Reproducing kernel Hilbert C^*-module for data analysis

Yuka Hashimoto

NTT / RIKEN AIP

March 25th, 2024

- Y. Hashimoto, I. Ishikawa, M. Ikeda, F. Komura, T. Katsura, and Y. Kawahara, JMLR, 22(267):1–56. (updated version : arXiv:2101.11410v2)
- Y. Hashimoto, F. Komura, and M. Ikeda, Matrix and Operator Equations, pp. 1–27.
- Y. Hashimoto, M. Ikeda, and H. Kadri, NeurIPS 2023.

Yuka Hashimoto

NTT / RIKEN AIP

- 2018-2023 Researcher at NTT Network Service Systems Laboratories
- 2022 Received Ph.D. from Keio University
- 2022- Visiting researcher at RIKEN AIP
- 2023- Distinguished researcher at NTT Network Service Systems Laboratories / NTT Communication Science Laboratories

Backgrounds / Interests

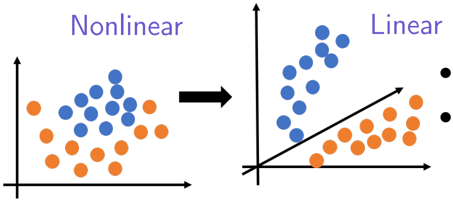
- Operator theoretic data analysis
- Kernel methods, neural networks
- Numerical linear algebra

Contents

Background: Kernel methods

$\mathcal{X}$
(Finite dimensional sp.)

RKHS[1]
(Infinite dimensional Hilbert sp.)



- Kernel PCA, Kernel SVM
- Learning complex-valued functions

Advantages of RKHS

- Nonlinearity in the original space is transformed into a linear one.
- We can compute inner products in RKHS exactly by computers.

[1]Schölkopf and Smola, MIT Press, Cambridge, 2001

Background: Reproducing kernel Hilbert space (RKHS)

Let $\mathcal{X}$ be a set. A map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is called a positive definite kernel if it satisfies:

1. $k(x, y) = \overline{k(y, x)}$ for $x, y \in \mathcal{X}$ and
2. $\sum_{t,s=1}^{n} \overline{c_t} k(x_t, x_s) c_s \geq 0$ for $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{C}$, $x_1, \ldots, x_n \in \mathcal{X}$.

$\phi(x) := k(\cdot, x)$ ($\phi : \mathcal{X} \to \mathbb{C}^{\mathcal{X}}$: feature map associated with k),

$$\mathcal{H}_{k,0} := \Big\{ \sum_{t=1}^{n} \phi(x_t) c_t \Big| \; n \in \mathbb{N},\; c_t \in \mathbb{C},\; x_t \in \mathcal{X} \Big\}. \tag{1}$$

We can define an inner product $\langle \cdot, \cdot \rangle_k : \mathcal{H}_{k,0} \times \mathcal{H}_{k,0} \to \mathbb{C}$ as

$$\Big\langle \sum_{s=1}^{n} \phi(x_s) c_s, \sum_{t=1}^{l} \phi(y_t) d_t \Big\rangle_k := \sum_{s=1}^{n} \sum_{t=1}^{l} \overline{c_s} k(x_s, y_t) d_t. \tag{2}$$

Reproducing property: $\langle \phi(x), v \rangle_k = v(x)$ for $v \in \mathcal{H}_k$ and $x \in \mathcal{X}$

RKHS $\mathcal{H}_k$: completion of $\mathcal{H}_{k,0}$

Background: Representer theorem in RKHSs

The representer theorem guarantees that solutions of a minimization problem are represented only with given samples[2].

$\mathcal{H}_k$: RKHS
$\mathbb{R}_+ := \{a \in \mathbb{R} \mid a \geq 0\}$

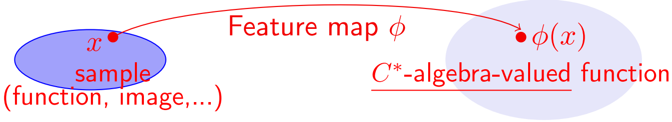
Theorem 1 Representer theorem in RKHSs

Let $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{C}$. Let $h : \mathcal{X} \times \mathbb{C}^2 \to \mathbb{R}_+$ be an error function and $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy $g(c) < g(d)$ for $c < d$. Then, any $u \in \mathcal{H}_k$ minimizing $\sum_{i=1}^n h(x_i, a_i, u(x_i)) + g(\|u\|_k)$ admits a representation of the form $\sum_{i=1}^n \phi(x_i) c_i$ for some $c_1, \ldots, c_n \in \mathbb{C}$.

The result can be applied to supervised problems.

[2]Schölkopf et al., COLT 2001.

Goal: Generalization of data analysis in RKHS to RKHM



$\mathcal{X}$
(Structured data sp.)

RKHM
(Infinite dimensional Hilbert C^*-module)

Nonlinear

Linear + C^*-algebra-valued inner product

Advantages of RKHM:

- C^*-algebra-valued inner products extract information of structures.

We constructed a framework of data analysis with RKHM.

- We can reconstruct existing RKHSs by using RKHMs.
- We have shown fundamental properties for data analysis in RKHMs (e.g. representer theorem, kernel mean embedding).

C^*-algebra and von Neumann-algebra

C^*-algebra : Banach space equipped with a product & an involution *
+ C^*-property

e.g.

- $C(\mathcal{Z})$ for a compact space $\mathcal{Z}$
  Norm : sup norm, **Product** : pointwise product,
  Involution : pointwise complex conjugate
- $\mathcal{K}(\mathcal{H}) = \{$compact operators on a Hilbert space $\mathcal{H}\}$
  Norm : operator norm, **Product** : composition, **Involution** : adjoint

Von Neumann-algebra : C^*-algebra that is closed in the strong operator topology

e.g.

- $L^\infty(\mathcal{Z})$ for a measure space $\mathcal{Z}$
- $\mathcal{B}(\mathcal{H}) = \{$bounded linear operators on a Hilbert space $\mathcal{H}\}$

Positivity and order in C^*-algebras

For optimization, we need the notion of "positive" and order.

$\mathcal{A}$: C^*-algebra

Definition 1 Positive

Let $a \in \mathcal{A}$. If $a = b^*b$ for some $b \in \mathcal{A}$, then a is called positive. We put $\mathcal{A}_+ = \{a \in \mathcal{A} \mid a \text{ is positive}\}$.

We can define a (partial) order $\leq_{\mathcal{A}}$ in $\mathcal{A}$ by
"$a \leq_{\mathcal{A}} b$ if and only if $b - a$ is positive".

We denote $a <_{\mathcal{A}} b$ if $b - a$ is positive and not zero.

We consider supremum, maximum, infimum, and minimum in $\mathcal{A}$ with respect to the order $\leq_{\mathcal{A}}$.

Hilbert C^*-module

$\mathcal{A}$: C^*-algebra
$\mathcal{M}$: right $\mathcal{A}$-module ($u \in \mathcal{M}$, $c \in \mathcal{A} \to uc \in \mathcal{M}$)

Definition 2 $\mathcal{A}$-valued inner product

A map $\langle \cdot, \cdot \rangle : \mathcal{M} \times \mathcal{M} \to \mathcal{A}$ is called an $\mathcal{A}$-valued inner product if it satisfies the following properties for $u, v, w \in \mathcal{M}$ and $c, d \in \mathcal{A}$:

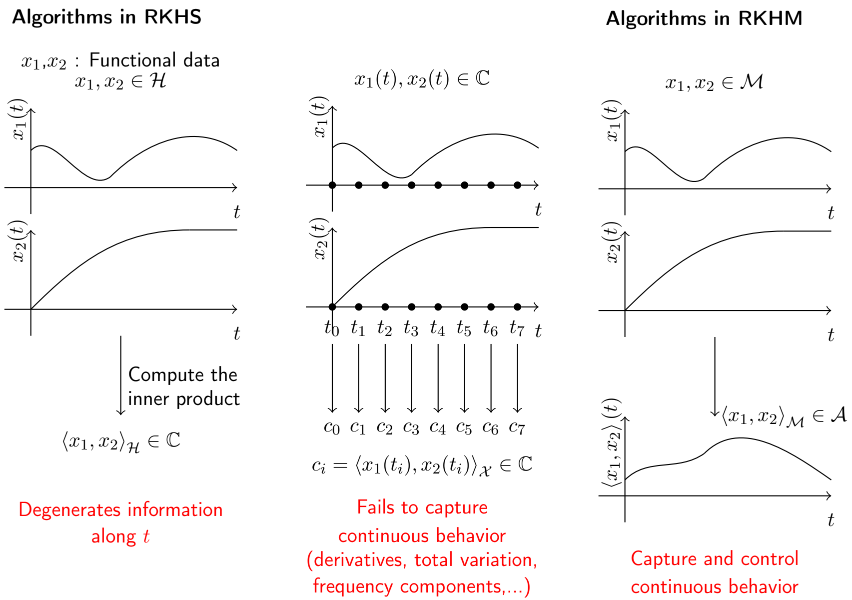
1. $\langle u, vc + wd \rangle = \langle u, v \rangle c + \langle u, w \rangle d$,
2. $\langle v, u \rangle = \langle u, v \rangle^*$,
3. $\langle u, u \rangle \geq 0$ (positive) and if $\langle u, u \rangle = 0$ then $u = 0$.

$\to$ *$\mathcal{A}$-valued absolute value* $|u| := \langle u, u \rangle^{1/2}$ $\to$ Norm $\|u\| := \| \langle u, u \rangle \|_{\mathcal{A}}^{1/2}$

Hilbert C^*-module $\mathcal{M}$[3]: complete $\mathcal{A}$-module equipped with an $\mathcal{A}$-valued inner-product

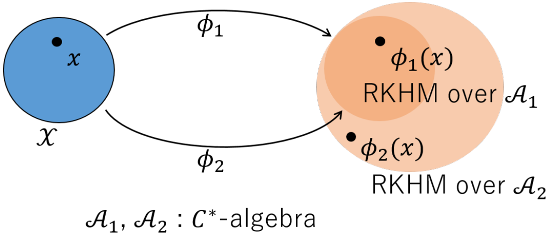
[3] Lance, Cambridge University Press, 1995.

Advantages of RKHM (functional data)



Advantages of RKHM

- Enlarge representation spaces using C^*-algebras (e.g. use the C^*-algebra of continuous functions for functional data).



- Make use of the product structure.
  e.g. polynomial kernel $k(x, y) = x^*y + x^*x^*yy$ $(x, y \in \mathcal{A}_1$ or $\mathcal{A}_2)$
- Use the operator norm to alleviate the dependency of the error on data dimension. (Explain later!)

Review of reproducing kernel Hilbert C^*-module

$\mathcal{A}$: C^*-algebra

RKHS ($\mathcal{H}_k$):

- $\mathbb{C}$-valued positive definite kernel k
- $\mathbb{C}$-valued functions
- $\mathbb{C}$-valued inner product

RKHM over $\mathcal{A}$ ($\mathcal{M}_k$):

- $\mathcal{A}$-valued positive definite kernel k
- $\mathcal{A}$-valued functions
- $\mathcal{A}$-valued inner product

Reproducing kernel Hilbert C^*-module (RKHM)

Let $\mathcal{X}$ be a set. A map $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is called an $\mathcal{A}$-valued positive definite kernel if it satisfies:

1. $k(x, y) = k(y, x)^*$ for $x, y \in \mathcal{X}$ and
2. $\sum_{t,s=1}^{n} c_t^* k(x_t, x_s) c_s \geq 0$ for $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathcal{A}$, $x_1, \ldots, x_n \in \mathcal{X}$.

$\phi(x) := k(\cdot, x)$ ($\phi : \mathcal{X} \to \mathcal{A}^{\mathcal{X}}$: feature map associated with k),

$$\mathcal{M}_{k,0} := \left\{ \sum_{t=1}^{n} \phi(x_t) c_t \,\middle|\, n \in \mathbb{N},\ c_t \in \mathcal{A},\ x_t \in \mathcal{X} \right\}. \tag{3}$$

We can define an $\mathcal{A}$-valued inner product $\langle \cdot, \cdot \rangle_k : \mathcal{M}_{k,0} \times \mathcal{M}_{k,0} \to \mathcal{A}$ as

$$\left\langle \sum_{s=1}^{n} \phi(x_s) c_s, \sum_{t=1}^{l} \phi(y_t) d_t \right\rangle_k := \sum_{s=1}^{n} \sum_{t=1}^{l} c_s^* k(x_s, y_t) d_t. \tag{4}$$

Reproducing property: $\langle \phi(x), v \rangle_k = v(x)$ for $v \in \mathcal{M}_k$ and $x \in \mathcal{X}$

RKHM $\mathcal{M}_k$: completion of $\mathcal{M}_{k,0}$

Orthonormality in Hilbert C^*-modules

To project a vector onto a finitely generated submodule, we introduce orthonormality.

Definition 3 Orthonormal

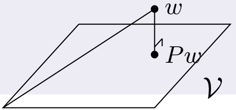
Let $\mathcal{M}$ be a Hilbert C^*-module.

1. A vector $q \in \mathcal{M}$ is said to be normalized if $0 \neq \langle q, q \rangle = \langle q, q \rangle^2$.
2. Two vectors $p, q \in \mathcal{M}$ are said to be orthogonal if $\langle p, q \rangle = 0$.

Theorem 2 Minimization property

Let $\mathcal{A}$ be a unital C^*-algebra and let $\mathcal{I}$ be a finite index set. Let $\mathcal{V}$ be the module spanned by an orthonormal system $\{q_i\}_{i \in \mathcal{I}}$ and let $P : \mathcal{M} \to \mathcal{V}$ be the projection operator. For $w \in \mathcal{M}$,

$$Pw = \arg\min_{v \in \mathcal{V}} |w - v|^2 \qquad (5)$$



Orthonormality in Hilbert C^*-modules (Proof)

Key point of the proof:
If $\mathcal{I}$ is finite, we can construct a projection onto $\mathcal{V}$.

Proof:
For $w \in \mathcal{M}$, define

$$Pw = \sum_{i \in \mathcal{I}} q_i \langle q_i, w \rangle_{\mathcal{M}}. \tag{6}$$

$P \colon \mathcal{M} \to \mathcal{M}$ is the orthogonal projection onto $\mathcal{V}$.
(Note: If $\mathcal{I}$ is infinite, the convergence is the strong convergence.)
Let $w \in \mathcal{M}$. For any $v \in \mathcal{V}$, we have

$$\begin{aligned} |w - v|_{\mathcal{M}}^2 &= |Pw + (I - P)w - v|_{\mathcal{M}}^2 \\ &= |Pw - v|_{\mathcal{M}}^2 + |(I - P)w|_{\mathcal{M}}^2 \geq |w - Pw|_{\mathcal{M}}^2. \end{aligned} \tag{7}$$

Thus, we have $|w - v|_{\mathcal{M}} \geq |w - Pw|_{\mathcal{M}}$.

Orthonormality in Hilbert C^*-modules (Proof)

Assume $v \in \mathcal{V}$ satisfies $|w - v|_{\mathcal{M}} = |w - Pw|_{\mathcal{M}}$. Since $v = Pv$ and $\langle w, Pw \rangle = \langle w, PPw \rangle = \langle Pw, Pw \rangle$, we have

$$
\begin{aligned}
|w - v|_{\mathcal{M}}^2 &= \langle w - v, w - v \rangle_{\mathcal{M}} \\
&= \langle w, w \rangle_{\mathcal{M}} - \langle w, v \rangle_{\mathcal{M}} - \langle v, w \rangle_{\mathcal{M}} + \langle v, v \rangle_{\mathcal{M}} \\
&= \langle w, w \rangle_{\mathcal{M}} - \langle w, Pv \rangle_{\mathcal{M}} - \langle Pv, w \rangle_{\mathcal{M}} + \langle v, v \rangle_{\mathcal{M}}, \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
|w - Pw|_{\mathcal{M}}^2 &= \langle w - Pw, w - Pw \rangle_{\mathcal{M}} \\
&= \langle w, w \rangle - \langle w, Pw \rangle - \langle Pw, w \rangle + \langle Pw, Pw \rangle \\
&= \langle w, w \rangle - \langle Pw, Pw \rangle - \langle Pw, Pw \rangle + \langle Pw, Pw \rangle . \qquad (9)
\end{aligned}
$$

Thus, we have

$$
\langle w, w \rangle_{\mathcal{M}} - \langle Pw, Pw \rangle_{\mathcal{M}} = \langle w, w \rangle_{\mathcal{M}} - \langle Pw, v \rangle_{\mathcal{M}} - \langle v, Pw \rangle_{\mathcal{M}} + \langle v, v \rangle_{\mathcal{M}} .
$$

Therefore, we have $|Pw - v|_{\mathcal{M}}^2 = 0$, which shows $Pw = v$.

Gram–Schmidt orthonormalization in Hilbert C^*-modules

$\mathcal{A}$: von Neumann-algebra

Proposition 1 Normalization

Let $\epsilon > 0$ and let $\hat{q} \in \mathcal{M}$ be a vector satisfying $\|\hat{q}\|_{\mathcal{M}} > \epsilon$. Then there exists $\hat{b} \in \mathcal{A}$ such that $\|\hat{b}\|_{\mathcal{A}} < 1/\epsilon$ and $q := \hat{q}\hat{b}$ is normalized. Moreover, there exists $b \in \mathcal{A}$ such that $\|\hat{q} - qb\|_{\mathcal{M}} \leq \epsilon$.

Proposition 2 Gram–Schmidt orthonormalization

Let $\{w_i\}_{i=1}^{\infty}$ be a sequence in $\mathcal{M}$. For $i = 1, 2, \ldots$ and $\epsilon > 0$, let

$$\hat{q}_j = w_j - \sum_{i=1}^{j-1} q_i \left\langle q_i, w_j \right\rangle_{\mathcal{M}}, \quad q_j = \hat{q}_j \hat{b}_j \quad \text{if } \|\hat{q}_j\|_{\mathcal{M}} > \epsilon, \tag{10}$$

$$q_j = 0 \quad \text{otherwise.} \tag{11}$$

Here, $\hat{b}_j$ is defined in the same manner as $\hat{b}$ in Proposition 1 by replacing $\hat{q}$ by $\hat{q}_j$. Then $\{q_j\}_{j=1}^{\infty}$ is an orthonormal system of $\mathcal{M}$. Moreover, any w_j is in the ϵ-neighborhood of the module generated by $\{q_j\}_{j=1}^{\infty}$.

Gram–Schmidt orthonormalization (Proof)

Key point of the proof:
If $\mathcal{A}$ is a von Neumann-algebra, we can apply the spectral decomposition.

Proof (Normalization):
$a := \langle \hat{q}, \hat{q} \rangle_{\mathcal{M}}$, $\sigma(a)$: spectrum of a
$a = \int_{\lambda \in \sigma(a)} \lambda \mathrm{d}E(\lambda)$: spectral decomposition of a
$\hat{b} := \int_{\lambda \in \sigma(a) \setminus B_{\epsilon^2}(0)} \lambda^{-1/2} \mathrm{d}E(\lambda) \in \mathcal{A}$, $B_\epsilon(0) := \{z \in \mathbb{C} \mid |z| \le \epsilon\}$
We have $\|\hat{b}\|_{\mathcal{A}} < 1/\epsilon$ and

$$\langle \hat{q}\hat{b}, \hat{q}\hat{b} \rangle_{\mathcal{M}} = \hat{b}^* a \hat{b} = \int_{\lambda \in \sigma(a) \setminus B_{\epsilon^2}(0)} \mathrm{d}E(\lambda).$$

Thus, $\langle \hat{q}\hat{b}, \hat{q}\hat{b} \rangle_{\mathcal{M}}$ is a nonzero orthogonal projection.

Gram–Schmidt orthonormalization (Proof)

$b := \int_{\lambda \in \sigma(a) \setminus B_{\epsilon^2}(0)} \lambda^{1/2} \mathrm{d}E(\lambda)$
Since $\hat{b}b = \int_{\lambda \in \sigma(a) \setminus B_{\epsilon^2}(0)} \mathrm{d}E(\lambda)$, we have

$$\langle \hat{q}, \hat{q}\hat{b}b \rangle = \langle \hat{q}, \hat{q} \rangle \hat{b}b = a\hat{b}b = \hat{b}ba\hat{b}b = \langle \hat{q}\hat{b}b, \hat{q}\hat{b}b \rangle \tag{12}$$

and obtain

$$\begin{aligned} \langle \hat{q} - qb, \hat{q} - qb \rangle_{\mathcal{M}} &= \langle \hat{q} - \hat{q}\hat{b}b, \hat{q} - \hat{q}\hat{b}b \rangle_{\mathcal{M}} = \langle \hat{q}, \hat{q} \rangle - \langle \hat{q}, \hat{q}\hat{b}b \rangle_{\mathcal{M}} \\ &= a(1_{\mathcal{A}} - \hat{b}b) = \int_{\lambda \in B_{\epsilon^2}(0)} \lambda \mathrm{d}E(\lambda). \end{aligned} \tag{13}$$

Thus, we have $\|\hat{q} - qb\|_{\mathcal{M}} \leq \epsilon$.

Representer theorem in RKHMs

To generalize complex-valued supervised problems to $\mathcal{A}$-valued ones, we show a representer theorem.

$\mathcal{M}_k$: RKHM over $\mathcal{A}$, $|\cdot|_k$: absolute value in $\mathcal{M}_k$
$\mathcal{A}_+ := \{a \in \mathcal{A} \mid {}^\exists b \in \mathcal{A} \text{ such that } a = b^*b\}$

Theorem 3 Representer theorem in RKHMs

Let $\mathcal{A}$ be a unital C^*-algebra, $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathcal{A}$. Let $h : \mathcal{X} \times \mathcal{A}^2 \to \mathcal{A}_+$ be an error function and $g : \mathcal{A}_+ \to \mathcal{A}_+$ satisfy $g(c) < g(d)$ for $c < d$. If $\mathrm{Span}_{\mathcal{A}}\{\phi(x_i)\}_{i=1}^n$ is closed, any $w \in \mathcal{M}_k$ minimizing $f(w) := \sum_{i=1}^n h(x_i, a_i, w(x_i)) + g(|w|_k)$ admits a representation of the form $\sum_{i=1}^n \phi(x_i)c_i$ for some $c_1, \ldots, c_n \in \mathcal{A}$.

Key point of the proof:
For a Hilbert C^*-module $\mathcal{M}$ over a unital C^*-algebra $\mathcal{A}$ and any finitely generated closed submodule $\mathcal{V}$ of $\mathcal{M}$, $w \in \mathcal{M}$ is decomposed into $w = w_1 + w_2$ where $w_1 \in \mathcal{V}$ and $w_2 \in \mathcal{V}^\perp$.

Approximate representer theorem in RKHMs

If $\mathcal{A}$ is a von Neumann algebra, we can show an approximate representer theorem under mild conditions.

Theorem 4 Approximate representer theorem in RKHMs

Let $\mathcal{A}$ be a von Neumann-algebra, $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathcal{A}$. Let $h : \mathcal{X} \times \mathcal{A}^2 \to \mathcal{A}_+$ be a Lipschitz continuous error function with Lipschitz constant L and $g : \mathcal{A}_+ \to \mathcal{A}_+$ satisfy $g(c) < g(d)$ for $c < d$. Assume $f(w) := \sum_{i=1}^n h(x_i, a_i, w(x_i)) + g(|w|_k)$ has a minimizer w. Then, for any $\epsilon > 0$, there exists $v \in \mathcal{M}_k$ of the form $\sum_{i=1}^n \phi(x_i) c_i$ such that $\|f(v) - f(w)\|_{\mathcal{A}} \leq Ln\epsilon \|w\|_{\mathcal{A}}$.

Key point of the proof:

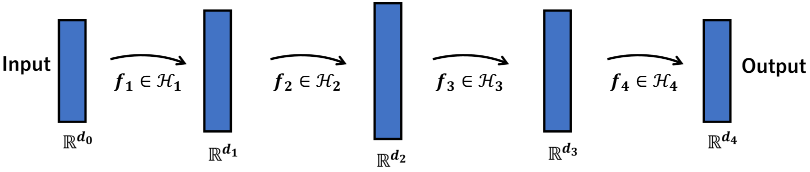
If $\mathcal{A}$ is a von Neumann-algebra, we can apply the Gram–Schmidt orthonormalization to construct a module approximating the module generated by $\{\phi(x_i)\}_{i=1}^n$.

Background: Deep learning with kernels

Combine the flexibility of deep neural networks with the representation power and solid theoretical understanding of kernel methods.

k_j : $\mathbb{R}^{d_j \times d_j}$-valued positive definite kernel
$\mathcal{H}_j$: vvRKHS a.w. k_j
$\mathcal{G}_j = \{f \in \mathcal{H}_j \mid \|f\|_{\mathcal{H}_j} \leq B_j\}$ $(j = 1, \ldots, L)$
$\mathcal{G}_L^{\mathrm{deep}} = \{f_L \circ \cdots \circ f_1 \mid f_j \in \mathcal{G}_j \; (j = 1, \ldots, L)\}$

$$\text{Deep RKHS} : f = f_1 \circ \cdots \circ f_L \tag{14}$$



Background: Perron–Frobenius operator on RKHM

$f : \mathcal{X} \to \mathcal{Y}$: nonlinear map
$\mathcal{M}_1$, $\mathcal{M}_2$: RKHMs on $\mathcal{X}$ and $\mathcal{Y}$ associated with feature maps ϕ_1 and ϕ_2

The Perron–Frobenius operator $P_f : \mathcal{M}_1 \to \mathcal{M}_2$ is an $\mathcal{A}$-linear operator satisfying

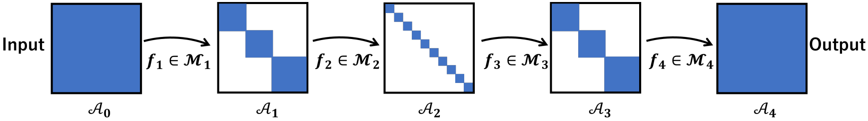
$$P_f \phi_1(x) = \phi_2(f(x)). \tag{15}$$

Remark
For the well-definedness of P_f, $\{\phi_1(x) \mid x \in \mathcal{X}\}$ should be $\mathcal{A}$-linearly independent.
(e.g. If $k = \tilde{k} I$ with a "good" $\mathbb{C}$-valued positive definite kernel $\tilde{k}$, the above condition is satisfied.)

Deep RKHM

$\mathcal{A} = \mathbb{C}^{d\times d}$, $\quad \mathcal{A}_j$: C^*-subalgebra of $\mathcal{A}$ $(j = 0, \ldots, L)$
k_j : $\mathcal{A}_j$-valued positive definite kernel
$\mathcal{M}_j$: RKHM a.w. k_j $(j = 1, \ldots, L)$
$\mathcal{F}_j = \{f \in \mathcal{M}_j \mid \|P_f\| \le B_j\}$ $(j = 1, \ldots, L-1)$
$\mathcal{F}_L = \{f \in \mathcal{M}_L \mid \|f\|_{\mathcal{M}_L} \le B_L\}$
$\mathcal{F}_L^{\mathrm{deep}} = \{f_L \circ \cdots \circ f_1 \mid f_j \in \mathcal{F}_j \ (j = 1, \ldots, L)\}$

$$\text{Deep RKHM} : f = f_L \circ \cdots \circ f_1 \in \mathcal{F}_L^{\mathrm{deep}} \tag{16}$$



Using the Perron–Frobenius operators and the reproducing property,

$$\begin{aligned} f(x) &= \langle \phi_L(f_{L-1} \circ \cdots \circ f_1(x)), f_L \rangle_{\mathcal{M}_L} \\ &= \left\langle P_{f_{L-1}} \cdots P_{f_1} \phi_L(x), f_L \right\rangle_{\mathcal{M}_L} \end{aligned} \tag{17}$$

Generalization bound with the operator norm

$\mathcal{G}(\mathcal{F}) := \{(x,y) \mapsto f(x) - y \mid f \in \mathcal{F},\ \|y\|_{\mathcal{A}} \le E\}$, n : number of samples

Theorem (Generalization bound)

Assume $^{\exists}D > 0$ s.t. $\|k_L(x,x)\|_{\mathcal{A}} \le D$ for any $x \in \mathcal{A}_{L-1}$. Let $\tilde{K} = 4\sqrt{2}(\sqrt{D}B_L + E)B_1 \cdots B_L$ ($B_1, \ldots, B_{L-1}$: norms of the Perron–Frobenius operators) and $\tilde{M} = 6(\sqrt{D}B_L + E)^2$. Then, for any $g \in \mathcal{G}(\mathcal{F}_L^{\mathrm{deep}})$ and any $\delta \in (0,1)$, with probability $\ge \delta$,

$$\|\mathrm{E}[|g(x,y)|_{\mathcal{A}}^2]\|_{\mathcal{A}} \le \Big\|\frac{1}{n}\sum_{i=1}^n |g(x_i,y_i)|_{\mathcal{A}}^2\Big\|_{\mathcal{A}} + \frac{\tilde{K}}{n}\Big(\sum_{i=1}^n \operatorname{tr} k_1(x_i,x_i)\Big)^{1/2} + \tilde{M}\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (18)$$

- Fix $p \in \mathbb{R}^d$ and upperbound $\|\mathrm{E}[|g(x,y)|_{\mathcal{A}}^2]^{1/2}p\|$ using the Rademacher complexity for vector-valued function spaces.
- Represent $f(x) = \left\langle P_{f_{L-1}} \cdots P_{f_1}\phi_L(x), f_L\right\rangle_{\mathcal{M}_L}$ and derive the product of the norms of the Perron–Frobenius operators.

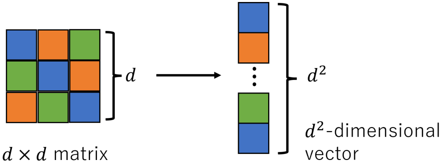
Comparison with vvRKHS

We can also flatten the matrices in $\mathcal{A} = \mathbb{C}^{d\times d}$ and regard them as d^2-dimensional vectors.
$\rightarrow$ Use vvRKHS to represent d^2-dimensional vector-valued functions.

In this case, the generalization bound is

$$\mathrm{E}[\|g(x,y)\|_{\mathrm{HS}}^2] \leq \frac{1}{n}\sum_{i=1}^n \|g(x_i,y_i)\|_{\mathrm{HS}}^2 + \frac{\tilde{K}}{n}\left(d\sum_{i=1}^n \operatorname{tr} k_1(x_i,x_i)\right)^{1/2} + \tilde{M}\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (19)$$

The operator norm alleviates the dependency of the generalization error on the output dimension.



Learning deep RKHM

$x_1, \ldots, x_n$: input training data
$y_1, \ldots, y_n$: output training data
Consider a minimization problem :

$$\min_{f_j \in \mathcal{M}_j} \left\| \frac{1}{n} \sum_{i=1}^{n} |f_L \circ \cdots \circ f_1(x_i) - y_i|_{\mathcal{A}}^2 \right\|_{\mathcal{A}} + \lambda_1 \| P_{f_{L-1}} \cdots P_{f_1}|_{\tilde{\mathcal{V}}(\mathbf{x})} \| + \lambda_2 \| f_L \|_{\mathcal{M}_L}, \qquad (20)$$

where $\tilde{\mathcal{V}}(\mathbf{x})$ is the Hilbert $\mathcal{A}$-module generated by $\phi_1(x_1), \ldots, \phi_1(x_n)$.

Proposition (Representer theorem)

A solution of the problem (20) is represented as $f_j = \sum_{i=1}^{n} \phi(x_i^{j-1}) c_{i,j}$ for some $c_{i,j} \in \mathcal{A}_j$, where $x_i^j = f_j \circ \cdots \circ f_1(x_i)$.

Connection with benign overfitting

$G_j \in \mathcal{A}^{n\times n}$: Gram matrix whose (i,l)-entry is $k_j(x_i^{j-1}, x_l^{j-1})$.

Proposition

We have

$$\|P_{f_{L-1}} \cdots P_{f_1}|_{\tilde{\mathcal{V}}(\mathbf{x})}\| = \|R_L^* G_L R_1\| \leq \|G_L^{-1}\|^{1/2}\|G_L\|\|G_1^{-1}\|^{1/2}. \quad (21)$$

To bound the norm of the Perron–Frobenius operator, we try to reduce $\|G_L^{-1}\|^{1/2}\|G_L\|$.

$\rightarrow$ Try to get the largest and the smallest eigenvalues of G_L closer.

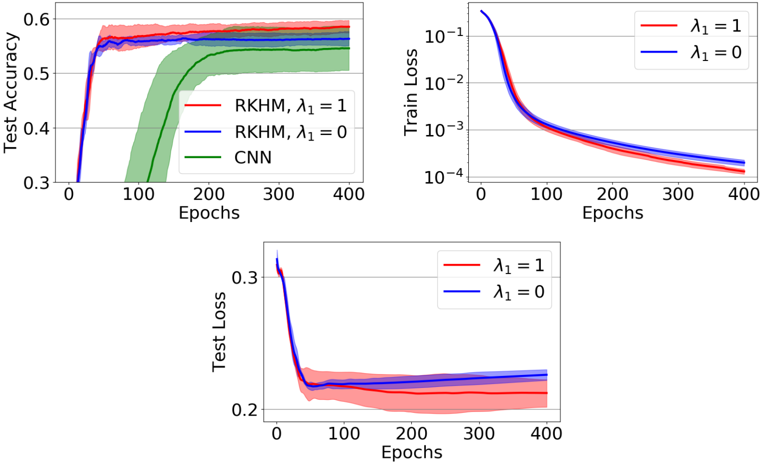
$\rightarrow$ According to the theory of overfitting for kernel regression[4], deep RKHM appreciates benign overfitting.

Benign overfitting: Networks predict well, even with a perfect fit to noisy training data. (Both training and test error decrease.)

[4] Mallinar et al., NeurIPS 2022

Numerical results

Classification task with MNIST with ($\lambda_1 = 1$) and without ($\lambda_1 = 0$) the Perron–Frobenius regularization, $d = 28$, $n = 20$, $L = 2$



Conclusion

- RKHM is a natural generalization of RKHS.
- We investigated properties related to the orthonormality in Hilbert C^*-modules.
- We showed a representer theorem and an approximate representer theorem in RKHMs and defined a kernel mean embedding in RKHMs.
- RKHMs are useful for analyzing image data and functional data.
- We proposed deep RKHM. We applied Perron–Frobenius operators and the operator norm to derive a generalization bound.