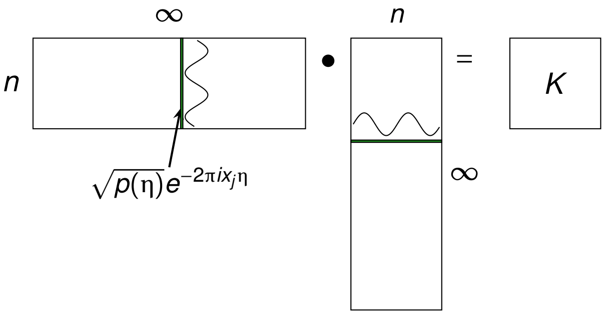


Our goal: find $Z \in \mathbb{R}^{n \times s}$, **s ≪ n** such that

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)?$$

Subspace embeddings for kernel matrices that can be applied implicitly to points $x_1, \ldots, x_n \in \mathbb{R}^d$?

Spectral approximations



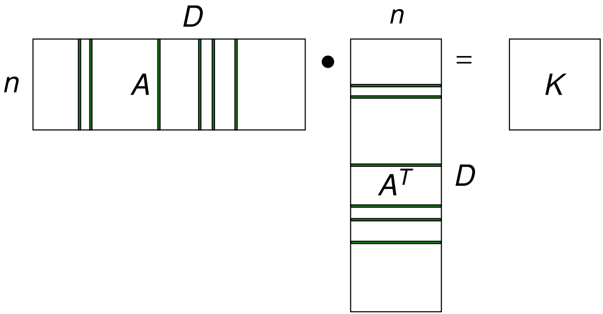
Our goal: find $Z \in \mathbb{R}^{n \times s}$, **s ≪ n** such that

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)?$$

Subspace embeddings for kernel matrices that can be applied implicitly to points $x_1, \ldots, x_n \in \mathbb{R}^d$?

Known for the polynomial kernel only: Avron et al., NIPS'2014 via TensorSketch

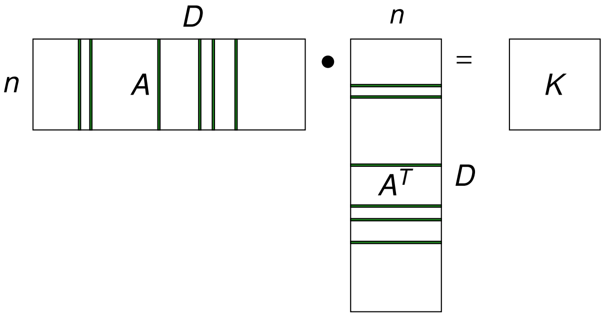
Spectral approximation via column sampling



For each $j = 1, \ldots, D$ compute sampling probability $\tau(j)$

Sample s columns independently from distribution τ, include j in Z with weight $\frac{1}{\sqrt{s \cdot \tau(j)}}$ if sampled.

Spectral approximation via column sampling



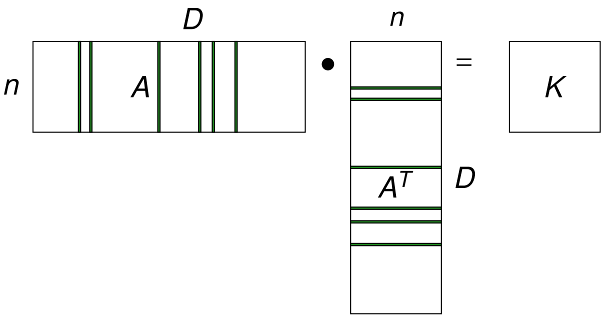
For each $j = 1, \ldots, D$ compute sampling probability $\tau(j)$

Sample s columns independently from distribution τ, include j in Z with weight $\frac{1}{\sqrt{s \cdot \tau(j)}}$ if sampled.

That way

$$\mathbf{E}[ZZ^T] = K$$

Spectral approximation via column sampling



For each $j = 1, \ldots, D$ compute sampling probability $\tau(j)$

Sample s columns independently from distribution τ, include j in Z with weight $\frac{1}{\sqrt{s \cdot \tau(j)}}$ if sampled.

That way

$$\mathbf{E}[ZZ^T] = K$$

Choose τ to ensure ZZ^T spectrally close to K whp?

Ridge leverage scores

Define λ-ridge leverage scores by

$$\tau_\lambda(j) := a_j^T (K + \lambda I)^+ a_j$$

Ridge leverage scores

Define λ-ridge leverage scores by

$$\tau_\lambda(j) := a_j^T (K + \lambda I)^+ a_j$$

The number of samples required $\approx$ statistical dimension of K

$$s_\lambda(K) := \mathrm{tr}((K + \lambda I)^+ K) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

Ridge leverage scores

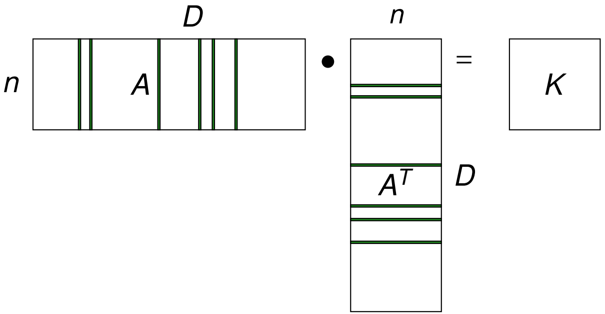
Define λ-ridge leverage scores by

$$\tau_\lambda(j) := a_j^T (K + \lambda I)^+ a_j$$

The number of samples required $\approx$ statistical dimension of K

$$s_\lambda(K) := \mathrm{tr}((K + \lambda I)^+ K) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

Statistical dimension$\approx$ # eigenvalues above λ
+(sum of eigenvalues below λ)$/\lambda$



Theorem (Folklore)

Suppose that

- *for each* $i = 1, \ldots, s$ *one has* $Z_i \sim a_j$ *with probability* $\sim \tau_\lambda(j)$ *independently;*
- $s = O(\varepsilon^{-2} s_\lambda \log s_\lambda)$.

Then

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)$$

with high probability.

Q1: does Fourier Features provide spectral guarantees with $\widetilde{O}(s_\lambda)$ samples?

Q1: does Fourier Features provide spectral guarantees with $\widetilde{O}(s_\lambda)$ samples?

This paper: NO, not even in dimension $d = 1$

Q1: does Fourier Features provide spectral guarantees with $\widetilde{O}(s_\lambda)$ samples?

This paper: NO, not even in dimension $d = 1$

Q1': how many samples are necessary and sufficient for spectral guarantees?

This paper: (essentially) tight bounds

Q1: does Fourier Features provide spectral guarantees with $\widetilde{O}(s_\lambda)$ samples?

This paper: NO, not even in dimension $d = 1$

Q1': how many samples are necessary and sufficient for spectral guarantees?

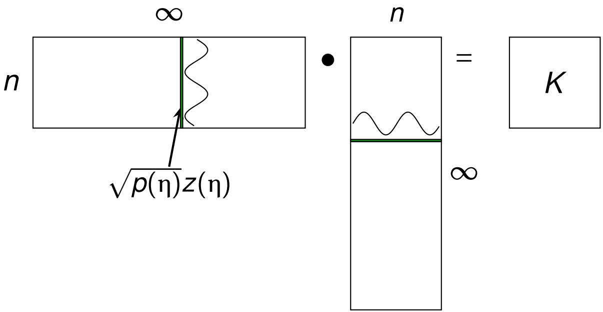
This paper: (essentially) tight bounds

Q2: a better sampling scheme with $\widetilde{O}(s_\lambda)$ samples?

This paper: YES, at least in constant dimensions for bounded datasets

- Leverage score density function
- Primal-dual characterization
- Tight lower bound for Fourier Features

- **Leverage score density function**
- Primal-dual characterization
- Tight lower bound for Fourier Features

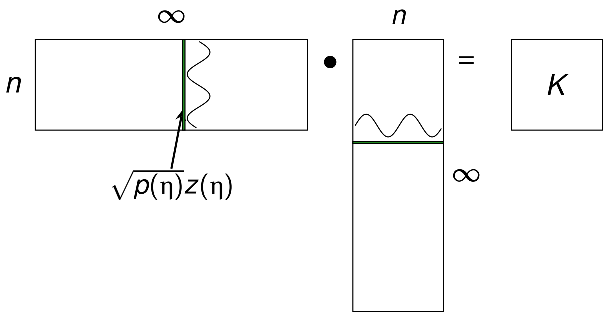


For each $\eta \in \mathbb{R}$ let

$$z(\eta)_j := e^{-2\pi x_j \eta}$$

and let $d\mu(\eta) := p(\eta)d\eta$ so that

$$K = \int_{\mathbb{R}} z(\eta)z(\eta)^* d\mu(\eta).$$



For each $\eta \in \mathbb{R}$ let

$$z(\eta)_j := e^{-2\pi x_j \eta}$$

and let $d\mu(\eta) := p(\eta)d\eta$ so that

$$K = \int_{\mathbb{R}} z(\eta)z(\eta)^* d\mu(\eta).$$

Define the ridge leverage score function

$$\tau_\lambda(\eta) := p(\eta)z(\eta)^*(K+\lambda I)^{-1}z(\eta)$$

Define the ridge leverage score function

$$\tau_\lambda(\eta) := p(\eta) z(\eta)^* (K + \lambda I)^{-1} z(\eta)$$

Lemma

For every $\eta \in \mathbb{R}$

$$\tau_\lambda(\eta) \leq p(\eta) \cdot \frac{n}{\lambda}$$

Define the ridge leverage score function

$$\tau_\lambda(\eta) := p(\eta) z(\eta)^* (K + \lambda I)^{-1} z(\eta)$$

Lemma

For every $\eta \in \mathbb{R}$

$$\tau_\lambda(\eta) \le p(\eta) \cdot \frac{n}{\lambda}$$

Proof:

$$\begin{aligned}
\tau_\lambda(\eta) &= p(\eta) z(\eta)^* (K + \lambda I)^{-1} z(\eta) \\
&\le p(\eta) z(\eta)^* z(\eta) / \lambda \\
&= p(\eta) ||z(\eta)||_2^2 / \lambda \\
&= p(\eta) \cdot \frac{n}{\lambda}
\end{aligned}$$

Theorem

For every kernel k, any dataset $x_1, \ldots, x_n$, any $\varepsilon \in (0, 1/2)$ if Z is a Fourier Features matrix with $s = O(\frac{1}{\varepsilon^2} \frac{\mathbf{n}}{\lambda} s_\lambda \log s_\lambda)$ columns, then

$$(1 - \varepsilon)(K + \lambda I) \prec ZZ^T + \lambda I \prec (1 + \varepsilon)(K + \lambda I)$$

with high probability.

Theorem

For every kernel k, any dataset $x_1,\ldots,x_n$, any $\varepsilon \in (0,1/2)$ if Z is a Fourier Features matrix with $s = O(\frac{1}{\varepsilon^2} \frac{\mathbf{n}}{\lambda} s_\lambda \log s_\lambda)$ columns, then

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)$$

with high probability.

Is this good? Usually $\lambda = \omega(1)$ (e.g. $\lambda = \sqrt{n}$), and definitely $\lambda \leq n$.

Theorem

For every kernel k, any dataset $x_1,\dots,x_n$, any $\varepsilon \in (0,1/2)$ if Z is a Fourier Features matrix with $s = O(\frac{1}{\varepsilon^2}\frac{\mathbf{n}}{\lambda} s_\lambda \log s_\lambda)$ columns, then

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)$$

with high probability.

Is this good? Usually $\lambda = \omega(1)$ (e.g. $\lambda = \sqrt{n}$), and definitely $\lambda \le n$.

Is this best possible? basically YES, even for 1d datasets!

Theorem

For every kernel k, any dataset $x_1, \ldots, x_n$, any $\varepsilon \in (0, 1/2)$ if Z is a Fourier Features matrix with $s = O(\frac{1}{\varepsilon^2} \frac{\mathbf{n}}{\lambda} s_\lambda \log s_\lambda)$ columns, then

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)$$

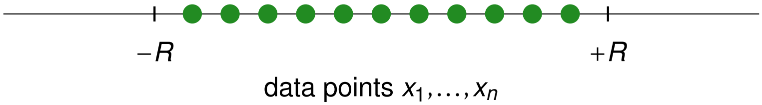
with high probability.

Is this good? Usually $\lambda = \omega(1)$ (e.g. $\lambda = \sqrt{n}$), and definitely $\lambda \leq n$.

Is this best possible? basically YES, even for 1d datasets!

Can we do better? YES, at least for bounded datasets in constant dimension

Assume: dimension d is constant (one in pictures), kernel is Gaussian, data points belong to $[-R, +R]$



Theorem (Upper bound, informal)

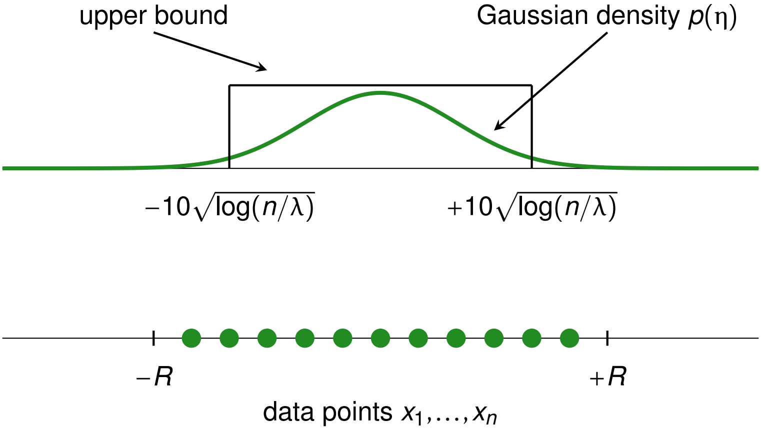
For every $|\eta| \le 10\sqrt{\log(n/\lambda)}$*:*

$$\tau_\lambda(\eta) \le 25 \max(R, 3000 \log^{1.5}(n/\lambda)).$$

Theorem (Upper bound, informal)

For every $|\eta| \le 10\sqrt{\log(n/\lambda)}$*:*

$$\tau_\lambda(\eta) \le 25\max(R, 3000\log^{1.5}(n/\lambda)).$$



Theorem (Lower bound, informal)

For integer n, regularization parameter λ, and radius R[1]*, there exist $x_1, \ldots, x_n \in [-R, R]$ such that for every $\eta \in [-100\sqrt{\log(n/\lambda)}, +100\sqrt{\log(n/\lambda)}]$*

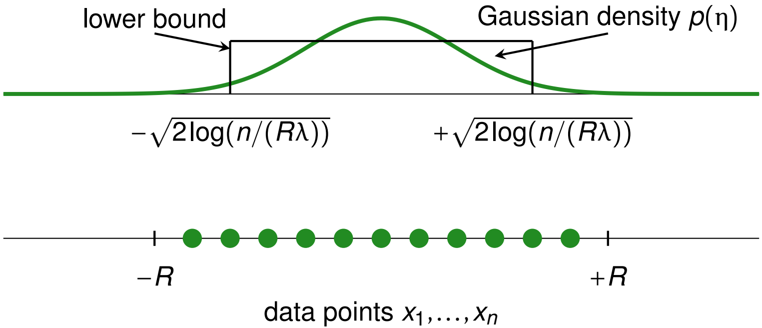
$$\tau_\lambda(\eta) \geq \frac{R}{150}\left(\frac{p(\eta)}{p(\eta) + 2R(\lambda/n)}\right).$$

[1] Restrictions apply

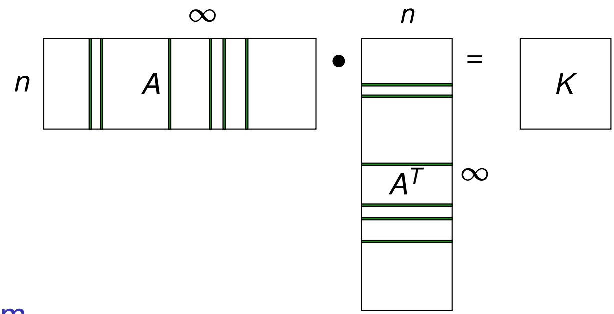
Theorem (Lower bound, informal)

For integer n, regularization parameter λ, and radius R[1], there exist $x_1, ..., x_n \in [-R, R]$ such that for every $\eta \in [-100\sqrt{\log(n/\lambda)}, +100\sqrt{\log(n/\lambda)}]$

$$\tau_\lambda(\eta) \geq \frac{R}{150}\left(\frac{p(\eta)}{p(\eta) + 2R(\lambda/n)}\right).$$



[1] Restrictions apply



Theorem

Suppose that

- *for each* $i = 1, \ldots, s$ *one has* $Z_i \sim a_\eta$ *with probability* $\tau_\lambda(\eta)d\eta$ *independently;*
- $s = O(\varepsilon^{-2} s_\lambda \log s_\lambda)$.

Then

$$(1-\varepsilon)(K+\lambda I) \prec ZZ^T + \lambda I \prec (1+\varepsilon)(K+\lambda I)$$

with high probability.

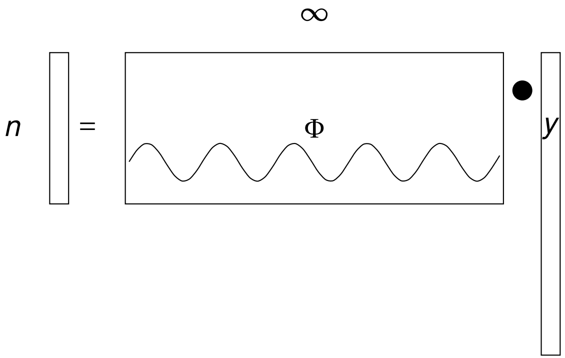
Statistical dimension of $s_\lambda(K) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \lambda}$

- Leverage score density function
- Primal-dual characterization
- Tight lower bound for Fourier Features

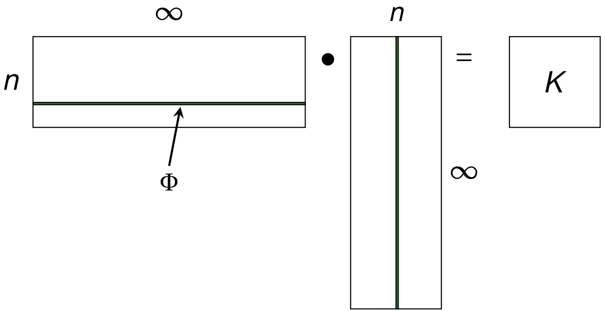
- Leverage score density function
- **Primal-dual characterization**
- Tight lower bound for Fourier Features

Define operator $\Phi : L_2(d\mu) \to \mathbb{C}^n$ by

$$\Phi y = \int_{\mathbb{R}} z(\xi) y(\xi) d\mu(\xi),$$



We have $\Phi\Phi^* = K$.

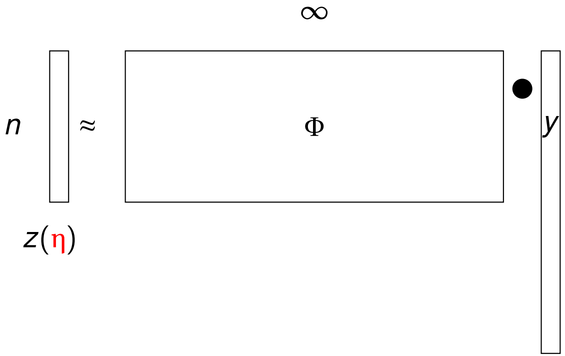


Lemma

The ridge leverage function can alternatively be defined as follows:

$$\tau_\lambda(\eta) = \min_{y \in L_2(d\mu)} \lambda^{-1} \|\Phi y - \sqrt{p(\eta)} z(\eta)\|_2^2 + \|y\|_{L_2(d\mu)}^2$$

Intuition: recombine many columns of Φ to get our column (i.e. frequency η), approximately

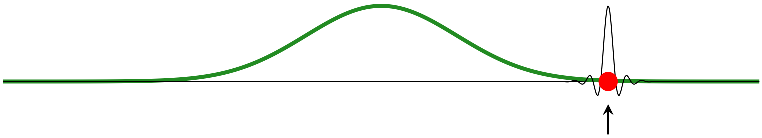


For a function $y \in L_2(d\mu)$

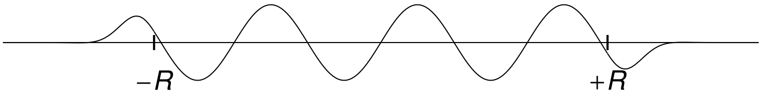
$$\Phi y = \int_{\mathbb{R}} z(\xi) y(\xi) d\mu(\xi)$$

Fix $\eta \in \mathbb{R}$. Want to upper bound

$$\tau_\lambda(\eta) = \min_{y \in L_2(d\mu)} \lambda^{-1} \|\Phi y - \sqrt{p(\eta)} z(\eta)\|_2^2 + \|y\|_{L_2(d\mu)}^2$$



Gaussian times **sinc**



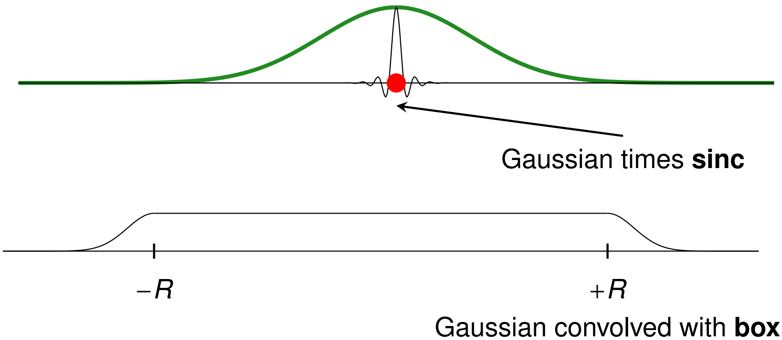
Gaussian convolved with **box**

For a function $y \in L_2(d\mu)$

$$\Phi y = \int_{\mathbb{R}} z(\xi) y(\xi) d\mu(\xi)$$

Fix $\eta \in \mathbb{R}$. Want to upper bound

$$\tau_\lambda(\eta) = \min_{y \in L_2(d\mu)} \lambda^{-1} ||\Phi y - \sqrt{p(\eta)} z(\eta)||_2^2 + ||y||_{L_2(d\mu)}^2$$

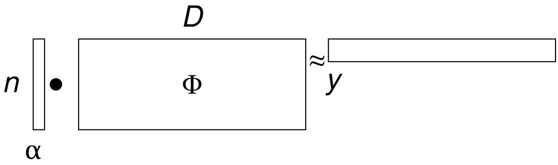


Lemma

The ridge leverage function can alternatively be defined as follows:

$$\tau_\lambda(\eta) = \max_{\alpha \in \mathbb{C}^n} \frac{p(\eta) \cdot |\alpha^* z(\eta)|^2}{\|\Phi^* \alpha\|_{L_2(d\mu)}^2 + \lambda \|\alpha\|_2^2}$$

Intuition: recombine rows of Φ to create a 'localized' vector



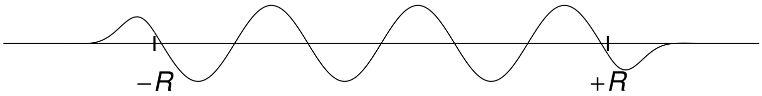
Similar construction of test functions

- Leverage score density function
- Primal-dual characterization
- **Tight lower bound for Fourier Features**

Tight lower bound – proof idea

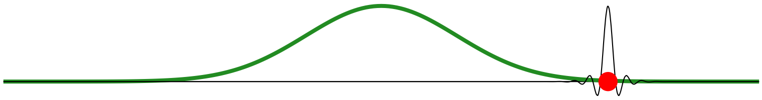
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K \alpha + \lambda \|\alpha\|_2^2 \in (1 \pm \varepsilon)(\alpha^T Z Z^T \alpha + \lambda \|\alpha\|_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

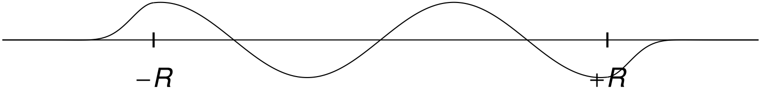
$$\alpha^T K \alpha = \int_{\mathbb{R}^d} p(\eta) \left| \sum_{j=1}^{n} e^{-2\pi i x_j \eta} \alpha_j \right|^2 d\eta$$



Tight lower bound – proof idea

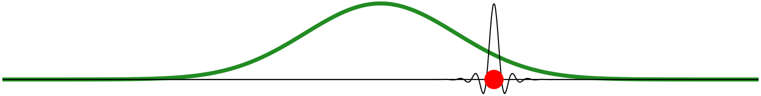
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha + \lambda||\alpha||_2^2 \in (1 \pm \varepsilon)(\alpha^T ZZ^T\alpha + \lambda||\alpha||_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

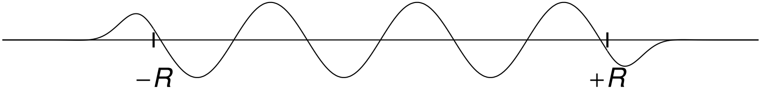
$$\alpha^T K\alpha = \int_{\mathbb{R}^d} p(\eta)\left|\sum_{j=1}^{n} e^{-2\pi i x_j \eta}\alpha_j\right|^2 d\eta$$



Tight lower bound – proof idea

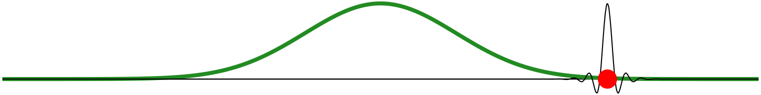
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha + \lambda||\alpha||_2^2 \in (1 \pm \varepsilon)(\alpha^T ZZ^T\alpha + \lambda||\alpha||_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

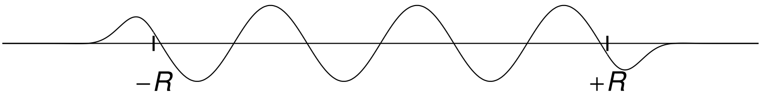
$$\alpha^T K\alpha = \int_{\mathbb{R}^d} p(\eta)\left|\sum_{j=1}^{n} e^{-2\pi i x_j \eta}\alpha_j\right|^2 d\eta$$



Tight lower bound – proof idea

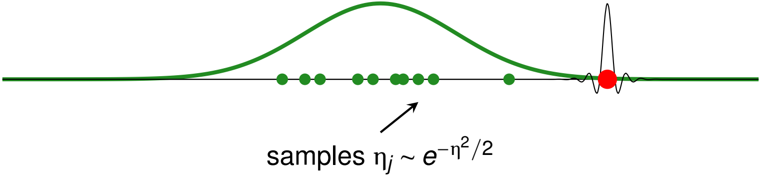
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha + \lambda||\alpha||_2^2 \in (1 \pm \varepsilon)(\alpha^T ZZ^T\alpha + \lambda||\alpha||_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

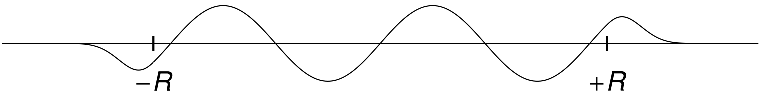
$$\alpha^T K\alpha = \int_{\mathbb{R}^d} p(\eta) \left| \sum_{j=1}^n e^{-2\pi i x_j \eta} \alpha_j \right|^2 d\eta$$



Tight lower bound – proof idea

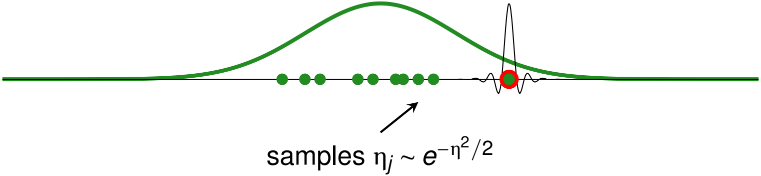
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha + \lambda\|\alpha\|_2^2 \in (1 \pm \varepsilon)(\alpha^T ZZ^T \alpha + \lambda\|\alpha\|_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

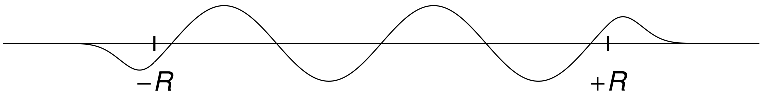
$$\alpha^T K\alpha = \int_{\mathbb{R}^d} p(\eta) \left| \sum_{j=1}^{n} e^{-2\pi i x_j \eta} \alpha_j \right|^2 d\eta$$



Tight lower bound – proof idea

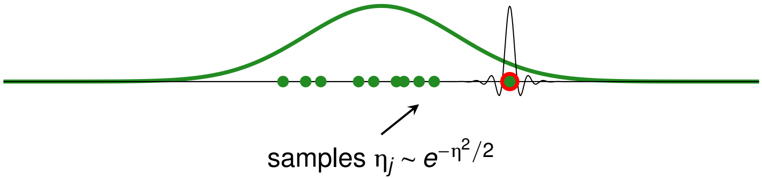
Need: for every $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha + \lambda\|\alpha\|_2^2 \in (1 \pm \varepsilon)(\alpha^T ZZ^T\alpha + \lambda\|\alpha\|_2^2)$$



For a vector $\alpha \in \mathbb{R}^n$

$$\alpha^T K\alpha = \int_{\mathbb{R}^d} p(\eta)\left|\sum_{j=1}^{n} e^{-2\pi i x_j \eta}\alpha_j\right|^2 d\eta \approx \frac{1}{s}\sum_{k=1}^{s} p(\eta_k)\left|\sum_{j=1}^{n} e^{-2\pi i x_j^T \eta_j}\alpha_j\right|^2$$



Experiments: one-dimensional

Sample from the function

$$f^\star(x) = \sin(6x) + \sin(60\exp(x))\,.$$

Use a 400-point uniform grid spanning $[-5/2\pi, +5/2\pi]$, and sample according to

$$y_i = f^\star(x_i) + \nu_i\,.$$

where ν_i is i.i.d. Gaussian noise.

Experiments: one-dimensional

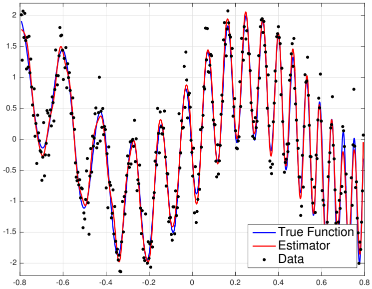
Sample from the function

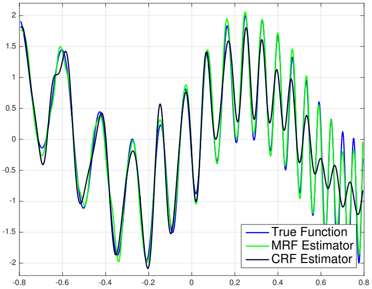
$$f^\star(x) = \sin(6x) + \sin(60\exp(x)).$$

Use a 400-point uniform grid spanning $[-5/2\pi, +5/2\pi]$, and sample according to

$$y_i = f^\star(x_i) + \nu_i.$$

where ν_i is i.i.d. Gaussian noise.





Experiments: two-dimensional

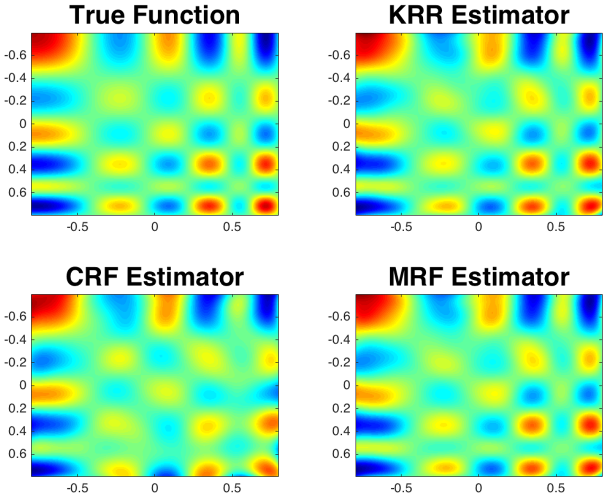
$$f^\star(x,z) = (\sin(x) + \sin(10\exp(x)))(\sin(z) + \sin(10\exp(z)))\,.$$

Sample points on a 40×40 uniform grid.

Experiments: two-dimensional

$$f^\star(x,z) = (\sin(x) + \sin(10\exp(x)))(\sin(z) + \sin(10\exp(z))).$$

Sample points on a 40×40 uniform grid.



Summary

Our results:

- tight bounds for Fourier Features for bounded datasets in constant dimension

Summary

Our results:

- tight bounds for Fourier Features for bounded datasets in constant dimension
- tight bounds on leverage score function for bounded datasets in any constant dimension

Summary

Our results:

- tight bounds for Fourier Features for bounded datasets in constant dimension
- tight bounds on leverage score function for bounded datasets in any constant dimension

Subspace embeddings with poly(d) dependence? Tight bounds for worst case datasets? Does Rahimi-Recht work on 'typical' datasets? Other kernels?

Thank you!