A Theoretical Cuntz-Algebraic Kernel Framework for Variable-Length Structured Data

PPMI FMIPA ITB 2026 Penelitian dan Inovasi KK

Intan Muchtadi-Alamsyah [1], Nurul Syafithri [1], Fajar Yuliawan [1], Mohammad Shane Alam [2]

[1] Algebra Research Group, FMIPA ITB

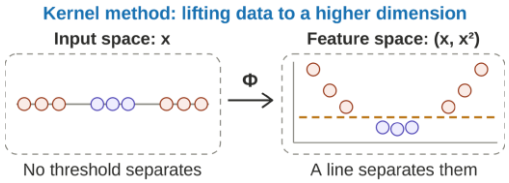
[2] Department of Mathematics, Aligarh Muslim University

Abstract

We construct positive definite kernels on finite words from feature maps in Hilbert modules over the Cuntz algebra O_n. We prove finite-sample prefix interpolation, truncation, and regularized learning bounds, and strong convergence of finite-rank word compressions. We also construct a machine-learning framework based on Cuntz-algebra-valued kernels.

On the full deduplicated SMS Spam Collection and AG News datasets, we compare learned states with matched diagonal controls, classical text baselines, and GRU and Transformer classifiers trained from scratch. The learned full state improves mean macro-F1 over its dephased control by 0.28 percentage points on SMS and scores 0.03 points lower on AG News.

Keywords: Cuntz algebra, Reproducing Kernel Hilbert Space, positive definite kernels.



Introduction

Many machine-learning problems involve data whose natural representation is not a vector of fixed length. For example, passages have different lengths. To work with such variable-length data, we want a framework that works well with finite words, including the operations on them.

The kernel method is a well-known technique in machine learning. Recent work generalizes the kernel method to the case where the kernels are C*-algebra-valued, instead of complex-valued. An example of C*-algebras is the Cuntz algebra O_n, which is generated by n isometries $S_1, \ldots, S_n$ satisfying $S_i^* S_j = \delta_{ij} 1$, $\sum S_i S_i^* = 1$. The Cuntz algebra can be used to represent finite words. Thus, we propose a machine learning framework based on the Cuntz algebra.

In this paper, we construct Cuntz-algebra-valued kernels, including the continuation kernel, and identify their scalarizations. We prove strict positivity and interpolation for the weighted prefix kernel on arbitrary finite sets of distinct words. We provide truncation bounds, establish strong convergence of finite-word compressions, and formulate models for multi-channel data. We also derive a supervised classifier from diagonal states on the same bigram features. Then we test classifiers based on Cuntz-algebra-valued kernels on the SMS Spam Collection and AG News datasets and compare the results with those from classical techniques.

Research Method

- We perform a literature study on classical machine learning techniques such as the kernel method and adapt them using Cuntz-algebra-valued kernels
- We provide an ablation study by performing kernel ridge regression (KRR) using the continuation kernel for 3 cases (learned diagonal state, learned full state, and dephased full state) and compare the results with baselines from classical methods. We use samples from the UCI SMS Spam Collection and the four-class AG News datasets.
- We perform KRR on the balanced 800-messages SMS pilot using various Cuntz-algebra-valued kernels and compare the results with baselines using classical kernels.

Discussion & Results

In the following discussion, let Σ_n^* be the monoid of finite words over the alphabet $\{1, \ldots, n\}$.

We construct kernels that map $\Sigma_n^* \times \Sigma_n^*$ to the Cuntz algebra O_n, including basic word kernels, weighted prefix kernels, occurrence-sensitive kernels, length-filtered kernels, and noncentral continuation kernels. We proved that all of these kernels are positive definite kernels.

We showed that O_n ($n \geq 2$) does not have a finite-dimensional *-representation. This implies that, for computational purposes, we need to consider finite truncations and compressions of those kernels. We showed that the truncated scalar prefix kernel converges to the actual scalar prefix kernel and that a compression of a Cuntz-algebra-valued kernel is also a positive definite kernel.

Table 1. Held-out classification results (%). SMS entries are the mean ± sample standard deviation over three overlapping splits. AG News uses a single fixed split.

Method	SMS Spam Collection		AG News	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Hellinger bigram	97.01 ± 0.63	93.29 ± 1.43	79.30	79.22
Learned diagonal state	96.98 ± 0.54	93.20 ± 1.27	79.30	79.22
Dephased full state	96.98 ± 0.64	93.20 ± 1.46	79.30	79.22
Learned full state	97.11 ± 0.59	93.48 ± 1.32	79.27	79.19
Character trigram spectrum	97.56 ± 0.59	94.25 ± 1.46	89.88	89.85
Word TF–IDF ridge	97.92 ± 0.37	95.13 ± 0.89	92.50	92.48
GRU	97.30 ± 0.59	93.77 ± 1.19	91.30	91.30
Transformer	97.40 ± 0.06	93.84 ± 0.36	91.51	91.49

Now we describe our experiments. From a density matrix D, we can construct a state φ_D: $O_n \to \mathbb{C}$. By applying φ_D to the continuation kernel we can obtain a scalar kernel that can be used in standard algorithms.

We tested KRR using the continuation kernel under 3 cases: φ_D is a learned full state, a learned diagonal state, and a dephased full state. We record the result in Table 1 alongside the baselines from other techniques. The result does not show a clear advantage in using the continuation kernel. Moreover, using a learned full state shows only a small, inconsistent empirical effect over using a dephased full state.

We also tested KRR using various kernels (Cuntz-algebra-valued and standard) and record the result in Table 2. The RBF character histogram—a classical method—performed best in this test.

Table 2. Text pilot: test percentages, mean ± sample standard deviation over three fixed splits. All models use the same kernel ridge regression protocol.

Kernel	Accuracy (%)	Macro-F1 (%)
Prefix (central lift)	80.4 ± 2.2	80.3 ± 2.3
Hellinger subword (central lift)	93.1 ± 1.1	93.1 ± 1.1
Classical spectrum	92.1 ± 1.4	92.1 ± 1.4
RBF character histogram	95.2 ± 2.0	95.2 ± 2.0
Prefix + Hellinger	94.2 ± 0.4	94.2 ± 0.4
Basic uniform state	52.5 ± 0.6	38.7 ± 1.3
Basic empirical state	52.5 ± 0.6	38.7 ± 1.3
Gauge delta	50.0 ± 0.0	33.3 ± 0.0

Conclusion

We developed a theoretical Cuntz-algebraic kernel framework for variable-length structured data. The main mathematical contribution is a family of O_n-valued positive definite kernels constructed from Hilbert C∗-module feature maps. Scalarization by positive functionals converts these kernels into ordinary positive definite kernels suitable for standard finite scalar kernel algorithms. The paper also shows that the implementations of Cuntz algebra in machine learning must be described as finite feature truncations, compressions of infinite-dimensional representations, or matrix-valued approximations rather than exact finite dimensional representations. We also discuss the limitations of these approaches. Finally, we performed experimental tests which show that using the learned full state when using the continuation kernel only shows a small, inconsistent advantage over using the dephased control and that using Cuntz algebra does not result in a considerable advantage over classical methods.

Output

I.Muchtadi-Alamsyah, N.Syafithri, F.Yuliawan, M. S. Alam, A Theoretical Cuntz-Algebraic Kernel Framework for Variable-Length Structured Data, AIMS Mathematics (accepted)