Journées de Statistique Mathématique et Data Science JSMDS2019, Hammamet 01–04 Novembre 2019.

Learning Via Reproducing Kernels and Nonparametic Regression.

Abderrazek Karoui
University of Carthage, Faculty of Sciences of Bizerte, Tunisia

In Collaboration with J. F. Dupuy, B. Bousselmi, A. Bonami and P. Jaming.

03 Novembre 2019

Outline

1. Reproducing kernel, positive-definite kernels and RKHS
   - Definitions and properties
   - The Legendre- and the Sinc-kernels cases
2. Kernels based scheme for nonparametric regression
   - Some concentration inequalities from Statistics
   - Nonparametric regression model
   - Nonparametric regression by Legendre-and Sinc-kernels
3. Tikhonov regularization and a learning scheme in a RKHS
   - Analysis of a RKHS-Regularization scheme for learning problem
   - Numerical Simulations

We first recall the concept of a positive definite kernel. In the sequel, we let $\mathcal{X}$ denote a non-empty set of $\mathbb{R}$.

Definition (Positive definite kernel)

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel if it is symmetric and for any integer $n \geq 1$ and any $x_1, \ldots, x_n \in \mathcal{X}$ and any $c_1, \ldots, c_n \in \mathbb{R}$, we have

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

Definition (reproducing kernel, RKHS)

Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space with an inner product $< \cdot, \cdot >$. A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel r.k. if:

1. $\forall x \in \mathcal{X}$, the function $K_x(\cdot) = K(x, \cdot) \in \mathcal{H}$.
2. For any $x \in \mathcal{X}$ and any $f \in \mathcal{H}$, we have $f(x) =< f, K_x >$. If $\mathcal{H}$ has a r.k., then it is called an reproducing kernel Hilbert space (RKHS).

Aronszajn's Theorem

There is an equivalence between the concepts of positive definite kernels and R.K. This is given by Aronszajn's theorem.

Theorem (Aronszajn's 1950)

A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is positive definite if and only if it is a reproducing kernel.*

Aronszajn's Theorem

There is an equivalence between the concepts of positive definite kernels and R.K. This is given by Aronszajn's theorem.

Theorem (Aronszajn's 1950)

A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if and only if it is a reproducing kernel.

From Riesz theorem, a Hilbert space of real-valued functions on X is an RKHS whenever the evaluation linear functional $L_x : f \to f(x)$ is continuous on $\mathcal{H}$ for every $x \in X$.

Aronszajn's Theorem

There is an equivalence between the concepts of positive definite kernels and R.K. This is given by Aronszajn's theorem.

Theorem (Aronszajn's 1950)

A kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is positive definite if and only if it is a reproducing kernel.

From Riesz theorem, a Hilbert space of real-valued functions on X is an RKHS whenever the evaluation linear functional $L_x : f \rightarrow f(x)$ is continuous on $\mathcal{H}$ for every $x \in X$.

Remark

Let $\mathcal{F}$ be a Hilbert space with an inner product $< \cdot, \cdot >_{\mathcal{F}}$ and let $\phi : \mathcal{X} \rightarrow \mathcal{F}$ be a map, then $K(x, y) =< \phi(x), \phi(y) >_{\mathcal{F}}$ is a positive definite kernel.

Constructions of R.K.

Let $\mathcal{H}_K$ be the RKHS, associated with the R.K. $K(\cdot,\cdot)$ and let K_1, K_2 be two R.K., then

- The sum kernel, $K(x,y) = K_1(x,y) + K_2(x,y),\ x,y \in \mathcal{X}$ is a R.K. and $\mathcal{H}_K = \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$.
- The product kernel, $K(x,y) = K_1(x,y)K_2(x,y),\ x,y \in \mathcal{X}$ is a R.K.
- Convolution kernel. By Bochner's Theorem, a continuous function f on $\mathbb{R}$ with $f(0) = 1$ is positive definite if and only if it is the Fourier transform of a probability measure μ on $\mathbb{R}$. Hence a convolution kernel $K(x,y) = k(x-y)$ is a R.K. whenever the function $k(\cdot)$ is the Fourier transform of a probability measure on $\mathbb{R}$.

Mercer's kernel

A Mercer's kernel K is a continuous and positive definite kernel. Moreover, if μ is a positive measure on $\mathcal{X}$, $K(\cdot,\cdot) \in L^2(\mathcal{X}\times\mathcal{X}, d\mu\otimes\mu)$ and φ_n, λ_n, $n \geq 0$ denote the orthonormal eigenfunctions and associated eigenvalues of the operator T_K, that is

$$T_K\varphi(x) = \int_X K(x,y)\varphi_n(y)\, d\mu(y) = \lambda_n\, \varphi_n(x), \quad x \in \mathcal{X}.$$

We have

$$K(x,y) = \sum_{n=0}^{\infty} \lambda_n\, \varphi_n(x)\varphi_n(y), \quad x,y \in X.$$

The associated RKHS is given by

$$\mathcal{H}_K = \{f \in L^2(X, d\mu),\ f = \sum_{n\geq 0} a_n(f)\varphi_n,\ \sum_{n\geq 0} \frac{|a_n(f)|^2}{\lambda_n} < +\infty\}.$$

with inner product, given by

$$< f, g >_K = \sum_{n\geq 0} \frac{a_n(f)a_n(g)}{\lambda_n}, \quad f = \sum_{n\geq 0} a_n(f)\varphi_n,\ g = \sum_{n\geq 0} a_n(g)\varphi_n.$$

Christoffel-Darboux Legendre Kernel

The Legendre polynomial P_n of degree n is given by the bounded solution of the differential equation

$$(1-x^2)\,y'' - 2x\,y' + \chi_n\,y = 0, \quad x \in I = [-1,1], \quad \chi_n = n(n+1).$$

The P_n, $n \geq 0$ an orthogonal basis of $L^2(I)$. We have $\int_{-1}^{1} P_n(x)P_m(x)dx = \frac{2}{2n+1}\delta_{n,m}$. The normalized Legendre polynomial is given by $\widetilde{P}_n(x) = \sqrt{n+\frac{1}{2}}\frac{1}{2^n\, n!}\frac{d^n}{dx^n}\left((x^2-1)^n\right)$. The Christoffel-Darboux Legendre kernel is given by $K_n(x,y) =$

$$\begin{aligned} = \sum_{j=0}^{n} \widetilde{P}_j(x)\widetilde{P}_j(y) &= \frac{2n+2}{(2n+1)^2}\frac{P_{n+1}(x)P_n(y) - P_n(x)P_{n+1}(y)}{x-y}, \, x \neq y \\ &= \frac{2n+2}{(2n+1)^2}\Big(P'_{n+1}(x)P_n(x) - P'_n(x)P_{n+1}(x)\Big), \, x = y. \end{aligned}$$

The Sinc-kernel

For a fixed real number $c > 0$, the Sinc-kernel is given by
$K_c(x,y) = \dfrac{\sin(c(x-y))}{\pi(x-y)}, \quad x, y \in \mathbb{R}$. Note that $K(x,y) = \dfrac{c}{\pi}\widehat{\mu}_c(x-y)$,
where μ_c is the uniform probability measure, given by
$\mu_c(x) = \dfrac{1}{2c}\mathbf{1}_{[-c,c]}(x)$, $x \in \mathbb{R}$. The associated RKHS is the Paley-Wiener space of $c-$bandlimited functions

$$\mathcal{B}_c = \{f \in L^2(\mathbb{R}),\ Supp^t \widehat{f} \in [-c,c]\}.$$

When the Sinc-kernel is restricted to the square $I^2 = [-1,1]^2$, one gets a Mercer's kernel written as

$$K_c(x,y) = \frac{\sin(c(x-y))}{\pi(x-y)} = \sum_{n=0}^{\infty} \lambda_n(c)\psi_{n,c}(x)\psi_{n,c}(y), \quad \forall\, x, y \in I = [-1,1]. \tag{1}$$

Here, the $\lambda_n(c)$ and $\psi_{n,c}$ are the positive eigenvalues and their associated eigenfunctions of the operator $\mathcal{Q}_c f(x) = \displaystyle\int_{-1}^{1} \frac{\sin(c(x-y))}{\pi(x-y)} f(y)\,dy$.

These $\psi_{n,c}$ are known as the prolate spheroidal wave functions (PSWFs). The theory of the PSWFs is due to the pioneer works of D.Slepian and his collaborators H.Landau and H.Pollack. Since the early 1960's, there has been a great and an extensive study for the decay and behaviour of the spectrum $\lambda_n(c), n \geq 0$, [Slepian , Widom, Landau, Fuchs, Bonami-Karoui 2017, ...]. Recently, in [Bonami-Karoui (2017)], it has been shown that for any $1 \leq a < \frac{4}{e}$, there exists $N_{c,a} \in \mathbb{N}$ such that

$$\lambda_n(c) \leq e^{-2n \log\left(\frac{an}{c}\right)}, \quad \forall\ n \geq N_{c,a}. \tag{2}$$

Also, from [Bonami-Jaming-Karoui (2018)], we have the following non-asymptotic behaviour of the $\lambda_n(c)$.

$$\lambda_n(c) \geq 1 - \frac{7}{\sqrt{c}} \frac{(2c)^n}{n!} e^{-c}, \quad \forall\, 0 \leq n < \frac{c}{2.7} \tag{3}$$

and

$$\lambda_n(c) \leq \exp\left(-(2n+1)\log(\frac{2}{ec}(n+1))\right), \quad \forall\, n \geq \max(\frac{ec}{2}, 2). \tag{4}$$

Some Concentration inequalities

The following is an extension of the original Bennett's concentration inequality [Bennett (1962) to vector-valued random variables, given in [Pinelis, (1994)]

Theorem (Bennett's inequality)

Let $\xi_1, \dots, \xi_n$ be independent random variables with values in a Hilbert space $\mathcal{H}$. Assume that $\|\xi_i\| \leq M$ almost surely. Let $\sigma^2 = \sum_{i=1}^{n} \mathbb{E}(\|\xi_i\|^2)$.

Then

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}(\xi_i))\right\| \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon}{2M}\log\left(1 + \frac{nM\varepsilon}{\sigma^2}\right)\right).$$

Theorem (Pinelis,(1994))

Let $\xi_1, ..., \xi_n$ *independent random variables with values in a separable Hilbert space such that* $E[\xi_i] = 0$ *and* $\|\xi_i\| \leq c, i = 1, ..., n$*, then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\| \leq \varepsilon\right) \geq 1 - 2e^{-\frac{n\varepsilon^2}{2c^2}}$$

Nonparametric regression model

For a given complete metric space $\mathcal{X}$ and for a given discrete set of $n-$random observations $\mathbf{Z} = \{(x_i, y_i),\ i = 1, \dots, n\}$, where $x_i \in \mathcal{X}$ and the responses $y_i \in \mathbb{R}$, the associated nonparametric regression model assumes that

$$y_i = f(x_i) + \eta_i, \quad i = 1, \dots, n, \tag{5}$$

for some i.i.d. random errors η_i with zero means. Also, we assume that the random errors η_i do not depend on the random samples x_i. The main issue related to this model is to estimate the function $f(\cdot)$ from the given set of $n-$random observations.

Also, for the random discrete data set $\mathbf{Z} = (x_i, y_i)_{1\leq i\leq n}$, drawn from a joint probability measure ρ on $\mathcal{X} \times \mathbb{R}$, the associated regression function f_ρ is defined on $\mathcal{X}$ by

$$f_\rho(x) = \mathbb{E}(y|x) = \int_{\mathbb{R}} y d\rho(y|x), \quad x \in \mathcal{X}, \tag{6}$$

where, $d\rho(y|x)$ is the conditional distribution at $x \in \mathcal{X}$. That is

$$d\rho(y|x) = \frac{\rho(x,y)}{\int \rho(x,y)\, dy}\, dy.$$

The conditional expectation $\mathbb{E}(y|x)$ is the minimum-mean-square-error prediction of y.

Nonparametric regression by the Legendre kernel

Let π_N be the projection operator on the span of the $N-$first Legendre polynomials defined by

$$\pi_N(f)(x) = < f, K_N(x,\cdot) >_{L^2(I)} = \sum_{k=0}^{N} < f, \widetilde{P}_k(y) > \widetilde{P}_k(x), \quad x \in I = [-1,1]. \tag{7}$$

For a positive integer n, define the associated empirical projection operator by

$$\widetilde{\pi}_{N,n}(y)(x) = \sum_{k=0}^{N} \left[\frac{2}{n} \sum_{i=1}^{n} y_i \widetilde{P}_k(x_i) \right] \widetilde{P}_k(x), \quad x \in I, \tag{8}$$

where the x_i are random samples following the uniform law on I. Based on Pinelis concentration inequality, we get

Theorem (Dupuy-Bousselmi-Karoui (2019))

Under the above notations and hypotheses, we have

$$\mathbb{E}[\widetilde{\pi}_{N,n}(y)(x)] = \pi_N(f)(x) + \frac{1}{n}\sum_{i=1}^{n}\eta_i, \quad x \in I. \tag{9}$$

In particular, if f is a bounded function belonging to $H^s(I)$ for some $s > 0$, then for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\|f - \widetilde{\pi}_{N,n}(y)\| \leq \varepsilon + \sqrt{2}\eta + c_1 N^{-s}\|f\|_{H^s}\right) \geq 1 - 2e^{-\dfrac{n\varepsilon^2}{2c_f^2}}, \quad \eta = \max_i |\eta_i|. \tag{10}$$

Here, $c_f = \left(4(N+1)^2(\|f\|_\infty + \eta)^2 + 2\|f\|^2\right)^{1/2} + \sqrt{2}\eta$ *and c_1 is a uniform constant.*

Remark

Once we have the expectation (9), it can be checked that if $f \in \mathcal{B}_c$, that is $c-$band-limited function for some $c > 0$, then by using the quality of approximation of functions from $\mathcal{B}_c$, given in [Jaming-Karoui-Spektor (2016)], one gets for any $N \geq \frac{ec}{2}$ and for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\|f - \widetilde{\pi}_{N,n}(y)\| \leq \varepsilon + \sqrt{2}\eta + c_2 e^{-(N+2)\log\left(\frac{2N+2}{ec}\right)} \|f\|_{L^2(\mathbb{R})} \right) \geq 1 - 2e^{-\frac{n\varepsilon^2}{2c_f^2}} \tag{11}$$

for some uniform constant c_2.

Nonparametric regression by the Sinc-kernel

Let $c > 0$ and let π_c be the projection operator over the span of the PSWFs $\psi_{n,c}$, that is

$$f(x) = \pi_c(f)(x) = \sum_{n=0}^{\infty} < f, \psi_{n,c} > \psi_{n,c}(x), \quad \forall f \in L^2(I). \tag{12}$$

The empirical projection operator associated with the Sinc-kernel and the regression model (5) is given by

$$\widetilde{\pi}_{c,n}(y)(x) = \frac{2}{n}\sum_{k=1}^{n} y_i K_c(x, x_i), \qquad K_c(x, y) = \frac{\sin c(x-y)}{\pi(x-y)}, \tag{13}$$

where the x_i are random samples following the uniform law on I. Note that

$$\mathbb{E}(\widetilde{\pi}_{c,n}(f)) = \sum_{k=0}^{\infty} \lambda_k(c) < f, \psi_{k,c} > \psi_{k,c}(x) \neq \pi_c(f).$$

We consider the weighted Sobolev space $\widetilde{H}^s(I)$, defined by

$$\widetilde{H}^s(I) = \{f \in L^2(I),\ \|f\|^2_{\widetilde{H}^s} = \sum_{k\geq 0}(\chi_k(c))^s| < f, \psi_{k,c} > |^2 < +\infty\}. \quad (14)$$

Here, the $\chi_k(c)$ are the eigenvalue of the differential operator associated with the $\psi_{k,c}$. Recall that from [Bonami-Jaming-Karoui (2018)], we have

$$\lambda_n(c) \geq 1 - \frac{7}{\sqrt{c}} \frac{(2c)^n}{n!} e^{-c}, \quad \forall\, 0 \leq n < \frac{c}{2.7}.$$

Lemma

For any real number $c \geq 6$ and any positive integer $N + 1 \leq \frac{c}{3}$, we have

$$\frac{e^{-c}}{\sqrt{c}} \sum_{k=0}^{N} \frac{(2c)^k}{k!} \leq \frac{1}{\sqrt{6}} \left(\frac{e^2}{6}\right)^{-c/3}. \quad (15)$$

Lemma

Under the above notation, for $f \in \widetilde{H}^s(I)$, $s > 0$ *and for any real* $c \geq 6$, *we have*

$$\| f - \mathbb{E}\left(\widetilde{\pi}_{c,n}(f)\right) \| \leq \frac{7}{\sqrt{6}} \left(\frac{e^2}{6}\right)^{-[c/3]} \|f\| + \left[\frac{c}{3}\right]^{-s} \|f\|_{\widetilde{H}^s}. \qquad (16)$$

Here, $[x]$ *denotes the integer part of* $x \in \mathbb{R}$.

We use the previous two lemmas and Pinelis inequality and get the quality of the nonparametric regression by the Sinc-kernel in the weighted Sobolev space.

Theorem (Dupuy-Bousselmi-Karoui (2019))

Let $c \geq 6$ and $s > 0$ be positive real numbers. Consider the regression model (5), with the assumption that $f \in \widetilde{H}^s(I)$ and bounded on I. Then, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\|f - \widetilde{\pi}_{c,n}(y)\| \leq \varepsilon + \sqrt{2}\eta + \epsilon_c\|f\| + \left[\frac{c}{3}\right]^{-s}\|f\|_{\widetilde{H}^s}\right) \geq 1 - 2e^{-\frac{n\varepsilon^2}{2c_f^2}}. \tag{17}$$

Here, $\widetilde{\pi}_{c,n}(y)$ is as given by (13) and

$$c_f = \sqrt{2}\frac{c}{\pi}\left(2\|f\|_\infty + (2+\sqrt{2})\eta + \|f\|\right), \quad \epsilon_c = \frac{7}{\sqrt{6}}\left(\frac{e^2}{6}\right)^{-[c/3]}, \quad \eta = \max_i |\eta_i|.$$

Tikhonov regularisation and nonparametric regression

The Tikhonov regularization algorithm (learning scheme) for the approximation of the regression function f_ρ by using the drawn data set $\mathbf{z} = (x_i, y_i)_{1 \leq i \leq n}$ is described as follows. For a regularization parameter $\lambda > 0$, find $f_{\mathbf{z},\lambda} \in \mathcal{H}_K$, solution of the minimization problem,

$$f_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(f(X_i) - y_i \Big)^2 + \lambda \|f\|_K^2 \right\}, \qquad (18)$$

where $\|\cdot\|_K$ is the norm associated with the RKHS $\mathcal{H}_K$, generated by a Mercer's kernel $K(\cdot,\cdot)$.

Representer Theorem

Theorem (Argyriou-Micchelli-Pontil (2009))

Let $\mathcal{X}$ be a nonempty set, K be a positive definite kernel on $\mathcal{X} \times \mathcal{X}$ and $\mathcal{H}_K$, be the associated RKHS. Let $R : \mathcal{H}_K \to \mathbb{R}$ be a differentiable function (in Gateaux sense). Given n random observation $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ and an arbitrary error function $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R}$. A minimizer

$$f^* = \arg \min_{f \in \mathcal{H}_K} \{E((x_1, y_1, f(x_1)), \dots, (x_1, y_1, f(x_1))) + R(f)\}$$

admits a representation of the form

$$f^*(\cdot) = \sum_{i=1}^{n} c_i K(\cdot, x_i), \quad c_i \in \mathbb{R}$$

if and only if there exists a nondecreasing function $h : [0, \infty) \to \mathbb{R}$, for which $R(f) = h(\|f\|)$.

In [Smale-Zhou (2007)], an analysis study of the following Tikhonov regularization algorithm for the approximation of f_ρ has been done. For a regularization parameter $\lambda > 0$, find $f_{\mathbf{z},\lambda} \in \mathcal{H}_K$, solution of

$$\textbf{Learning Scheme} \quad f_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(f(x_i) - y_i \Big)^2 + \lambda \|f\|_K^2 \right\}, \tag{19}$$

where $\|\cdot\|_K$ is the norm associated with the RKHS $\mathcal{H}_K$, generated by the Mercer's kernel $K(\cdot,\cdot)$. The random samples set $\mathbf{z} = (x_i, y_i)$, $i = 1, \dots, n$ is drawn according to a joint distribution $\rho(x, y)$.
The solution of the previous problem is given by

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{n} c_i K(x_i, x), \tag{20}$$

where $\mathbf{C} = (c_i)_{1 \leq i \leq n}$ is given as solution of the system

$$\left[\Big[K(x_i, x_j) \Big]_{1 \leq i,j \leq n} + n\lambda I_n \right] \mathbf{C} = \mathbf{G}_\lambda \mathbf{C} = \mathbf{Y}, \quad \mathbf{Y} = (y_i)_{1 \leq i \leq n}. \tag{21}$$

Given a positive definite Mercer's kernel K, so that the RKHS $\mathcal{H}_K \subset C(\mathcal{X})$, let $\kappa^2 = \sup_{x \in \mathcal{X}} K(x,x)$ and let L_K, the integral operator, defined on $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ by

$$L_K(f)(x) = \int_{\mathcal{X}} K(x,y) f(y) d\rho_{\mathcal{X}}(y).$$

Then, by using Bennett's inequality and some involved estimates of operators and functions norms, the following error analysis of the previous learning scheme has been given in [Smale-Zhou (2007)].

Theorem (Smale-Zhou (2007))

Under the previous notation, assume that for some $\frac{1}{2} < r \leq 1$, we have $L_K^{-r} f_\rho \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$. Then, with confidence $1 - \delta$, we have

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq 2 \log(2/\delta) \left(\frac{3\kappa M}{\sqrt{n}\lambda} + \lambda^{r-\frac{1}{2}} \|L_K^{-r} f_\rho\|_\rho \right). \qquad (22)$$

Remark

The previous Tikhonov regularization scheme with Mercer's kernel has the advantage to work with random sampling set X_i, $i = 1, \ldots, n$ drawn from a fairly general probability measure ρ_X on a compact metric space X.

Remark

The previous Tikhonov regularization scheme with Mercer's kernel has the advantage to work with random sampling set X_i, $i = 1, \dots, n$ *drawn from a fairly general probability measure* ρ_X *on a compact metric space* X.

Remark

The previous scheme has the drawback that the random Gram-matrix $\Big[K(X_i, X_j)]_{1 \le i,j \le n}$ *has a large condition number. Consequently, for for very small values of the parameter* λ, *the matrix* $\mathbf{G}_\lambda$ *will have a large condition number that generate a numerical instability in computing the expansions coefficients vector* $\mathbf{C}$ *by using the Tikhonov regularization scheme.*

Why regularisation is essential

The answer follows from the combinaison of Weyl's perturbation theorem with quality of approximation of the spectrum of the random Gramian matrices by the spectrum of the associated integral operator.

Theorem (Weyl's perturbation theorem)

Let H be an $n \times n$ Hermitian matrix and let E be a Hermitian perturbation matrix. We let $\mu_1(H+E) \geq \mu_2(H+E) \geq \cdots \mu_n(H+E)$,

$$\lambda_1(E) \geq \lambda_2(E) \geq \cdots \geq \lambda_n(E), \;\; \rho_1(E) \geq \rho_2(E) \geq \cdots \rho_n(E),$$

denote the eigenvalues of the different matrices arranged in the decreazing order. Then, the following inequalities hold

$$\lambda_i(H) + \rho_n \leq \mu_i(H+E) \leq \lambda_i(H) + \rho_1, \quad i = 1, \ldots, n.$$

Why regularisation is essential

Note that if the matrices H and E are positive definite, $\rho_n > 0$ and the $2-$condition number of the matrix $H + E$ is bounded above by

$$\kappa_2(H+E) \leq \frac{\lambda_1(H) + \rho_1}{\rho_n} \tag{23}$$

Consider the special case of the Sinc-kernel case $K_c(x,y) = \dfrac{\sin(c(x-y))}{\pi(x-y)}$, and $H = \Big[K_c(x_i, x_j)\Big]_{1\leq i,j\leq n}$, with x_i are random samples following the uniform law on $[-1,1]$ and $E = \lambda I_n$.

• By combining the non-asymptotic estimate of the $\lambda_n(c)$, given in [Bonami-Jaming-Karoui (2018)], that is

$$\lambda_n(c) \leq \exp\left(-(2n+1)\log(\frac{2}{ec}(n+1))\right), \quad \forall\, n \geq \max(\frac{ec}{2}, 2),$$

$$\lambda_k(c) > 1 - \frac{7}{\sqrt{c}}\frac{(2c)^k}{k!}e^{-c}, \quad \forall\, 0 \leq k < \frac{c}{2.7},$$

the ℓ_2-estimate of the spectrum of H and the spectrum of T_{K_c}, given in [Bonami-Karoui (2017)] and the estimates given in [Shawe-Taylor et al. (2005)] for $1 \leq d \leq n$,

$$\mathbb{E}\Big(\sum_{j<d} \lambda_j(H)\Big) \geq \sum_{j<d} \lambda_j(T_{K_c}) \tag{24}$$

$$\mathbb{E}\Big(\sum_{j\geq d} \lambda_j(H)\Big) \leq \sum_{j\geq d} \lambda_j(T_{K_c}). \tag{25}$$

We conclude that for $n \geq \frac{ec}{2}$ and with high probability, we have

$$\kappa_2(\mathbf{G}) = \frac{\lambda_1(G)}{\lambda_n(\mathbf{G})} \geq \frac{\lambda_0(\mathcal{Q}_c)}{\lambda_{n-1}(\mathcal{Q}_c)} \gtrsim e^{2n\log((2n)/(ec))}$$

while

$$\kappa_2(\mathbf{G}_\lambda) = \frac{\lambda_0(G_\lambda)}{\lambda_n(\mathbf{G}_\lambda)} \lesssim \frac{1}{\lambda}.$$

Numerical Simulations

Consider the noisy nonparametric model given by

$$y_i = \widehat{\varphi}_1(x_i) = \varphi_1(x_i) + \eta_i, \qquad \varphi_1(x) = \frac{\sin(20x)}{20x}, \quad x \in I,$$

with random samples x_i following the uniform distribution on I and the η_i are of Gaussian white noise type given by $\eta_i = 0.1 Z_i$, with Z_i following the standard normal distribution. We have computed the empirical projections $\widetilde{\pi}_{N,n}(\widehat{\varphi})$ and $\widetilde{\pi}_{c,n}(\widehat{\varphi})$, with $n = 1000$ and associated with the Legendre kernel with $N = 20$, and the Sinc-kernel with $c = 20$.

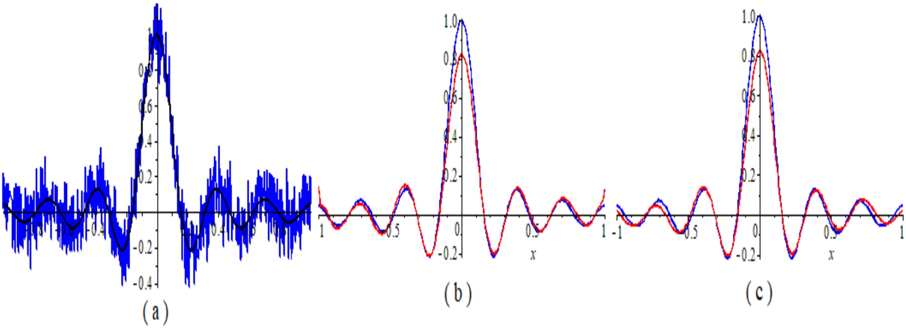


Figure : (a) Graphs of φ_1 and $\widehat{\varphi}_1$, (b) Graph of $\widetilde{\pi}_{N,n}(\widehat{\varphi}_1)$ (c) Graph of $\widetilde{\pi}_{c,n}(\widehat{\varphi}_1)$.

Consider the noisy nonparametric model given by

$$y_i = \widehat{\varphi}(x_i) = \varphi(x_i) + \eta_i, \quad \varphi(x) = \exp(3x^2)\big(\cos(40x) + \sin(40x)\big), \quad x \in I,$$

where $\eta_i = 0.1Z_i$. We have applied the Tikhonov regularization with Sinc-kernel based scheme for the reconstruction of φ_2 with $n = 70$ random samples x_i, the bandwidth $c = 50$ and the reguarisation parameter $\lambda = 1E-08$. The following figures illustrate the obtained numrical results.

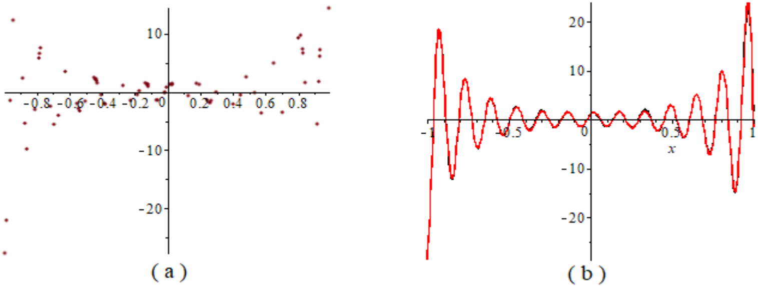
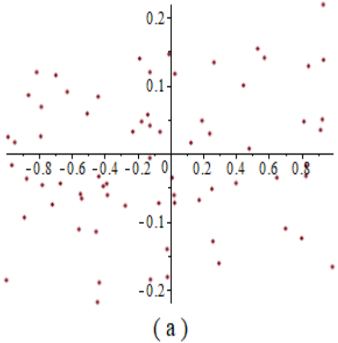


Figure : (a) Plots of the random observations set $\mathbf{z}$ (b) Graph of the approximation $\varphi_{\mathbf{z},\lambda}$.

Numerical Simulations



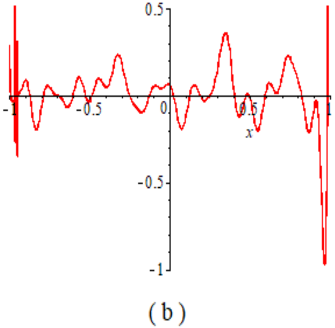


Figure : (a) Graph of the added Gaussian white noise (b) Graph of the reconstruction error $\varphi(x) - \varphi_{\mathbf{z},\lambda}(x)$.

References

[1] A. Argyriou, C. A. Micchelli and M. Pontil, When is there a representer Theorem ? Vector versus matrix regularizers? *Journal of Machine Learning Research*, **10** (2009) 2507–2529.

References

[1] A. Argyriou, C. A. Micchelli and M. Pontil, When is there a representer Theorem ? Vector versus matrix regularizers? *Journal of Machine Learning Research*, **10** (2009) 2507–2529.

[2] N. Aronszajn, Theory of Reproducing Kernels, *Transactions of the American Mathematical Society*, **68** (3) (1950), 337–404.

References

[1] A. Argyriou, C. A. Micchelli and M. Pontil, When is there a representer Theorem ? Vector versus matrix regularizers? *Journal of Machine Learning Research,* **10** (2009) 2507–2529.
[2] N. Aronszajn, Theory of Reproducing Kernels, *Transactions of the American Mathematical Society,* **68** (3) (1950), 337–404.
[3] A. Bonami and A. Karoui, Spectral decay of time and frequency limiting operator, *Appl. Comput. Harmonic Anal.,* **42** (2017), 1–20.

References

[1] A. Argyriou, C. A. Micchelli and M. Pontil, When is there a representer Theorem ? Vector versus matrix regularizers? *Journal of Machine Learning Research*, **10** (2009) 2507–2529.
[2] N. Aronszajn, Theory of Reproducing Kernels, *Transactions of the American Mathematical Society*, **68** (3) (1950), 337–404.
[3] A. Bonami and A. Karoui, Spectral decay of time and frequency limiting operator, *Appl. Comput. Harmonic Anal.*, **42** (2017), 1–20.
[4] A. Bonami and A. Karoui, Random Discretization of the Finite Fourier Transform and Related Kernel Random Matrices, preprint version 2 (2017), available at arXiv:1703.10459.

References

[1] A. Argyriou, C. A. Micchelli and M. Pontil, When is there a representer Theorem ? Vector versus matrix regularizers? *Journal of Machine Learning Research,* **10** (2009) 2507–2529.

[2] N. Aronszajn, Theory of Reproducing Kernels, *Transactions of the American Mathematical Society,* **68** (3) (1950), 337–404.

[3] A. Bonami and A. Karoui, Spectral decay of time and frequency limiting operator, *Appl. Comput. Harmonic Anal.,* **42** (2017), 1–20.

[4] A. Bonami and A. Karoui, Random Discretization of the Finite Fourier Transform and Related Kernel Random Matrices, preprint version 2 (2017), available at arXiv:1703.10459.

[5] B. Bousselmi, J. F. Dupuy and A. Karoui, Learning Schemes Via Reproducing Kernels, Spectral Approximation and Nonparametric Estimation of a Zero-Inflated Poisson Model, Work in progress (2019).

[6] P. Jaming, A. Karoui and S. Spektor, The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases, *J. Approx. Theory,* **212** (2016), 41–65.

[6] P. Jaming, A. Karoui and S. Spektor, The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases, *J. Approx. Theory,* **212** (2016), 41–65.
[7] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *Ann. Probab.,* **22** (1994), 1679–1706.

[6] P. Jaming, A. Karoui and S. Spektor, The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases, *J. Approx. Theory,* **212** (2016), 41–65.
[7] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *Ann. Probab.,* **22** (1994), 1679–1706.
[8] J. Shawe-Taylor, K. I. Williams, N. Cristianini and J. Kandola, On the Eigenspectrum of the Gram Matrix and the Generalized Error of Kernel-PCA, *IEEE Trans. Inf. Theory,* **51** no. 7, (2005), 2510–2522.

[6] P. Jaming, A. Karoui and S. Spektor, The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases, *J. Approx. Theory,* **212** (2016), 41–65.
[7] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *Ann. Probab.,* **22** (1994), 1679–1706.
[8] J. Shawe-Taylor, K. I. Williams, N. Cristianini and J. Kandola, On the Eigenspectrum of the Gram Matrix and the Generalized Error of Kernel-PCA, *IEEE Trans. Inf. Theory,* **51** no. 7, (2005), 2510–2522.
[9] D. Slepian, H. O. Pollak, H. Landau, Prolate spheroidal wave functions, Fourier analysis, and uncertainty-I, -II, -III, -IV, Bell Syst. Tech. J. (1961 - 1964).

[6] P. Jaming, A. Karoui and S. Spektor, The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases, *J. Approx. Theory,* **212** (2016), 41–65.
[7] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *Ann. Probab.,* **22** (1994), 1679–1706.
[8] J. Shawe-Taylor, K. I. Williams, N. Cristianini and J. Kandola, On the Eigenspectrum of the Gram Matrix and the Generalized Error of Kernel-PCA, *IEEE Trans. Inf. Theory,* **51** no. 7, (2005), 2510–2522.
[9] D. Slepian, H. O. Pollak, H. Landau, Prolate spheroidal wave functions, Fourier analysis, and uncertainty-I, -II, -III, -IV, Bell Syst. Tech. J. (1961 - 1964).
[10] S. Smale and D. X. Zhou, Learning Theory Estimates via Integral Operators and Their Approximations, Constructive Approximation, **26** (2) (2007), 153–172.