

Kernel Analog Forecasting and its applications

Reviewing the state of theoretical research and applying the techniques
to a range of datasets

by

Kieran John Taylor

Thesis for the

Masters of Mathematics

Under the supervision of Prof. Dr. Anke Pohl

at

UNIVERSITY OF BREMEN

ABSTRACT

This is where I will write my abstract.

CONTENTS

I	INTRODUCTION	I
2	KERNEL ANALOG FORECASTING	2
2.1	Forecasting Situation	2
2.1.1	Dynamical System	2
2.1.2	Koopman Operator	3
2.1.3	Sub Sigma-Algebra	4
2.1.4	Regression Function	5
2.2	Reproducing Kernel Hilbert Spaces	5
2.2.1	RKHS	6
2.2.2	Ideal Target Function	10
2.2.3	Moore-Penrose Pseudoinverse	12
2.3	Data-Driven Target Function	13
2.3.1	Supervised Learning Scenario	13
2.3.2	Empirical Error Minimization	14
2.3.3	Main Theory	15
	ACRONYMS	16
	GLOSSARY	17
	BIBLIOGRAPHY	18

I INTRODUCTION

We will first work through proving the main theorems of [\[1\]](#).

2 KERNEL ANALOG FORECASTING

2.1 FORECASTING SITUATION

2.1.1 DYNAMICAL SYSTEM

Let $(\Omega, \mathcal{F}, \mu)$ be a **Probability Space**, with Ω being the space of all possible states, \mathcal{F} a σ -algebra of distinguished subsets of Ω , and $\mu : \mathcal{F} \rightarrow [0, 1]$ a **Probability Measure**. We now define a **Dynamical System** on Ω through the semigroup of measurable maps: $\{\Phi^t : \Omega \rightarrow \Omega\}_{t \geq 0}$, which evolve any initial state $\omega_0 \in \Omega$ to $\omega_t \in \Omega$. While the semigroup property stipulates that we can split these maps $\Phi^{t_1+t_2} = \Phi^{t_1} \circ \Phi^{t_2}$, measurability ensures that for each $t \geq 0$ and $\forall A \in \mathcal{F}$

$$\Phi^{-t}A = (\Phi^t)^{-1}A := \{\omega \in \Omega : \Phi^t(\omega) \in A\} \in \mathcal{F}$$

Furthermore, we assume that these maps are also measure-preserving;

$$\forall B \in \mathcal{F}, t \geq 0, \mu(\Phi^{-t}B) = \mu(B)$$

Define the functional space $L^2(\mu)$ as the space of square integrable functions from $\Omega \rightarrow \mathbb{C}$, with respect to the probability measure μ . Using the standard representation, we then define $L^2(\mu)$ as the Hilbert Space of μ -a.e. equivalence functions in $L^2(\mu)$. The space is equipped with the inner product $\langle f_1, f_2 \rangle_{L^2(\mu)} = \int_{\Omega} \overline{f_1(\omega)} f_2(\omega) d\mu(\omega)$, which is conjugate-linear in the first argument.

Having defined our general probability space, we now move onto the forecasting task. We define a **Measurable Covariate Space** $(\mathcal{X}, \Sigma_{\mathcal{X}})$ as the tuple of a space, \mathcal{X} , which, while not necessarily linear, has the structure of a metric space, coupled with a Borel σ -algebra $\Sigma_{\mathcal{X}}$.

Furthermore, a **Measurable Response Space** $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ is a Hilbert Space over the complex numbers \mathcal{Y} , with a Borel σ -algebra $\Sigma_{\mathcal{Y}}$. In order to use the kernel trick, which is central for KAF, \mathcal{Y} is assumed to be \mathbb{C} , with $\Sigma_{\mathcal{Y}}$ being the Borel σ -algebra generated by the topology induced by open balls. The inner product $\langle y_1, y_2 \rangle_{\mathcal{Y}} = \overline{y_1} y_2$ is taken to be conjugate-linear in the first argument, as $L^2(\mu)$ above.

2 Kernel Analog Forecasting

For any $\omega \in \Omega$, we then define the data-producing measurable covariance function:

$$\begin{aligned} X : \Omega &\rightarrow \mathcal{X} \\ \omega &\mapsto x_\omega := X(\omega) \end{aligned}$$

and, similarly, the data-producing measurable response function:

$$\begin{aligned} Y : \Omega &\rightarrow \mathcal{Y} = \mathbb{C} \\ \omega &\mapsto y_\omega := Y(\omega) \end{aligned}$$

Our aim is, for any given **Lead Time** $\tau > 0$, to find a target function $f_\tau : \mathcal{X} \rightarrow \mathbb{C}$, such that $f_\tau \circ X$ approximates $Y \circ \Phi^\tau$. We use the norm induced by the inner-product of $L^2(\mu)$ to evaluate the accuracy of such a target function:

$$\|f_\tau \circ X - Y \circ \Phi^\tau\|_{L^2(\mu)}^2 = \int_{\Omega} \|(f_\tau \circ X_t)(\omega) - (Y \circ \Phi^\tau)(\omega)\|_{\mathbb{C}}^2 d\mu(\omega) \quad (1)$$

2.1.2 KOOPMAN OPERATOR

We are now presented with the problem that for any $t \geq 0$, Φ^t is not an intrinsically linear operator. We therefore use the **Koopman Operator** $U^t : L^2(\mu) \rightarrow L^2(\mu)$, named after the author of the paper in which it was first introduced [2]. The Koopman Operator allows us to express, for any $f \in L^2(\mu)$, the dynamical action in Ω as:

$$U^t f = f \circ \Phi^t$$

Unlike Φ^t , U^t is an intrinsically linear operator on $L^2(\mu)$, as $\forall a, b \in \mathbb{C}, f, g \in L^2(\mu)$

$$U^\tau(af + bg) = aU^\tau f + bU^\tau g$$

Lemma 2.1. *As the dynamical system is measure-preserving, the Koopman operator is a unitary operator on $L^2(\mu)$.*

Proof. TODO

□

As U^t are unitary operators for all $t \geq 0$, we can freely shift the starting point of our dynamical system ω_0 forwards by any arbitrary amount $\tau \geq 0$:

$$\begin{aligned}\|f_t \circ X - U^t Y\|_{L^2(\mu)}^2 &= \|U^\tau(f_t \circ X - U^t Y)\|_{L^2(\mu)}^2 \\ &= \|f_t \circ X \circ \Phi^\tau - U^t Y \circ \Phi^\tau\|_{L^2(\mu)}^2\end{aligned}$$

As such, the starting point ω_0 for our KAF predictions need not be fixed.

2.1.3 SUB SIGMA-ALGEBRA

Let us now define a measure μ_X on \mathcal{X} , as the pushforward of μ onto \mathcal{X} through X :

$$\forall S \in \Sigma_X : \mu_X(S) = \mu(X^{-1}S) \quad (2)$$

This is well-defined, as X is a μ -measurable function. We can therefore also define the Hilbert-Space $L^2(\mu_X)$ of equivalence classes of square-integrable functions from \mathcal{X} to \mathbb{C} . Furthermore, X also induces a Sub σ -Algebra $\mathcal{G} \subseteq \mathcal{F}$ on Ω , with $\mathcal{G} := X^{-1}(\Sigma_X)$. We now define $L_{\mathcal{G}}^2(\mu)$ (and subsequently $L_{\mathcal{G}}^2(\mu)$) as the set of square-integrable functions that take constant values on μ -measurable subsets of Ω where X is constant. We label such a function $g \in L_{\mathcal{G}}^2(\mu)$ as being **\mathcal{G} -measurable** and

$$\forall T \in \mathcal{F} \setminus \mathcal{G}, g(T) = 0$$

Lemma 2.2. *For any \mathcal{G} -measurable function $g \in L_{\mathcal{G}}^2(\mu)$, there exists an $f : \mathcal{X} \rightarrow \mathcal{Y}$ s.t. $g = f \circ X$.*

Proof. TODO □

We can therefore define an **Isometric Embedding** with range $L_{\mathcal{G}}^2(\mu)$:

$$\begin{aligned}\Xi : L^2(\mu_X) &\rightarrow L^2(\mu) \\ f &\mapsto f \circ X\end{aligned}$$

We also define the accompanying adjoint operator:

$$\Xi^* : L^2(\mu) \rightarrow L^2(\mu_X)$$

Such that $\forall g \in L_{\mathcal{G}}^2(\mu), g := f \circ X, \Xi^*(g) = f$ and $\ker \Xi^* = (L_{\mathcal{G}}^2(\mu))^\perp$

2.1.4 REGRESSION FUNCTION

Lemma 2.3. For $U^\tau \in L^2(\mu)$ there exists a unique \mathcal{G} -measurable element $Z_\tau \circ X \in L^2(\mu)$ such that for all $g \in L^2_{\mathcal{G}}(\mu)$

$$\langle g, U^\tau Y \rangle_{L^2(\mu)} = \langle g, Z_\tau \circ X \rangle_{L^2(\mu)} = \int_{\Omega} \overline{g(\omega)} \cdot (Z_\tau \circ X)(\omega) d\mu(\omega)$$

Proof. TODO

- Paper says to use Radon-Nikodym Theorem □

We define the **Orthogonal Projection** $\Pi_{\mathcal{G}} : L^2(\mu) \rightarrow L^2(\mu)$ as the mapping into the closed and convex subspace $L^2_{\mathcal{G}} \subseteq L^2(\mu)$. We can therefore also observe the unique element $Z_\tau \circ X$, which we label the **Conditional Expectation**, as the orthogonal projection of $U^\tau Y$ onto $L^2_{\mathcal{G}}(\mu)$.

$$Z_\tau \circ X = \Pi_{\mathcal{G}} U^\tau Y \tag{3}$$

Furthermore, as $Z_\tau \circ X \in L^2_{\mathcal{G}}$, then $\Xi Z_\tau = Z_\tau \circ X$. Consequently, the adjoint of the isometric embedding Ξ^* defines a unique $Z_\tau \in L^2(\mu_X)$, which from now on will be known as the **Regression Function**:

$$Z_\tau = \Xi^* U^\tau Y \tag{4}$$

The regression function is the unique element that minimizes the Mean Squared Error in Equation 1:

$$\arg \min_{f \in L^2_{\mathcal{G}}(\mu)} \|f \circ X - U^\tau Y\|_{L^2(\mu)} = Z_\tau$$

We define the **Intrinsic System Error** $\sigma_\tau = \|Z_\tau \circ X - U^\tau Y\|_{L^2(\mu)}$ as the error that is inherent to our system, or the amount of information that is lost due to our choice of covariate and response functions. This error will always appear and is independent of our search for suitable forecasting functions. Consequently, our task can be formulated as trying to produce functions that consistently approximate the regression function Z_τ .

2.2 REPRODUCING KERNEL HILBERT SPACES

In the following section we define and construct a subspace of functions $H \subseteq L^2(\mu)$ with a convenient structure to simplify the search for optimal approximations of the regression function. We must first, however, define the **Quotient Map**:

$$q : L^2(\mu_X) \rightarrow L^2(\mu_X)$$

2 Kernel Analog Forecasting

This linear map associates a function f to a representative of its equivalence class $qf \in [f]$. We may now use this to adapt the mean squared error (Equation 1) to the **Generalization Error** $\mathcal{E}_\tau : \mathbb{L}^2(\mu_X) \rightarrow \mathbb{R}$:

$$\mathcal{E}_\tau(f) = \|qf \circ X - U^\tau Y\|_{L^2(\mu)}^2 \quad (5)$$

To ease the formulation of the following lemma, let us define the **Excess Generalization Error** as:

$$\mathcal{A}_\tau(f) = \|qf - Z_\tau\|_{L^2(\mu_X)}^2 \quad (6)$$

Lemma 2.4. *We can decompose the generalization error into the excess generalization error and the intrinsic system error:*

$$\mathcal{E}_\tau(f) = \mathcal{A}_\tau(f) + \sigma_\tau$$

Proof. TODO □

As the intrinsic system error is independent of our prediction function f , the task of minimizing \mathcal{E}_τ is equivalent to minimizing \mathcal{A}_τ .

2.2.1 RKHS

Definition 2.5 (RKHS). *For each $x \in \mathcal{X}$, let $L_x : \mathcal{H} \rightarrow \mathbb{C}$ be the evaluation function defined by $L_x f = f(x)$. The space \mathcal{H} is an RKHS if L_x is bounded at every $x \in \mathcal{X}$.*

Theorem 2.6 (Riesz Representation Theorem). *For every $x \in \mathcal{X}$, there exists some function $k_x \in \mathcal{H}$ such that $\forall f \in \mathcal{H}$*

$$L_x f = \langle k_x, f \rangle_{\mathcal{H}} = f(x)$$

Proof. BLACK BOX □

Due to this reproducing property of \mathcal{H} , we are able to define the **Reproducing Kernel** function:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{C} \\ (x_1, x_2) &\mapsto \langle k_{x_1}, k_{x_2} \rangle_{\mathcal{H}} \end{aligned}$$

Lemma 2.7. *This reproducing kernel function k is*

- (i) *Conjugate Symmetric*
- (ii) *Positive Definite*

Proof. (i) $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}} = \overline{\langle k_y, k_x \rangle_{\mathcal{H}}} = \overline{k(y, x)}$

(ii) TODO □

Let us define the quotient embedding of the RKHS into $L^2(\mu_X)$, similarly to how we defined the quotient map above, as the identity mapping of a function in \mathcal{H} to a representative of its equivalence class in $L^2(\mu_X)$:

$$\begin{aligned} \iota : \mathcal{H} &\rightarrow L^2(\mu_X) \\ f &\mapsto \iota f \end{aligned}$$

We assume that this operator is continuous, meaning that every $f \in \mathcal{H}$ maps to exactly one representative of an equivalence class in $K := \iota\mathcal{H} \subseteq L^2(\mu_X)$. The adjoint of the identity embedding then immediately presents the utility of the RKHS method, as we observe that $\iota^* : L^2(\mu_X) \rightarrow \mathcal{H}$ represents the integral operator:

$$\begin{aligned} \iota^* f(x) &= \langle k_x, \iota^* f \rangle_{\mathcal{H}} \\ &= \langle \iota k_x, f \rangle_{L^2(\mu_X)} \\ &= \int_{\mathcal{X}} \overline{k(x, y)} f(y) d\mu_X(y) \\ &= \int_{\mathcal{X}} k(y, x) f(y) d\mu_X(y) \end{aligned}$$

Next, let us define the following linear operator on \mathcal{H}

$$\begin{aligned} T_k^{\mathcal{H}} : \mathcal{H} &\rightarrow \mathcal{H} \\ f &\mapsto \iota^* \iota f \end{aligned}$$

and similarly the following on $L^2(\mu_X)$:

$$\begin{aligned} T_k^{L^2} : L^2(\mu_X) &\rightarrow L^2(\mu_X) \\ f &\mapsto \iota^* \iota f \end{aligned}$$

Both of these operators are continuous, positive semi-definite, self-adjoint operators, which, if we assume that $k \in L^2(\mathcal{X} \times \mathcal{X}, \mu_X \otimes \mu_X)$, are then also compact [3].

Lemma 2.8 (Spectral Theorem for Compact, Self-Adjoint operators). *There exists an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ of $L^2(\mu_X)$ consisting of the eigenfunctions of $T_k^{L^2}$, with non-negative corresponding eigenvalues λ_i .*

Proof. BLACK BOX □

Theorem 2.9. Due to Lemma 2.8, $T_k^{L^2}$ has an orthonormal basis of eigenfunctions $\{\phi_i\}_i \subset L^2(\mu_X)$ with corresponding non-negative eigenvalues $\{\lambda_i\}$. Furthermore, the reproducing kernel can now be expressed through the series expansion:

$$k(x, y) = \sum_{i: \lambda_i > 0} \lambda_i \overline{\phi_i(x)} \phi_i(y)$$

with convergence being absolute and uniform on $L^2(\mathcal{X} \times \mathcal{X})$.

Proof. TODO or BLACK BOX □

So, we know that $\forall i : \lambda_i > 0, T_k^{L^2, L^2} \phi_i = \lambda_i \phi_i$ and we now wish to form an orthonormal basis $\{\psi_i\}_i$ of \mathcal{H} , such that $T_k^{\mathcal{H}} \psi_i = \lambda_i \psi_i$.

Lemma 2.10. The set $\left\{ \psi_i : \psi_i = \iota^* \frac{\phi_i}{\sqrt{\lambda_i}} \right\}_i$ are eigenfunctions of $T_k^{\mathcal{H}}$ with corresponding eigenvalues λ_i .

Proof. For any arbitrary i , such that $\lambda_i > 0$, let us observe $\psi_i(\cdot) = \iota^* \frac{\phi_i(\cdot)}{\sqrt{\lambda_i}} = \int_{\mathcal{X}} k(x, \cdot) \frac{\phi_i(x)}{\sqrt{\lambda_i}} d\mu_X(x)$.

We can see that

$$\begin{aligned}
 T_k^{\mathcal{K}} \psi_i(x) &= \langle k_x, T_k^{\mathcal{K}} \psi_i \rangle_{\mathcal{H}} \\
 &= \langle k_x, \iota^* \psi_i \rangle_{\mathcal{H}} \\
 &= \langle \iota k_x, \psi_i \rangle_{L^2(\mu_X)} \\
 &= \left\langle \iota k_x, \iota^* \frac{\phi_i}{\sqrt{\lambda_i}} \right\rangle_{L^2(\mu_X)} \\
 &= \left\langle \iota k_x, T_k^{L^2} \frac{\phi_i}{\sqrt{\lambda_i}} \right\rangle_{L^2(\mu_X)} \\
 &= \left\langle \iota k_x, \lambda_i \frac{\phi_i}{\sqrt{\lambda_i}} \right\rangle_{L^2(\mu_X)} \\
 &= \langle \iota k_x, \sqrt{\lambda_i} \phi_i \rangle_{L^2(\mu_X)} \\
 &= \int_X \overline{k(x, y)} \sqrt{\lambda_i} \phi_i(y) d\mu_X(y) \\
 &= \int_X k(y, x) \sqrt{\lambda_i} \phi_i(y) d\mu_X(y) \\
 &= \sqrt{\lambda_i} \int_X k(y, x) \phi_i(y) d\mu_X(y) \\
 &= \lambda_i \int_X k(y, x) \frac{\phi_i(y)}{\sqrt{\lambda_i}} d\mu_X(y) \\
 &= \lambda_i \psi_i(x)
 \end{aligned}$$

So, $\{\psi_i\}_i$ are eigenfunctions of $T_k^{\mathcal{K}}$ with corresponding eigenvalues λ_i . □

Theorem 2.11. *The eigenfunctions $\{\psi_i\}_i$ form an orthonormal basis of \mathcal{H}*

Proof. The eigenfunctions are orthonormal as

$$\begin{aligned}
 \langle \psi_i, \psi_j \rangle_{\mathcal{H}} &= \left\langle \iota^* \frac{\phi_i}{\sqrt{\lambda_i}}, \iota^* \frac{\phi_j}{\sqrt{\lambda_j}} \right\rangle_{\mathcal{H}} \\
 &= \left\langle \frac{\phi_i}{\sqrt{\lambda_i}}, u^* \frac{\phi_j}{\sqrt{\lambda_j}} \right\rangle_{L^2} \\
 &= \left\langle \frac{\phi_i}{\sqrt{\lambda_i}}, T_k^{L^2, L^2} \frac{\phi_j}{\sqrt{\lambda_j}} \right\rangle_{L^2} \\
 &= \left\langle \frac{\phi_i}{\sqrt{\lambda_i}}, \sqrt{\lambda_j} \phi_j \right\rangle_{L^2} \\
 &= \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}} \langle \phi_i, \phi_j \rangle_{L^2} \\
 &= \delta_{i,j}
 \end{aligned}$$

Again due to Lemma 2.8, $\{\psi_i\}_i$ form an orthonormal basis of \mathcal{H} . □

Now, we order the eigenvalues in decreasing order, such that, due to the compactness of $T_k^{L^2}$, the sequence $\lambda_1, \lambda_2, \dots$ accumulates only at 0.

For any $l > 0$, such that $\lambda_l > 0$, we define the l -dimensional **hypothesis space** $\mathcal{H}_l \subseteq \mathcal{H}$ as:

$$\mathcal{H}_l = \text{span}\{\phi_1, \dots, \phi_l\} \tag{7}$$

2.2.2 IDEAL TARGET FUNCTION

We now aim to find the minimizer of the error within this hypothesis space. Let us first define the restricted quotient map on \mathcal{H}_l :

$$\begin{aligned}
 \iota|_{\mathcal{H}_l} : \mathcal{H}_l &\rightarrow L^2(\mu_X) \\
 f &\mapsto \iota|_{\mathcal{H}_l} f
 \end{aligned}$$

Consequently, the image of \mathcal{H}_l under $L^2(\mu_X)$ inclusion can be defined as $H_l := \iota|_{\mathcal{H}_l} \mathcal{H}_l \subseteq L^2(\mu_X)$.

Lemma 2.12. $H_l \subseteq L^2(\mu_X)$ is closed and convex.

Proof. TODO □

As H_l is closed and convex, we can define the orthogonal projection

$$\Pi_l : L^2(\mu_X) \rightarrow L^2(\mu_X)$$

that maps into H_l . This allows us to further deconstruct the excess generalization error \mathcal{A}_τ defined in Equation 6 for any $f \in \mathcal{H}_l$.

$$\mathcal{A}_\tau(f) = \|\iota|_{\mathcal{H}_l} f - \Pi_l Z_\tau\|_{L^2(\mu_X)}^2 + \|(I - \Pi_l)Z_\tau\|_{L^2(\mu_X)}^2 \quad (8)$$

The task of finding the minimizer of the error within the l -dimensional hypothesis space \mathcal{H}_l is therefore reduced to minimizing $\|qf - \Pi_l Z_\tau\|_{L^2(\mu_X)}^2$. If $\iota|_{\mathcal{H}_l}$ is injective, then it is invertible, and there exists a unique minimizer in \mathcal{H}_l , which we call the **Ideal Target Function**, as it is independent of training data:

$$f_{\tau,l} := (\iota|_{\mathcal{H}_l})^{-1} \Pi_l Z_\tau \quad (9)$$

The remaining error of such an ideal target function would then be

$$\mathcal{E}_\tau(f_{\tau,l}) = \sigma_\tau + \|(I - \Pi_l)Z_\tau\|_{L^2(\mu_X)}^2$$

Lemma 2.13. *When is $\iota|_{\mathcal{H}_l}$ injective?*

Proof. TODO □

By defining for each i

$$\alpha_i(\tau) := \langle \phi_i \circ X, U^\tau Y \rangle_{L^2(\mu)} \quad (10)$$

we are able to rewrite $Z_\tau \in L^2(\mu_X)$ in terms of the orthonormal basis $\{\phi_i\}_i$ of $L^2(\mu_X)$:

$$Z_\tau = \sum_i \alpha_i(\tau) \phi_i \quad (11)$$

We can therefore reform our ideal target function in terms of the orthonormal basis $\{\psi_i\}_i$ of \mathcal{K} :

$$\begin{aligned} f_{\tau,l} &= (\iota|_{\mathcal{H}_l})^{-1} \Pi_l Z_\tau \\ &= (\iota|_{\mathcal{H}_l})^{-1} \Pi_l \sum_i \alpha_i(\tau) \phi_i \\ &= (\iota|_{\mathcal{H}_l})^{-1} \sum_{i=1}^l \alpha_i(\tau) \phi_i \\ &= \sum_{i=1}^l \frac{\alpha_i(\tau)}{\sqrt{\lambda_i}} \psi_i \end{aligned}$$

2.2.3 MOORE-PENROSE PSEUDOINVERSE

We define $T := (\iota|_{\mathcal{H}_l})^{-1} \Pi_l$ as the pseudoinverse of $\iota|_{\mathcal{H}_l}$.

T reduces to the **Moore-Penrose Pseudoinverse** ι^+ if \mathcal{H} is a Hilbert Space.

USE LEMMA 5 and LEMMA 6 FROM PAPER TO PROVE FOLLOWING THEOREM.

Theorem 2.14 (Dense Image). *K is dense in $L^2(\mu_X)$ if, and only if, $T_k^{L^2}$ is a strictly-positive operator*

Proof. TODO □

Theorem 2.15. *Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ be an $L^2(\mu_X)$ -strictly-positive kernel with corresponding RKHS \mathcal{K} . Then, for any response variable $U^\tau Y \in L^2(\mu)$ and lead time $\tau \geq 0$, as $l \rightarrow \infty$, the target functions $f_{\tau,l}$ converge to the conditional expectation $Z_\tau \circ X$, in the sense of the excess generalization error (Equation 5)*

$$\lim_{l \rightarrow \infty} \|q f_{\tau,l} - Z_\tau\|_{L^2(\mu_X)}^2 = 0.$$

Convergence with respect to the (stronger) RKHS \mathcal{K} norm is obtained if $Z_\tau \in K$

Proof. If, and only if, K is a dense subspace of $L^2(\mu_X)$, then we have that

$$\mathcal{A}_\tau(f_{\tau,l}) = \sum_{i=l+1}^{\infty} |\alpha_i(\tau)|^2$$

TODO □

2.3 DATA-DRIVEN TARGET FUNCTION

2.3.1 SUPERVISED LEARNING SCENARIO

We now turn to a situation in which we have access to a training dataset consisting of pairs:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

where $x_i = X(\omega_i)$ and $y_i = Y(\omega_i)$ on an unknown collection of points $\omega_1, \dots, \omega_n \in \Omega$.

These ω_i are on a single trajectory, and at a fixed sampling interval $\Delta t > 0$:

$$\omega_i = \Phi^{t_i}(\omega_1), t_i = (i - 1)\Delta t \quad (12)$$

Unlike in well-known methods such as kernel SVMs, KAF does not use the *kernel trick*, instead running a more computationally costly method to compute the orthogonal eigenfunctions. This has the benefit of being able to control the parameter $l \ll n$ to control and prevent overfitting to the training data. For any training dataset, we define an **Empirical Probability Measure** on the σ -algebra \mathcal{F} of Ω :

$$\begin{aligned} \mu_n : \mathcal{F} &\rightarrow [0, 1] \\ \omega &\mapsto \sum_{i=1}^n \frac{\delta_{\omega_i}(\omega)}{n} \end{aligned}$$

Where δ_{ω_i} is the Dirac-delta measure on $\{\omega_i\} \in \Omega$. Similarly, we can define an **Empirical Covariant Probability Measure** on $\Sigma_{\mathcal{X}}$:

$$\begin{aligned} \mu_{X,n} : \Sigma_{\mathcal{X}} &\rightarrow [0, 1] \\ x &\mapsto \sum_{i=1}^n \frac{\delta_{x_i}(x)}{n} \end{aligned}$$

Based on these new finite measures, we can define the **Empirical Hilbert Spaces** $L^2(\mu_n)$ and $L^2(\mu_{X,n})$. $L^2(\mu_n)$ consists of equivalence classes of complex-valued, measurable functions on Ω that have common values at ω_i . It therefore has, at most, a dimension of n (less if the points ω_i are not unique) and can therefore be embedded into \mathbb{C}^n . Any $f \in L^2(\mu_n)$ can be mapped to its $L^2(\mu_n)$ equivalence class by q_n :

$$\begin{aligned} q_n : L^2(\mu_n) &\rightarrow L^2(\mu_n) \\ f(\omega_1) &\mapsto \mathbf{f}(\omega_1) := [f(\omega_1), \dots, f(\omega_n)] \end{aligned}$$

with the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle = \frac{\mathbf{f} \cdot \mathbf{g}}{n}.$$

The same can be said of $L^2(\mu_{X,n})$, with operators on $L^2(\mu_n)$ and $L^2(\mu_{X,n})$ being $n \times n$ complex matrices.

We then define $L^2_{\mathcal{G}}(\mu_n)$ as the set of square-integrable functions that take constant values on μ_n -measurable subsets of Ω where X is constant. Again as in the previous section, we define the **Empirical Isometric Embedding** with range $L^2_{\mathcal{G}}(\mu_n)$ through the following covariate map:

$$\begin{aligned} \Xi_n : L^2(\mu_{X,n}) &\rightarrow L^2(\mu_n) \\ f &\mapsto f \circ X \end{aligned}$$

As before, let $\Pi_{\mathcal{G},n} : L^2(\mu_n) \rightarrow L^2(\mu_n)$ be the orthogonal projection into the closed and convex $L^2_{\mathcal{G}}(\mu_n) \subset L^2(\mu_n)$. Note that, when the training data samples are distinct (the typical situation), then $\Pi_{X,n}$ is the identity and Ξ_n is unitary, even if X is non-injective on sets of positive μ -measure.

We now wish to define a Koopman Operator on $L^2(\mu_n)$, however the composition operator with respect to dynamic flow does not lift to an operator on the function equivalence classes. This is because Φ^τ does not preserve null sets with respect to μ_n , meaning that two functions $f, f' : \Omega \rightarrow \mathbb{C}$ that lie in the same $L^2(\mu_n)$ equivalence class may lie in different equivalence classes after a certain lead time; $[f \circ \Phi^t]_{L^2(\mu_n)} \neq [f' \circ \Phi^t]_{L^2(\mu_n)}$.

Using the notation that \mathbf{f}_i is the i 'th element of $\mathbf{f} \in L^2(\mu_n)$, for any $p > 0$, we define the following **Shift Operator** U_n^p , which preserves null sets through the filling of the shifted vector with 0s:

$$\begin{aligned} U_n^p : L^2(\mu_n) &\rightarrow L^2(\mu_n) \\ \mathbf{f}_i &\mapsto \begin{cases} \mathbf{f}_{i+p}, & \text{if } i + p \leq n \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Subsequently, for $\tau = p\Delta t$, we define the shifted vector $\mathbf{y}_\tau \in \mathbb{C}^n$ as the **Analog Vector**:

$$\mathbf{y}_\tau = U_n^p q_n Y = [\mathbf{y}_{1+p}, \dots, \mathbf{y}_n, 0, \dots, 0]$$

2.3.2 EMPIRICAL ERROR MINIMIZATION

With this previous definition, we observe the following **Empirical Generalization Error** $\mathcal{E}_{\tau,n} : L^2(\mu_n) \rightarrow \mathbb{R}$:

$$\mathcal{E}_{\tau,n}(f) := \|q_n f \circ X - U_n^p q_n Y\|_{L^2(\mu_n)} \quad (13)$$

Analogously to the previous chapter, we know that there is a unique $Z_{\tau,n} \in L^2(\mu_{X,n})$ that satisfies $Z_{\tau,n} = \arg \min_{g \in L^2(\mu_{X,n})} \|g \circ X - U_n^p q_n Y\|_{L^2(\mu_n)}$. We can then also split the empirical generalization error

$$\mathcal{E}_{\tau,n}(f) = \mathcal{A}_{\tau,n}(f) + \sigma_{\tau,n}$$

into the **Empirical Excess Generalization Error**

$$\mathcal{A}_{\tau,n}(f) := \|q_n f - Z_{\tau,n}\|_{L^2(\mu_{X,n})}^2$$

and the **Empirical Intrinsic System Error**

$$\sigma_{\tau,n} := \|Z_{\tau,n} \circ X - U_n^p q_n Y\|_{L^2(\mu_n)}^2$$

2.3.3 MAIN THEORY

Theorem 2.16 (Uniform Convergence of KAF Target Function). *Under the basic assumptions for convergence, for every lead time $\tau = q\Delta t, q \in \mathbb{N}_0$, and hypothesis space dimension l such that $\lambda_l > 0$, the KAF target function $f_{\tau,l,n} \in \mathcal{K}_n$ converges as $n \rightarrow \infty$ to the ideal target function $f_{\tau,l} \in \mathcal{K}$, uniformly on \mathcal{U} . Moreover, if the reproducing kernel k of \mathcal{K} is $L^2(\mu_X)$ -strictly-positive-definite, then by Theorem 2.15, $f_{\tau,l,n}$ converges to the regression function Z_τ associated with the conditional expectation, $Z_\tau \circ X$, in the sense of the iterated limit*

$$\lim_{l \rightarrow \infty} \lim_{n \rightarrow \infty} f_{\tau,l,n} = \lim_{l \rightarrow \infty} f_{\tau,l} = Z_\tau$$

Here, the $n \rightarrow \infty$ and $l \rightarrow \infty$ limits are taken in $C(\mathcal{U})$ and $L^2(\mu_X)$ norm, respectively. Moreover, the convergence is uniform with respect to τ lying in compact sets.

Proof. TODO

□

ACRONYMS

KAF	Kernel Analog Forecasting
RKHS	Reproducing Kernel Hilbert Space

GLOSSARY

\mathbb{R}	The set of real numbers
$(\Omega, \mathcal{F}, \mu)$	Probability space
Ω	Space of all possible initial states
\mathcal{F}	σ -algebra of distinguished subsets of Ω
$\mu : \mathcal{F} \rightarrow \infty$	Probability Measure

BIBLIOGRAPHY

1. R. Alexander and D. Giannakis. “Operator-Theoretic Framework for Forecasting Nonlinear Time Series with Kernel Analog Techniques”. *Physica D: Nonlinear Phenomena* 409, 2020, p. 132520. ISSN: 0167-2789. DOI: <https://doi.org/10.1016/j.physd.2020.132520>. URL: <https://www.sciencedirect.com/science/article/pii/S016727891930377X>.
2. B. O. Koopman. “Hamiltonian Systems and Transformation in Hilbert Space”. *Proceedings of the National Academy of Sciences* 17:5, 1931, pp. 315–318. DOI: [10.1073/pnas.17.5.315](https://doi.org/10.1073/pnas.17.5.315). eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.17.5.315>. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.17.5.315>.
3. M. Wnuk. *A short note on compact embeddings of reproducing kernel Hilbert spaces in L^2 for infinite-variate function approximation*. 2022. arXiv: [2206.07432](https://arxiv.org/abs/2206.07432) [math.FA]. URL: <https://arxiv.org/abs/2206.07432>.