Autoencoding any Data through Kernel Autoencoders

Pierre Laforgue* **Stephan Clémençon*** **Florence d'Alché-Buc***
*LTCI, Télécom ParisTech, Université Paris Saclay, Paris, France

Abstract

This paper investigates a novel algorithmic approach to data representation based on kernel methods. Assuming that the observations lie in a Hilbert space $\mathcal{X}$, the introduced Kernel Autoencoder (KAE) is the composition of mappings from vector-valued Reproducing Kernel Hilbert Spaces (vv-RKHSs) that minimizes the expected reconstruction error. Beyond a first extension of the autoencoding scheme to possibly infinite dimensional Hilbert spaces, KAE further allows to autoencode any kind of data by choosing $\mathcal{X}$ to be itself a RKHS. A theoretical analysis of the model is carried out, providing a generalization bound, and shedding light on its connection with Kernel Principal Component Analysis. The proposed algorithms are then detailed at length: they crucially rely on the form taken by the minimizers, revealed by a dedicated Representer Theorem. Finally, numerical experiments on both simulated data and real labeled graphs (molecules) provide empirical evidence of the KAE performances.

1 INTRODUCTION

As experienced by any practitioner, data representation is critical to the application of Machine Learning, whatever the targeted task, supervised or unsupervised. An answer to this issue consists in feature engineering, a step that requires time-consuming interactions with domain experts. To overcome these limitations, Representation Learning (RL) (Bengio et al., 2013) aims at building automatically new features in an unsupervised fashion. Recent applications to neural nets pre-training, image denoising and semantic hashing have renewed a strong interest in RL, now a proper research field. Among successful RL approaches, mention has to be made of Autoencoders (AEs) (Vincent et al., 2010), and their generative variant, Deep Boltzman Machines (Salakhutdinov and Hinton, 2009).

AEs attempt to learn a pair of encoding/decoding functions under structural constraints so as to capture the most important properties of the data (Alain and Bengio, 2014). If they have mostly been studied under the angle of neural networks (Baldi, 2012) and deep architectures (Vincent et al., 2010), the concepts underlying AEs are very general and go beyond neural implementations. In this work, we develop a general framework inspired from AEs, and based on Operator-Valued Kernels (OVKs) (Senkene and Tempel'man, 1973) and vector-valued Reproducing Kernel Hilbert Spaces (vv-RKHSs). Mainly developed for supervised learning, OVKs provide a nonparametric way to tackle complex output prediction problems (Álvarez et al., 2012), including multi-task regression, structured output prediction (Brouard et al., 2016b), or functional regression (Kadri et al., 2016). This work is a first contribution to combine OVKs with AEs, enlarging the latters' applicability scope - so far restricted to $\mathbb{R}^d$ - to any data described by a similarity matrix.

We start from the simplest formulation in which a Kernel Autoencoder (KAE) is a pair of encoding/decoding functions lying in two different vv-RKHSs, and whose composition approximates the identity function. This approach is further extended to a general framework involving the composition of an arbitrary number of mappings, defined and valued on Hilbert spaces. A crucial application of KAEs arises if the input space is itself a RKHS: it allows to perform autoencoding on any type of data, by first mapping it to the RKHS, and then applying a KAE. The solutions computation, even in infinite dimensional spaces, is made possible by a Representer Theorem and the use of the kernel trick. This unlocks new applications on structured objects for which feature vectors are missing or too complex (*e.g.* in chemoinformatics).

Kernelizing an AE criterion has also been proposed by Gholami and Hajisami (2016). But their approach differs from ours in many key aspects: it is restricted

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

to AEs with 2 layers and composed of linear maps only; it relies on semi-supervised information; it comes with no theoretical analysis, and within a hashing perspective solely. Despite a similar title, the work by Kampffmeyer et al. (2017) has no connection with ours. It uses standard AEs, and regularize the learning by aligning the latent code with some predetermined kernel. In the experimental section, we implement autoencoding on graphs, which cannot be done by means of standard AEs. Graph AEs (Kipf and Welling, 2016) do not autoencode graphs, but $\mathbb{R}^d$ points with an additive graph characterizing the data structure.

The rest of the article is structured as follows. The novel kernel-based framework for RL is detailed in Section 2. A generalization bound and a strong connection with Kernel PCA are established in Section 3, whereas Section 4 describes the algorithmic approach based on a Representer Theorem. Illustrative numerical experiments are displayed in Section 5, while concluding remarks are collected in Section 6. Finally, technical details are deferred to the Appendix.

2 THE KERNEL AUTOENCODER

In this section, we introduce a general framework for building AEs based on vv-RKHSs. Here and throughout the paper, the set of bounded linear operators mapping a vector space E to itself is denoted by $\mathcal{L}(E)$, and the set of mappings from a set A to an ensemble B by $\mathcal{F}(A, B)$. The adjoint of an operator M is denoted by M^*. Finally, $[\![n]\!]$ denotes the set $\{1, \dots, n\}$ for any integer $n \in \mathbb{N}^*$.

2.1 Background on vv-RKHSs

Vv-RKHSs allow to cope with the approximation of functions from an input set $\mathcal{X}$ to some output Hilbert space $\mathcal{Y}$ (Senkene and Tempel'man, 1973; Caponnetto et al., 2008). Vv-RKHS can be defined from an OVK, which extends the classic notion of positive definite kernel. An OVK is a function $K\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, that satisfies the following two properties:

$$\forall (x, x') \in \mathcal{X} \times \mathcal{X}, \qquad K(x, x') = K(x', x)^*,$$

and $\forall n \in \mathbb{N}^*, \forall\, \{(x_i, y_i)\}_{1 \le i \le n} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\sum_{1 \le i,j \le n} \langle y_i, K(x_i, x_j) y_j \rangle_{\mathcal{Y}} \ge 0.$$

A simple example of OVK is the *separable kernel* such that: $\forall\ (x, x') \in \mathcal{X} \times \mathcal{X}$, $K(x, x') = k(x, x')A$, where k is a positive definite scalar-valued kernel, and A is a positive semi-definite operator on $\mathcal{Y}$. Its relevance for multi-task learning has been highlighted for instance by Micchelli and Pontil (2005).

Let K be an OVK, and for $x \in \mathcal{X}$, let $K_x\colon y \in \mathcal{Y} \mapsto K_x y \in \mathcal{F}(X, Y)$ the linear operator such that:

$$\forall x' \in \mathcal{X},\ (K_x y)(x') = K(x', x) y.$$

Then, there is a unique Hilbert space $\mathcal{H}_K \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ called the vv-RKHS associated to K, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and norm $\|\cdot\|_{\mathcal{H}_K}$, such that $\forall x \in \mathcal{X}$:

- K_x spans the space $\mathcal{H}_K$ ($\forall y \in \mathcal{Y}\colon K_x y \in \mathcal{H}_K$)
- K_x is bounded for the uniform norm
- $\forall f \in \mathcal{H},\ f(x) = K_x^* f$ (*i.e.* reproducing property)

2.2 Input Output Kernel Regression

Now, let us assume that the output space $\mathcal{Y}$ is chosen itself as a RKHS, say $\mathcal{H}$, associated to the positive definite scalar-valued kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, with $\mathcal{Z}$ a non-empty set. Working in the vv-RKHS $\mathcal{H}_K$ associated to an OVK $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H})$ opens the door to a large family of learning tasks where the output set $\mathcal{Z}$ can be a set of complex objects such as nodes in a graph, graphs (Brouard et al., 2016a) or functions (Kadri et al., 2016). Following the work of Brouard et al. (2016b), we refer to these methods as Input Output Kernel Regression (IOKR). IOKR has been shown to be of special interest in case of Ridge Regression, where closed-form solutions are available besides classical gradient descent algorithms. Note that in a general supervised setting, learning a function $f \in \mathcal{H}_K$ is not sufficient to provide a prediction in the output set, and a pre-image problem has to be solved. In sections 2.5 and 4.3, a similar idea is applied at the last layer of our KAE, allowing for auto-encoding non-vectorial data while avoiding complex pre-image problems.

2.3 The 2-layer Kernel Autoencoder (KAE)

Let $S = (x_1, \dots, x_n)$ denote a sample of n independent realizations of a random vector X, valued in a separable Hilbert space $(\mathcal{X}_0, \|\cdot\|_{\mathcal{X}_0})$ with unknown probability distribution P, and such that there exists $M < +\infty, \|X\|_{\mathcal{X}_0} \le M$ almost surely. On the basis of the training sample S, we are interested in constructing a pair of encoding/decoding mappings $(f : \mathcal{X}_0 \to \mathcal{X}_1,\ g : \mathcal{X}_1 \to \mathcal{X}_0)$, where $(\mathcal{X}_1, \|\cdot\|_{\mathcal{X}_1})$ is the (Hilbert) *representation space*. Just as for standard AEs, we regard as good internal representations the ones that allow for an accurate recovery of the original information in expectation. The problem to be solved states as follows:

$$\min_{\substack{(f,g) \in \mathcal{H}_1 \times \mathcal{H}_2 \\ \|f\|_{\mathcal{H}_1} \le s,\ \|g\|_{\mathcal{H}_2} \le t}} \epsilon(f, g) \coloneqq \mathbb{E}_{X \sim P} \|X - g \circ f(X)\|_{\mathcal{X}_0}^2, \tag{1}$$

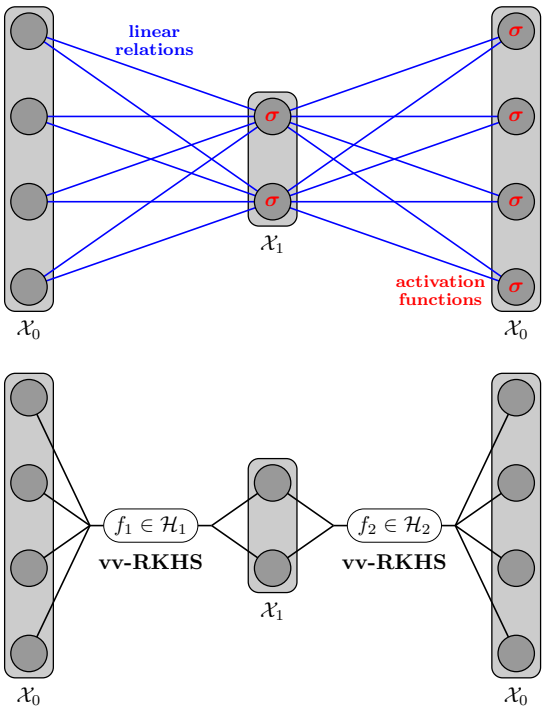


Figure 1: Standard and Kernel 2-layer Autoencoders

where $\mathcal{H}_1$ and $\mathcal{H}_2$ are two vv-RKHSs, and s and t two positive constants. $\mathcal{H}_1$ is associated to an OVK $\mathcal{K}_1 : \mathcal{X}_0 \times \mathcal{X}_0 \to \mathcal{L}(\mathcal{X}_1)$, while $\mathcal{H}_2$ is associated to $\mathcal{K}_2 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathcal{L}(\mathcal{X}_0)$. Figure 1 illustrates the parallel and differences between standard and kernel 2-layer Autoencoders.

Following the Empirical Risk Minimization (ERM) paradigm, the true risk (1) is replaced by its empirical version

$$\hat{\epsilon}_n(f, g) \coloneqq \frac{1}{n} \sum_{i=1}^{n} ||x_i - g \circ f(x_i)||^2_{\mathcal{X}_0},$$

and a penalty term $\Omega(f, g) \coloneqq \lambda \|f\|^2_{\mathcal{H}_1} + \mu \|g\|^2_{\mathcal{H}_2}$ is added instead of the norm constraints (see Theorem 1). Solutions to the following regularized ERM problem shall be referred to as *2-layer KAE*:

$$\min_{(f,g) \in \mathcal{H}_1 \times \mathcal{H}_2} \hat{\epsilon}_n(f, g) + \Omega(f, g). \tag{2}$$

2.4 The Multi-layer KAE

Like for standard AEs, the model previously described can be directly extended to more than 2 layers. Let $L \geq 3$, and consider a collection of Hilbert spaces $\mathcal{X}_0, \ldots, \mathcal{X}_L$, with $\mathcal{X}_L = \mathcal{X}_0$. For $0 \leq l \leq L-1$, the space $\mathcal{X}_l$ is supposed to be endowed with an OVK $\mathcal{K}_{l+1} : \mathcal{X}_l \times \mathcal{X}_l \to \mathcal{L}(\mathcal{X}_{l+1})$, associated to a vv-RKHS $\mathcal{H}_{l+1} \subset \mathcal{F}(\mathcal{X}_l, \mathcal{X}_{l+1})$. We then want to minimize $\epsilon(f_1, \ldots, f_L)$ over $\prod_{l=1}^{L} \mathcal{H}_l$. Setting $\Omega(f_1, \ldots, f_L) \coloneqq \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_L}$ allows for a direct extension of (2):

$$\min_{f_l \in \mathcal{H}_l} \frac{1}{n} \sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}. \tag{3}$$

2.5 The General Hilbert KAE and the K^{2}AE

So far, and up to the regularization term, the main difference between standard and kernel AEs is the function space on which the reconstruction criterion is optimized: respectively neural functions or RKHS ones. But what should also be highlighted is that RKHS functions are valued in general Hilbert spaces, while neural functions are restricted to $\mathbb{R}^d$. As shall be seen in section 4.3, this enables KAEs to handle data from infinite dimensional Hilbert spaces (*e.g.* function spaces), what standard AEs are unable to do. To our knowledge, this first extension of the autoencoding scheme is novel.

But even more interesting is the possible extension when the input/output Hilbert space is chosen to be itself a RKHS. Indeed, let $\mathcal{X}_0$ denote now any set (without the Hilbert assumption). In the spirit of IOKR, let us first map $x \in \mathcal{X}_0$ to the RKHS $\mathcal{H}$ associated to some scalar kernel k, and its canonical feature map ϕ. Since the $\phi(x_i)'s$ are by definition valued in a Hilbert, KAE can be applied. This way, we have extended the autoencoding paradigm to any set, and finite dimensional representations can be extracted from all types of data. Again, such extension is novel to our knowledge. Figure 2 depicts the procedure, referred to as K^{2}AE, since the new criterion is a *kernelization* of the KAE that reads:

$$\frac{1}{n} \sum_{i=1}^{n} \|\phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i))\|^2_{\mathcal{H}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}. \tag{4}$$

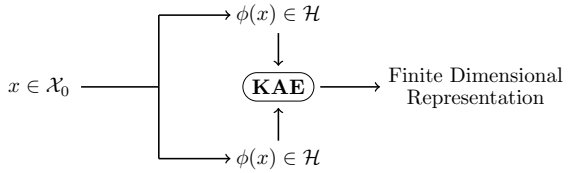


Figure 2: Autoencoding any data thanks to K^{2}AE

3 THEORETICAL ANALYSIS

It is the purpose of this section to investigate theoretical properties of the introduced model, its capacity to be learnt from training data with a controlled generalization error, and the connection between K^{2}AE and Kernel PCA (KPCA) namely.

3.1 Generalization Bound

While the algorithmic formulation aims at minimizing the regularized risk (2), the subsequent theoretical analysis focuses on the constrained problem (1). Theorem 1 relates the solutions from the two approaches to each other, so that bounds derived in the latter setting also apply to numerical solutions of the first one.

Theorem 1. *Let $V : \mathcal{H}_1 \times \ldots \times \mathcal{H}_L \to \mathbb{R}$ be an arbitrary function. Consider the two problems:*

$$\min_{f_l \in \mathcal{H}_l} \left\{ V(f_1, \ldots, f_L) + \sum_{l=1}^{L} \lambda_l \|f_l\|_{\mathcal{H}_l}^2 \right\}, \tag{5}$$

$$\min_{\substack{f_l \in \mathcal{H}_l \\ \|f_l\|_{\mathcal{H}_l} \leq s_l}} V(f_1, \ldots, f_L). \tag{6}$$

Then, for any $(\lambda_1, \ldots, \lambda_L) \in \mathbb{R}_+^L$, there exists $(s_1, \ldots, s_L) \in \mathbb{R}_+^L$ such that any (respectively, local) solution to problem (5) is also a (respectively, local) solution to problem (6).

Refer to Appendix A.1 for the proof and a discussion on the converse statement.

In order to establish generalization bound results for empirical minimizers in the present setting, we now define two key quantities involved in the proof, *i.e.* Rademacher and Gaussian averages for classes of Hilbert-valued functions.

Definition 2. *Let $\mathcal{X}$ be any measurable space, and H a separable Hilbert space. Consider a class $\mathcal{C}$ of measurable functions $h : \mathcal{X} \to H$. Let $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n$ be $n \geq 1$ independent H-valued Rademacher variables and define:*

$$\widehat{\mathscr{R}}_n(\mathcal{C}(S)) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\sigma}_i, h(x_i) \rangle_H \right].$$

If $H = \mathbb{R}$, it is the classical Rademacher average (see *e.g.* Mohri et al. (2012) p.34), while, when $H = \mathbb{R}^p$, it corresponds to the expectation of the supremum of the sum of the Rademacher averages over the p components of h (see Definition 2.1 in Maurer and Pontil (2016)). If H is an infinite dimensional Hilbert space with countable orthonormal basis $(e_k)_{k \in \mathbb{N}}$, we have:

$$\widehat{\mathscr{R}}_n(\mathcal{C}(S)) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{\infty} \sigma_{i,k} \langle h(x_i), e_k \rangle_H \right].$$

The Gaussian counterpart of $\widehat{\mathscr{R}}_n(\mathcal{C}(S))$, obtained by replacing Rademacher random variables/processes with standard H-valued Gaussian ones, is denoted by $\widehat{\mathscr{G}}_n(\mathcal{C}(S))$ throughout the paper.

For the sake of simplicity, results in the rest of the subsection are derived in the 2-layer case solely, with $\mathcal{X}_1$ finite dimensional (*i.e.* $\mathcal{X}_1 = \mathbb{R}^p$), although the approach remains valid for deeper architectures.

Let $\mathcal{H}_{1,s} := \{f \in \mathcal{H}_1 : \|f\|_{\mathcal{H}_1} \leq s\}$, and similarly $\mathcal{H}_{2,t} := \{g \in \mathcal{H}_2 : \|g\|_{\mathcal{H}_2} \leq t, \sup_{y \in \mathbb{R}^p} \|g(y)\|_{\mathcal{X}_0} \leq M\}$. We shall use the notation $\mathcal{H}_{s,t} \subset \mathcal{F}(\mathcal{X}_0, \mathcal{X}_0)$ to mean the space of composed functions $\mathcal{H}_{1,s} \circ \mathcal{H}_{2,t} := \{h \in \mathcal{F}(\mathcal{X}_0, \mathcal{X}_0) : \exists (f, g) \in \mathcal{H}_{1,s} \times \mathcal{H}_{2,t},\ h = g \circ f\}$. To simplify the notation, ϵ (and $\hat{\epsilon}_n$) may be abusively considered as a functional with one or two arguments: $\epsilon(f, g) = \epsilon(g \circ f) = \mathbb{E}_{X \sim P} \|X - g \circ f(X)\|_{\mathcal{X}_0}^2$. Finally, let $\hat{h}_n$ denote the minimizer of $\hat{\epsilon}_n$ over $\mathcal{H}_{s,t}$, and ϵ^* the infimum of ϵ on the same functional space.

Assumption 3. *There exists $K < +\infty$ such that:*

$$\forall x \in \mathcal{X}_0, \quad \mathbf{Tr}\big(\mathcal{K}_1(x, x)\big) \leq Kp.$$

Assumption 4. *There exists $L < +\infty$ such that for all y, y' in $\mathbb{R}^p$:*

$$\mathbf{Tr}\big(\mathcal{K}_2(y, y) - 2\mathcal{K}_2(y, y') + \mathcal{K}_2(y', y')\big) \leq L^2 \|y - y'\|_{\mathbb{R}^p}^2.$$

Theorem 5. *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be OVKs satisfying Assumptions 3 and 4 respectively. Then, there exists a universal constant $C_0 < +\infty$ such that, for any $0 < \delta < 1$, we have with probability at least $1 - \delta$:*

$$\epsilon(\hat{h}_n) - \epsilon^* \leq C_0 LMst \sqrt{\frac{Kp}{n}} + 24M^2 \sqrt{\frac{\log(2)/\delta}{2n}}.$$

The proof relies on a Rademacher bound, which is in turn upper bounded using Corollary 4 in Maurer (2016), an extension of Theorem 2 in Maurer (2014) proved in the Supplementary Material, and several intermediary results derived from the stipulated assumptions. Technical details are deferred to Appendix A.2.

Attention should be paid to the fact that constants in Theorem 5 appear in a very interpretable fashion: the less spread the input (the smaller the constant M), the more restrictive the constraints on the functions (the smaller K, L, s and t), and the smaller the internal dimension p, the sharper the bound.

3.2 K^{2}AE and Kernel PCA: a Connection

Just as Bourlard and Kamp (1988) have shown a mere equivalence between PCA and standard 2-layer AEs, a similar link can be established between 2-layer K^{2}AE and Kernel PCA. Throughout the analysis, a 2-layer K^{2}AE is considered, with decomposable kernels made

of linear scalar kernels and identity operators. Also, there is no penalization (*i.e.* $\lambda_1 = \lambda_2 = 0$). We want to autoencode data into $\mathbb{R}^p$, after a first embedding through the feature map ϕ, like in (4).

3.2.1 Finite Dimensional Feature Map

Let us assume first that ϕ is valued in $\mathbb{R}^d$, with $p < d < n$. Let $\Phi = (\phi(x_1), \dots, \phi(x_n))^T \in \mathbb{R}^{n \times d}$ denote the matrix storing the $\phi(x_i)^T$ to autoencode in rows. Note that $K_\phi = \Phi\Phi^T \in \mathbb{R}^{n \times n}$ corresponds to the Gram matrix associated to ϕ. As shall be seen in Theorem 6, the optimal f and g have a specific form, so that they only depend on two coefficient matrices, $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{n \times d}$ respectively. Equipped with this notation, one has: $Y = f_A(\Phi) = \Phi\Phi^T A \in \mathbb{R}^{n \times p}$, and $\widetilde{\Phi} = g_B(Y) = YY^T B \in \mathbb{R}^{n \times d}$. Without penalization, the goal is then to minimize in A and B:

$$\|\Phi - \widetilde{\Phi}\|_{Fr}^2.$$

$\widetilde{\Phi}$ being at most of rank p, we know from Eckart-Young Theorem that the best possible $\widetilde{\Phi}$ is given by $\Phi^* = U\Sigma_p V^T$, where $\left(U \in \mathbb{R}^{n \times d}, \Sigma \in \mathbb{R}^{d \times d}, V^T \in \mathbb{R}^{d \times d}\right)$ is the *thin* Singular Value Decomposition (SVD) of Φ such that $\Phi = U\Sigma V^T$, and Σ_p is equal to Σ, but with the $d - p$ smallest singular values zeroed.

Let us now prove that there exists a couple of matrices (A^*, B^*) such that $g_{B^*} \circ f_{A^*}(\Phi) = \Phi^*$. One can verify that $(A^* = U_p \overline{\Sigma}_p^{-3/2}, B^* = UV^T)$, with $U_p \in \mathbb{R}^{n \times p}$ storing only the p largest eigenvectors of K_ϕ, and $\overline{\Sigma}_p \in \mathbb{R}^{p \times p}$ the $p \times p$ top left block of Σ_p, satisfy it. Finally, the optimal encoding returned is $Y^* = f_{A^*}(\Phi) = \left(\sqrt{\sigma_1}u_1, \dots, \sqrt{\sigma_p}u_p\right)$, with $u_1, \dots, u_p$ the p largest eigenvectors of K_ϕ, while the KPCA's new representation is $(\sigma_1 u_1, \dots, \sigma_p u_p)$.

We have shown that a specific instance of K^{2}AE can be solved explicitly using a SVD, and that the optimal coding returned is close to the one output by KPCA.

3.2.2 Infinite Dimensional Feature Map

Let us assume now that ϕ is valued in a general Hilbert space $\mathcal{H}$. Φ is now seen as the linear operator from $\mathcal{H}$ to $\mathbb{R}^n$ such that $\forall \alpha \in \mathcal{H}$, $\Phi\alpha = (\langle \alpha, \phi(x_1) \rangle_{\mathcal{H}}, \dots, \langle \alpha, \phi(x_n) \rangle_{\mathcal{H}}) \in \mathbb{R}^n$. Since Theorem 1 makes no assumption on the dimensionality, everything stated in the finite dimensional scenario applies, except that $B \in \mathcal{L}(\mathcal{H}, \mathbb{R}^n)$, and that we minimize the Hilbert-Schmidt norm: $\|\Phi - \widetilde{\Phi}\|_{HS}^2$. We then need an equivalent of Eckart-Young Theorem. It still holds since its proof only requires the existence of an SVD for any operator, which is granted in our case since we deal with compact operators (they have finite rank n). The end of the proof is analogous to the finite dimensional case.

4 THE KAE ALGORITHMS

This section describes at length the algorithms we propose to solve problems (3) and (4). They raise two major issues as their objective functions are non-convex, and their search spaces are infinite dimensional. However, this last difficulty is solved by Theorem 6.

4.1 A Representer Theorem

Theorem 6. *Let $L_0 \in [\![L]\!]$, and $V : \mathcal{X}_{L_0}^n \times \mathbb{R}_+^{L_0} \to \mathbb{R}$ a function of $n + L_0$ variables, strictly increasing in each of its L_0 last arguments. Suppose that $(f_1^*, \dots, f_{L_0}^*)$ is a solution to the optimization problem:*

$$\min_{f_l \in \mathcal{H}_l} V\Big((f_{L_0} \circ \dots \circ f_1)(x_1), \dots, (f_{L_0} \circ \dots \circ f_1)(x_n), \|f_1\|_{\mathcal{H}_1}, \dots, \|f_{L_0}\|_{\mathcal{H}_{L_0}}\Big).$$

Let $x_i^{*(l)} := f_l^* \circ \dots \circ f_1^*(x_i)$, with $x_i^{*(0)} := x_i$. Then, $\exists \left(\varphi_{1,1}^*, \dots, \varphi_{1,n}^*, \dots, \varphi_{L_0,n}^*\right) \in \mathcal{X}_1^n \times \dots \times \mathcal{X}_{L_0}^n$:

$$\forall\, l \in [\![L_0]\!], \quad f_l^*(\cdot) = \sum_{i=1}^n \mathcal{K}_l\left(\ \cdot\ , x_i^{*(l-1)}\right)\varphi_{l,i}^*.$$

Proof. Refer to Appendix A.3 □

This Theorem exhibits a very specific structure for the minimizers, as each layer's support vectors are the images of the original points by the previous layer.

4.2 Finite Dimension Case

In this section, let us assume that $\mathcal{X}_l = \mathbb{R}^{d_l}$ for $l \in [\![L]\!]$. The objective function of (3), viewed as a function of $(f_L \circ \dots \circ f_1)(x_1), \dots, (f_L \circ \dots \circ f_1)(x_n), \|f_1\|_{\mathcal{H}_1}, \dots, \|f_L\|_{\mathcal{H}_L}$ satisfies the condition on V involved in Theorem 6. After applying it (with $L_0 = L$), problem (3) boils down to the problem of finding the $\varphi_{l,i}^*$'s, which are finite dimensional. This crucial observation shows that our problem can be solved in a computable manner. However, its convexity still cannot be ensured (see Appendix A.4).

The objective only depending on the $\varphi_{l,i}$'s, problem (3) can be approximately solved by Gradient Descent (GD). We now specify the gradient derivation in the decomposable OVKs case, *i.e.* for any layer l there exists a scalar kernel k_l and $A_l \in \mathcal{L}(X_l)$ positive semidefinite such that $\mathcal{K}_l(x, x') = k_l(x, x')A_l$. All detailed computations can be found in Appendix B. Let $\Phi_l := (\varphi_{l,1}, \dots, \varphi_{l,n})^T \in \mathbb{R}^{n \times d_l}$ storing the coefficients $\varphi_{l,i}$ in rows, and $K_l \in \mathbb{R}^{n \times n}$ such that $[K_l]_{i,i'} = k_l\left(x_i^{(l-1)}, x_{i'}^{(l-1)}\right)$. Let $(l_0, i_0) \in [\![L]\!] \times [\![n]\!]$, the gradient of the distortion term reads:

$$\left(\nabla_{\varphi_{l_0,i_0}} \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0}\right)^T \tag{7}$$
$$= -\frac{2}{n}\sum_{i=1}^{n}\left(x_i - {x_i}^{(L)}\right)^T \mathbf{Jac}_{{x_i}^{(L)}}(\varphi_{l_0,i_0}).$$

On the other hand, $\|f_l\|^2_{\mathcal{H}_l}$ may be rewritten as:

$$\|f_l\|^2_{\mathcal{H}_l} = \sum_{i,i'=1}^{n} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\langle \varphi_{l,i}, A_l\varphi_{l,i'}\rangle_{\mathcal{X}_l}, \tag{8}$$

so that it may depend on φ_{l_0,i_0} in two ways: 1) if $l_0 = l$, there is a direct dependence of the second quadratic term, 2) but note also that for $l_0 < l$, the $\varphi_{l_0,i}$ have an influence on the ${x_i}^{(l-1)}$ and so on the first term. This remark leads to the following formulas:

$$\nabla_{\Phi_l}\ \|f_l\|^2_{\mathcal{H}_l} = 2\ K_l\Phi_l A_l, \tag{9}$$

with $\nabla_{\Phi_l}F \coloneqq \left((\nabla_{\varphi_{l,1}}F)^T, \ldots, (\nabla_{\varphi_{l,n}}F)^T\right)^T \in \mathbb{R}^{n\times d_l}$ storing the gradients of any real-valued function F with respect to the $\varphi_{l,i}$ in rows. And when $l_0 < l$:

$$\left(\nabla_{\varphi_{l_0,i_0}}\|f_l\|^2_{\mathcal{H}_l}\right)^T = 2\sum_{i,i'=1}^{n}\Big\{ \tag{10}$$
$$[N_l]_{i,i'}\left(\nabla^{(1)}k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0})\Big\},$$

where $\nabla^{(1)}k_l(x,x')$ denotes the gradient of $k_l(\cdot,\cdot)$ with respect to the 1^{st} coordinate evaluated in (x,x'), and N_l the $n\times n$ matrix such that $[N_l]_{i,i'} = \langle\varphi_{l,i}\ ,\ A_l\ \varphi_{l,i'}\rangle_{\mathcal{X}_l}$. Again, assuming the matrices $\mathbf{Jac}_{{x_i}^{(L)}}(\varphi_{l_0,i_0})$ are known, the norm part of the gradient is computable. Combining expressions (7), (9) and (10) using the linearity of the gradient leads readily to the complete formula.

If n, L, and p denote respectively the number of samples, the number of layers, and the size of the largest latent space, the algorithm complexity is no more than $\mathcal{O}(n^2Lp)$ for objective evaluation, and $\mathcal{O}(n^3L^2p^3)$ for gradient derivation. Hence, it appears natural to consider stochastic versions of GD. But as shown by equation (10), the norms gradients involve the computation of many Jacobians. Selecting a mini-batch does not affect these terms, which are the most time consuming. Thus, the expected acceleration due to stochasticity must not be so important. Nevertheless, a *doubly stochastic* scheme where both the points on which the objective is evaluated, as well as the coefficients to be updated, are chosen randomly at each iteration, might be of high interest since it would dramatically decrease the number of Jacobians computed. However, this approach goes beyond the scope of this paper, and is left for future work.

4.3 General Hilbert Space Case

In this section, $\mathcal{X}_0$ (and so $\mathcal{X}_L$) are supposed to be infinite dimensional. Despite this relaxation, KAEs remains computable. As Theorem 6 makes no assumption on the dimensionality of $\mathcal{X}_0$, it can be applied. The only difference is that coefficients $\varphi_{L,i}$'s $\in \mathcal{X}_L^n$ are infinite dimensional, preventing from the use of a global GD. But assuming the $\varphi_{L,i}$'s to be fixed, a GD can still be performed on the $\varphi_{l,i}$'s, $l \in [\![L-1]\!]$. On the other hand, if one assumes these coefficients fixed, the optimal $\varphi_{L,i}$'s are the solutions to a Kernel Ridge Regression (KRR). Consequently, a hybrid approach alternating GD and KRR is considered. Two issues remain to be addressed: 1) how to compute the KRR in $\mathcal{X}_L$, 2) how to propagate the gradients through $\mathcal{X}_L$.

From now, A_L is assumed to be the identity operator. If the $\varphi_{l,i}$'s, $l \in [\![L-1]\!]$ are fixed, then the best $\varphi_{L,i}$'s shall satisfy (Micchelli and Pontil, 2005) for all $i \in [\![n]\!]$:

$$\sum_{i'=1}^{n}\left(\mathcal{K}_L\left({x_i}^{(L-1)}, {x_{i'}}^{(L-1)}\right) + n\lambda_L\delta_{ii'}\right)\varphi_{L,i'} = x_i. \tag{11}$$

In particular, the computation of N_L becomes explicit (Appendix B.5) as long as we know the dot products $\langle x_j, x_{j'}\rangle_{\mathcal{X}_0}$. In the case of the K^{2}AE, these dot products are the input Gram matrix K_{in}. Let N_{KRR} be the function that computes N_L from the $\varphi_{l,i}$'s, $l \in [\![L-1]\!]$, K_{in} and λ_L. What is remarkable is that knowing N_L (and not each $\varphi_{L,i}$ individually) is enough to propagate the gradient through the infinite dimensional layer.

Indeed, let us assume now that N_L is fixed. All spaces but $\mathcal{X}_L$ remaining finite dimensional, changes in the gradients only occur where the last layer is involved, namely for the distortion and for $\|f_L\|^2_{\mathcal{H}_L}$. As for the gradients of $\|f_L\|^2_{\mathcal{H}_L}$, equation (10) remain true. If N_L is given, there is no difficulty. As for the distortion, the use of the differential (see Appendix B.6) gives:

$$\nabla_{\varphi_{l_0,i_0}}\left\|x_i - x_i^{(L)}\right\|^2_{\mathcal{X}_0}{}^T = -2\sum_{i'=1}^{n}\Big\{ \tag{12}$$
$$\left\langle x_i - x_i^{(L)}, \varphi_{L,i'}\right\rangle_{\mathcal{X}_L}\left(\nabla_{\varphi_{l_0,i_0}}k_L\left(x_i^{(L-1)}, x_{i'}^{(L-1)}\right)\right)^T\Big\}.$$

It is a direct extension of (7), where $\mathbf{Jac}_{{x_i}^{(L)}}(\varphi_{l_0,i_0})$, has been replaced using the definition of ${x_i}^{(L)}$. Using again (11), $\langle x_i - x_i^{(L)}, \varphi_{L,i'}\rangle_{\mathcal{X}_L}$ can be rewritten as $n\lambda_L\langle\varphi_{L,i}, \varphi_{L,i'}\rangle_{\mathcal{X}_L} = n\lambda_L[N_L]_{i,i'}$, and infinite dimensional objects are dealt with. The crux of the algorithm is that infinite dimensional coefficients $\varphi_{L,i}$'s are never computed, but only their scalar products. Not knowing the $\varphi_{L,i}$'s is of no importance, as we are interested in the encoding function, which does not rely on them. Let T be a number of epochs, and γ_t a step size rule, the approach is summarized in Algorithm 1.

Algorithm 1 General Hilbert KAE and K^{2}AE

input : Gram matrix K_{in}
init : $\Phi_1 = \Phi_1^{init}, \dots, \Phi_{L-1} = \Phi_{L-1}^{init}$,
$N_L = N_{\text{KRR}}(\Phi_1, \dots, \Phi_{L-1}, K_{in}, \lambda_L)$
for *epoch* t *from* 1 *to* T **do**
  // inner coefficients updates at fixed N_L
  for *layer* l *from* 1 *to* $L-1$ **do**
    $\Phi_l = \Phi_l - \gamma_t \nabla_{\Phi_l}(\hat{\epsilon}_n + \Omega \mid N_L)$
  // N_L update
  $N_L = N_{\text{KRR}}(\Phi_1, \dots, \Phi_{L-1}, K_{in}, \lambda_L)$
return $\Phi_1, \dots, \Phi_{L-1}$

5 NUMERICAL EXPERIMENTS

Numerical experiments have been run in order to assess the ability of KAEs to provide relevant data representations. We used decomposable OVKs with the identity operator as A, and the Gaussian kernel as k. First, we present insights on the interesting properties of the KAEs via a 2D example. Then, we describe more involved experiments on the NCI dataset to measure the power of KAEs.

5.1 Behavior on a 2D problem

Let us first consider three noisy concentric circles such as in Figure 3(a). Although the main strength of KAEs is to perform autoencoding on complex data (Section 2.5), they can still be applied on real-valued points. Figures 3(b) and 3(c) show the reconstructions obtained after fitting respectively a 2-1-2 standard and kernel AE. Since the latent space is of dimension 1, the 2D reconstructions are manifolds of the same dimension, hence the curve aspect. What is interesting though is that the KAE learns a much more complex manifold than the standard AE. Due to its linear limitations (the nonlinear activation functions did not help much in this case), the standard AE returns a line, far from the original data, while the KAE outputs a more complex manifold, much closer to the initial data.

Apart from a good reconstruction, we are interested in finding representations with attractive properties. The 1D feature found by the previous KAE is interesting, as it is a discriminative one with respect to the original clusters: points from different circles are mapped around different values (Figure 3(d)). Interestingly, after a few iterations, some variability is introduced around these *cluster values*, so that all codes shall not be mapped back to the same point (Figure 3(e)).

Finally, a KAE with 1 hidden layer of size 2 gives the internal representation shown in Figure 3(f). This new 2D representation has a disentangling effect: the circle structure is kept so as to preserve the intra-cluster specificity, while the inter-cluster differentiation is ensured by the circles' dissociation. These visual 2D examples give interesting insights on the good properties of the KAE representations: discrimination, disentanglement (see further experiments in Appendix C.1).

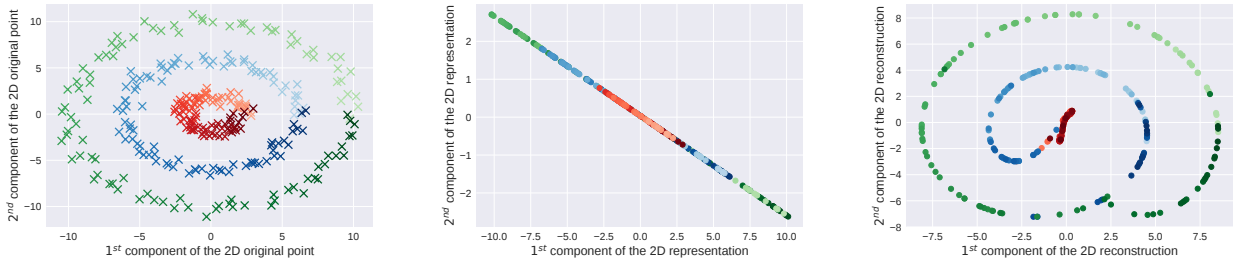
5.2 Representation Learning on Molecules

We now present an application of KAEs in the context of chemoinformatics. The motivation is triple. First, such complex data cannot be handled by standard AEs. Second, kernel methods being prominent in the field, data are often stored as Gram matrices, suiting perfectly our framework. Third, finding a compressed representation of a molecule is a problem of highest interest in Drug Discovery. We considered two different problems, one supervised, one unsupervised.

As for the supervised one, we exploited the dataset of Su et al. (2010) from the NCI-Cancer database: it consists in a Gram matrix comparing 2303 molecules by the mean of a Tanimoto kernel (a linear path kernel built using the presence or absence of sequences of atoms in the molecule), as well as the molecules activities in the presence of 59 types of cancer. The dataset containing no vectorial representations of the molecules (but only Gram matrices), only kernel methods were possible to benchmark. As a good representation is supposed to facilitate ulterior learning tasks, we assess the goodness of the representations through the regression scores obtained by Random Forests (RFs) from scikit-learn (Pedregosa et al., 2011) fed with it.

2-layer K^{2}AEs with respectively 5, 10, 25, 50 and 100 internal dimension were run, as well as Kernel Principal Component Analyses (KPCAs) with the same number of components. Finally, these representations were given as inputs to RFs. KRR was also added to the comparison. The Normalized Mean Squared Errors (NMSEs), averaged on 10 runs, for 5 strategies and on the first 10 cancers are stored in Table 1 (the complete results are available in Appendix C.2). A visualization with all strategies is also proposed in Figure 8. Clearly, methods combining a data representation step followed by a prediction one performs better. But the good performance of our approach should not be attributed to the use of RFs only, since the same strategy run with KPCA leads to worse results. Indeed, the K^{2}AE 50 + RF strategy outperforms all other procedures on all problems, managing to extract compact and useful feature vectors from the molecules.

The data for the unsupervised problem is taken from Brouard et al. (2016a). It is composed of two sets (a train set of size 5579, and a test set of size 1395), each one containing metabolites under the form of 4136-long binary vectors (called fingerprints), as well as a Gram matrix comparing them. 2-layer standard AEs



(a) 2D Input Data (b) AE Reconstruction (c) KAE Reconstruction

(d) 1D Representation (5 it.) (e) 1D Representation (20 it.) (f) 2D Re-Representation

Figure 3: KAE Performance on Noisy Concentric Circles

Table 1: NMSEs on Molecular Activity for Different Types of Cancer

	KRR	KPCA 10 + RF	KPCA 50 + RF	K^{2}AE 10 + RF	K^{2}AE 50 + RF
Cancer 01	0.02978	0.03279	0.03035	0.03097	**0.02808**
Cancer 02	0.03004	0.03194	0.02978	0.03099	**0.02775**
Cancer 03	0.02878	0.03155	0.02914	0.02989	**0.02709**
Cancer 04	0.03003	0.03274	0.03074	0.03218	**0.02924**
Cancer 05	0.02954	0.03185	0.02903	0.03065	**0.02754**
Cancer 06	0.02914	0.03258	0.03083	0.03134	**0.02838**
Cancer 07	0.03113	0.03468	0.03207	0.03257	**0.03018**
Cancer 08	0.02899	0.03162	0.02898	0.03065	**0.02770**
Cancer 09	0.02860	0.02992	0.02804	0.02872	**0.02627**
Cancer 10	0.02987	0.03291	0.03111	0.03170	**0.02910**

Table 2: MSREs on Test Metabolites

Dimension	AE (sigmoid)	AE (relu)	KAE
5	99.81	96.62	**76.38**
10	87.36	84.02	**65.76**
25	72.31	68.77	**51.63**
50	63.00	58.29	**40.72**
100	55.43	48.63	**36.27**

from Keras (Chollet et al., 2015) with sigmoid and relu activation functions, and 2-layer KAEs with internal layer of size 5, 10, 25, 50 and 100, were trained. In absence of a supervised task, we measured the Mean Squared Reconstruction Errors (MSREs) induced on the test set, and stored them in Table 2. Again, the KAE approach shows a systematic improvement.

6 CONCLUSION

We introduce a new framework for AEs, based on vv-RKHSs and OVKs. The use of RKHS functions enables KAEs to handle data from possibly infinite dimensional Hilbert spaces, and then to extend the autoencoding scheme to any kind of data. A generalization bound and a strong connection to KPCA are established, while the underlying optimization problem is tackled by a Representer Theorem and the kernel trick. Beyond a detailed description, the behavior of the algorithm is carefully studied on simulated data, and yields relevant performances on graph data, that standard AEs are typically unable to handle. Further research may consider a semi-supervised approach, that would ideally tailor the representation according to the future targeted task.

Acknowledgment This work has been funded by the Industrial Chair *Machine Learning for Big Data* from Télécom ParisTech, Paris, France.

References

Alain, G. and Bengio, Y. (2014). What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.*, 15(1):3563–3593.

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 37–49.

Bauschke, H. H., Combettes, P. L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

Bengio, Y., Courville, A., Vincent, P., and Umanità, V. (2013). Representation learning: a review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294.

Brouard, C., Shen, H., Dührkop, K., d'Alché-Buc, F., Böcker, S., and Rousu, J. (2016a). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):28–36.

Brouard, C., Szafranski, M., and d'Alché-Buc, F. (2016b). Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, 17:176:1–176:48.

Caponnetto, A., Micchelli, C. A., , M., and Ying, Y. (2008). Universal multitask kernels. *Journal of Machine Learning Research*, 9:1615–1646.

Chollet, F. et al. (2015). Keras, https://keras.io.

Gholami, B. and Hajisami, A. (2016). Kernel autoencoder for semi-supervised hashing. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. (2016). Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17:20:1–20:54.

Kampffmeyer, M., Løkse, S., Bianchi, F. M., Jenssen, R., and Livi, L. (2017). Deep kernelized autoencoders. In *Scandinavian Conference on Image Analysis*, pages 419–430. Springer.

Kipf, T. N. and Welling, M. (2016). Variational graph autoencoders. *NIPS Workshop on Bayesian Deep Learning*.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media.

Maurer, A. (2014). A chain rule for the expected suprema of gaussian processes. In *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014, Proceedings*, volume 8776, page 245. Springer.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer.

Maurer, A. and Pontil, M. (2016). Bounds for vector-valued function estimation. *arXiv preprint arXiv:1606.01487*.

Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural computation*, 17(1):177–204.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT press.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pisier, G. (1986). Probabilistic methods in the geometry of banach spaces. In *Probability and analysis*, pages 167–241. Springer.

Salakhutdinov, R. and Hinton, G. (2009). Deep boltzmann machines. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455. PMLR.

Senkene, E. and Tempel'man, A. (1973). Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13(4):665–670.

Su, H., Heinonen, M., and Rousu, J. (2010). Structured output prediction of anti-cancer drug activity. In Dijkstra, T., Tsivtsivadze, E., Marchiori, E., and Heskes, T., editors, *Pattern Recognition in Bioinformatics - 5th IAPR International Conference, PRIB 2010, Proceedings*, volume 6282 of *Lecture Notes in Computer Science*, pages 38–49. Springer.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

A TECHNICAL PROOFS

A.1 Proof of Theorem 1

Let $(\lambda_1, \ldots, \lambda_L) \in \mathbb{R}_+^L$ and $(f_1^*, \ldots, f_L^*)$ a solution to problem (5). Let $s_l = \|f_l^*\|_{\mathcal{H}_l}^2 \ \forall\ l \in [\![L]\!]$. We shall prove that $(f_1^*, \ldots, f_L^*)$ is also a solution to problem (6) for this choice of $(s_1, \ldots, s_L)$. Consider $(f_1, \ldots, f_L)$ satisfying problem (6)'s constraints. $\forall\ l \in [\![L]\!]$, $\|f_l\|_{\mathcal{H}_l}^2 \leq s_l = \|f_l^*\|_{\mathcal{H}_l}^2$. Hence, we have $\sum_{l=1}^L \lambda_l \|f_l\|_{\mathcal{H}_l}^2 \leq \sum_{l=1}^L \lambda_l \|f_l^*\|_{\mathcal{H}_l}^2$. On the other hand, by definition of the f_l^*'s, it holds :

$$V(f_1, \ldots, f_L) + \sum_{l=1}^L \lambda_l \|f_l\|_{\mathcal{H}_l}^2 \ \geq\ V(f_1^*, \ldots, f_L^*) + \sum_{l=1}^L \lambda_l \|f_l^*\|_{\mathcal{H}_l}^2.$$

Thus, we necessarily have: $V(f_1, \ldots, f_d) \ \geq\ V(f_1^*, \ldots, f_d^*)$.

A similar argument can be used for local solutions, details are left to the reader. □

Although this result may appear rather simple, we thought it was worth mentioning as our setting is particularly unfriendly: the objective function V is not assumed to be convex, and the variables f_l are infinite dimensional. As a consequence, in absence of additional assumptions the converse statement (that solutions to problem (6) are also solutions to problem (5) for a suitable choice of λ_l's) is not guaranteed. The proof indeed rely on the existence of Lagrangian multipliers, which has been shown when the variables are finite dimensional (KKT conditions), or when the objective function is assumed to be convex (Bauschke et al., 2011), but is not ensured in our case.

A.2 Proof of Theorem 5

The technical proof is structured as follows.

A.2.1 Standard Rademacher Generalization Bound

Let loss ℓ denote the squared norm on $\mathcal{X}_0$: $\forall x \in \mathcal{X}_0, \ell(x) = \|x\|_{\mathcal{X}_0}^2$. Notice that, on the set considered, the mapping ℓ is $2M$-Lipschitz, and: $\ell(x_i - h(x_i)) - \ell(x_{i'} - h(x_{i'})) \leq 4M^2$. Hence, by applying McDiarmid's inequality, together with standard arguments in the statistical learning literature (symmetrization/randomization tricks, see *e.g.* Theorem 3.1 in Mohri et al. (2012)), one may show that, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\frac{1}{2}\Big(\epsilon(\hat{h}_n) - \epsilon^*\Big) \leq \sup_{h \in \mathcal{H}_{s,t}} |\epsilon(h) - \hat{\epsilon}_n(h)| \leq 2\widehat{\mathscr{R}}_n\Big(\big(\ell \circ (\mathrm{id} - \mathcal{H}_{s,t})\big)(S)\Big) + 12M^2\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \tag{13}$$

The subsequent results shall provide tools to bound the quantity $\widehat{\mathscr{R}}_n\Big(\big(\ell \circ (\mathrm{id} - \mathcal{H}_{s,t})\big)(S)\Big)$ properly.

A.2.2 Operations on the Rademacher Average

As a first go, we state a preliminary lemma that establishes a comparison between Rademacher and Gaussian averages.

Lemma 7. *We have:* $\forall n \geq 1$,

$$\widehat{\mathscr{R}}_n(\mathcal{C}(S)) \leq \sqrt{\frac{\pi}{2}}\ \widehat{\mathscr{G}}_n(\mathcal{C}(S)).$$

Proof. The proof is based on the fact that $\gamma_{i,k}$ and $\sigma_{i,k}\,|\gamma_{i,k}|$ have the same distribution, combined with Jensen's inequality. See also Lemma 4.5 in Ledoux and Talagrand (1991). □

Hence, the application of the lemma above yields:

$$\widehat{\mathscr{R}}_n\Big((\ell \circ (\mathrm{id} - \mathcal{H}_{s,t}))(S)\Big) \leq 2\sqrt{2}M\ \widehat{\mathscr{R}}_n\Big((\mathrm{id} - \mathcal{H}_{s,t})(S)\Big), \tag{14}$$

$$\leq 2\sqrt{2}M\ \left[\widehat{\mathscr{R}}_n\Big(\{\mathrm{id}\}(S)\Big) + \widehat{\mathscr{R}}_n\Big(\mathcal{H}_{s,t}(S)\Big)\right],$$

$$\leq 2\sqrt{2}M\ \widehat{\mathscr{R}}_n\Big(\mathcal{H}_{s,t}(S)\Big),$$

$$\widehat{\mathscr{R}}_n\Big((\ell \circ (\mathrm{id} - \mathcal{H}_{s,t}))(S)\Big) \leq 2\sqrt{\pi}M\ \widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S)\Big), \tag{15}$$

where (14) directly results from Corollary 4 in Maurer (2016) (observing that, even if they do not take their values in $\ell_2(\mathbb{N})$ but in the separable Hilbert space $\mathcal{X}_0$, the functions $h(x)$ can replaced by the square-summable sequence $(\langle h(x), e_k\rangle)_{k\in\mathbb{N}}$) and (15) is a consequence of Lemma 7.

It now remains to bound $\widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S)\Big)$ using an extension of a result established in Maurer (2014) and applying to classes of functions valued in $\mathbb{R}^m$ only, while functions in $\mathcal{H}_{s,t}$ are Hilbert-valued.

A.2.3 Extension of Maurer's Chain Rule

The result stated below extends Theorem 2 in Maurer (2014) to the Hilbert-valued situation.

Theorem 8. *Let H be a Hilbert space, X a H-valued Gaussian random vector, and $f : H \to \mathbb{R}$ a L-Lipschitz mapping. We have:*

$$\forall t > 0, \qquad \mathbb{P}\Big(|f(X) - \mathbb{E}f(X)| > t\Big) \leq \exp\left(-\frac{2t^2}{\pi^2 L^2}\right).$$

Proof. It is a direct extension of Corollary 2.3 in Pisier (1986), which states the result for $H = \mathbb{R}^N$ only, observing that the proof given therein actually makes no use of the assumption of finite dimensionality of H, and thus remains valid in our case. Up to constants, it can also be viewed an extension of Theorem 4 in Maurer (2014). □

We now introduce quantities involved in the rest of the analysis, see Definition 1 in Maurer (2014).

Definition 9. *Let $Y \subset \mathbb{R}^n$, H be a Hilbert space, $Z \subset H$, and γ be a H-valued standard Gaussian variable/process. We set:*

$$D(Y) = \sup_{y,y'\in Y} \|y - y'\|_{\mathbb{R}^n},$$

$$G(Z) = \sup_{z\in Z} \mathbb{E}_\gamma\left[\langle \gamma, z\rangle_H\right].$$

If $\mathcal{H}$ a class of functions from Y to H, we set:

$$L(\mathcal{H}, Y) = \sup_{h\in\mathcal{H}} \sup_{y,y'\in Y,\ y\neq y'} \frac{\|h(y) - h(y')\|_H}{\|y - y'\|_{\mathbb{R}^n}},$$

$$R(\mathcal{H}, Y) = \sup_{y,y'\in Y,\ y\neq y'} \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{h\in\mathcal{H}} \frac{\langle \boldsymbol{\gamma}, h(y) - h(y')\rangle_H}{\|y - y'\|_{\mathbb{R}^n}}\right].$$

The next result establishes useful relationships between the quantities introduced above.

Theorem 10. *Let $Y \subset \mathbb{R}^n$ be a finite set, H a Hilbert space and $\mathcal{H}$ a finite class of functions $h : Y \to H$. Then, there are universal constants C_1 and C_2 such that, for any $y_0 \in Y$:*

$$G(\mathcal{H}(Y)) \leq C_1 L(\mathcal{H}, Y) G(Y) + C_2 R(\mathcal{H}, Y) D(Y) + G(\mathcal{H}(y_0)).$$

Proof. This result is a direct extension of Theorem 2 in Maurer (2014) for H-valued functions. The only part in the proof depending on the dimensionality of H is Theorem 4 in the same paper, whose extension to any Hilbert space in Theorem 8 is proved in the present paper. Indeed, considering $X_y = (\sqrt{2}/\pi L(F, Y)) \sup_{f\in F} \langle \gamma, f(y)\rangle$ (using the same notation as in Maurer (2014)) allows to finish the proof like in the finite dimensional case. □

Let $\mathcal{H}'_{1,s}$ be the set of functions from $(\mathcal{X}_0)^n$ to $\mathbb{R}^{np}$ that take as input $S = (x_1, \ldots, x_n)$ and return $(f(x_1), \ldots, f(x_n))$, $f \in \mathcal{H}_{1,s}$. Let $Y = \mathcal{H}'_{1,s}(S) \subset \mathbb{R}^{np}$, and $H = (\mathcal{X}_0)^n$, which is a Hilbert space. Let $\mathcal{H} = \mathcal{H}'_{2,t}$ be the set of functions from $\mathbb{R}^{np}$ to $(\mathcal{X}_0)^n$ that take as input $(y_1, \ldots, y_n)$ and return $(g(y_1), \ldots, g(y_n))$, $g \in \mathcal{H}_{2,t}$. Finally, let $y_0 = (0_{\mathbb{R}^p}, \ldots, 0_{\mathbb{R}^p})$ (it actually belongs to $\mathcal{H}'_{1,s}(S)$ since the null function is in $\mathcal{H}'_{1,s}$). Theorem 10 entails that:

$$G\Big(\mathcal{H}'_{2,t}(\mathcal{H}'_{1,s}(S))\Big) \leq C_1 L\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big) G\Big(\mathcal{H}'_{1,s}(S)\Big) + C_2 R\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big) D\Big(\mathcal{H}'_{1,s}(S)\Big) + G\Big(\mathcal{H}'_{2,t}(0)\Big),$$

and

$$\widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S)\Big) \leq C_1 L\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big) \widehat{\mathscr{G}}_n\Big(\mathcal{H}_{1,s}(S)\Big) + \frac{C_2}{n} R\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big) D\Big(\mathcal{H}'_{1,s}(S)\Big) + \frac{1}{n} G\Big(\mathcal{H}'_{2,t}(0)\Big). \quad (16)$$

We now bound each term appearing on the right-hand side.

Bounding $L\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big)$. Consider the following hypothesis, denoting by $\|.\|_*$ the operator norm of any bounded linear operator.

Assumption 11. *There exists a constant $L < +\infty$ such that: $\forall (y, y') \in \mathbb{R}^p$,*

$$\big\|\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\big\|_* \leq L^2 \ \|y - y'\|^2_{\mathbb{R}^p}.$$

This assumption is not too much compelling since it is enough for $\mathcal{K}_2$ to be the sum of M decomposable kernels $k_m(\cdot,\cdot)A_m$ such that the scalar feature maps ϕ_m are L_m-Lipschitz (the feature map of the Gaussian kernel with bandwidth $1/(2\sigma^2)$ has Lipschitz constant $1/\sigma$ for instance), and the A_m operators have finite operator norms σ_m. Indeed, we would have then: $\forall z \in \mathcal{X}_0$,

$$\begin{aligned}
\Big\|\Big(\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\Big)z\Big\|_{\mathcal{X}_0} &= \left\|\left(\sum_{m=1}^M \|\phi_m(y) - \phi_m(y')\|^2 A_m\right) z\right\|_{\mathcal{X}_0}, \\
&\leq \sum_{m=1}^M \|\phi_m(y) - \phi_m(y')\|^2 \sigma_m \ \|z\|_{\mathcal{X}_0}, \\
\Big\|\Big(\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\Big)z\Big\|_{\mathcal{X}_0} &\leq \left(\sum_{m=1}^M L_m^2 \sigma_m\right) \|y - y'\|^2_{\mathbb{R}^p} \ \|z\|_{\mathcal{X}_0}, \\
\|\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\|_* &\leq \left(\sum_{m=1}^M L_m^2 \sigma_m\right) \|y - y'\|^2_{\mathbb{R}^p}.
\end{aligned}$$

Let $\mathcal{K}_2$ satisfy Assumption 11, $g \in \mathcal{H}'_{2,t}$ and $(\boldsymbol{y}, \boldsymbol{y'}) \in \mathbb{R}^{np}$. We have:

$$
\begin{aligned}
\|g(\boldsymbol{y}) - g(\boldsymbol{y'})\|^2_{(\mathcal{X}_0)^n} &= \sum_{i=1}^n \|g(y_i) - g(y'_i)\|^2_{\mathcal{X}_0}\,, \\
&= \sum_{i=1}^n \langle g(y_i) - g(y'_i), g(y_i) - g(y'_i)\rangle_{\mathcal{X}_0}\,, \\
&= \sum_{i=1}^n \langle \mathcal{K}_{2y_i}(g(y_i) - g(y'_i)), g\rangle_{\mathcal{H}_2} - \left\langle \mathcal{K}_{2y'_i}(g(y_i) - g(y'_i)), g\right\rangle_{\mathcal{H}_2}\,, \qquad (17) \\
&\leq \|g\|_{\mathcal{H}_2} \sum_{i=1}^n \left\| \mathcal{K}_{2y_i}(g(y_i) - g(y'_i)) - \mathcal{K}_{2y'_i}(g(y_i) - g(y'_i))\right\|_{\mathcal{H}_2}\,, \qquad (18) \\
&\leq t \sum_{i=1}^n \sqrt{\langle g(y_i) - g(y'_i), \big(\mathcal{K}_2(y_i, y_i) - 2\mathcal{K}_2(y_i, y'_i) + \mathcal{K}_2(y'_i, y'_i)\big)(g(y_i) - g(y'_i))\rangle_{\mathcal{X}_0}}, \qquad (19) \\
&\leq Lt \sum_{i=1}^n \|g(y_i) - g(y'_i)\|_{\mathcal{X}_0} \|y_i - y'_i\|_{\mathbb{R}^p}, \qquad (20) \\
\|g(\boldsymbol{y}) - g(\boldsymbol{y'})\|^2_{(\mathcal{X}_0)^n} &\leq Lt \ \|g(\boldsymbol{y}) - g(\boldsymbol{y'})\|_{(\mathcal{X}_0)^n} \|\boldsymbol{y} - \boldsymbol{y'}\|_{\mathbb{R}^{np}}\,, \qquad (21) \\
\|g(\boldsymbol{y}) - g(\boldsymbol{y'})\|_{(\mathcal{X}_0)^n} &\leq Lt \ \|\boldsymbol{y} - \boldsymbol{y'}\|_{\mathbb{R}^{np}}\,,
\end{aligned}
$$

where (17) results from the reproducing property in vv-RKHSs (see Eq. (2.1) in Micchelli and Pontil (2005)), (18) follows from Cauchy-Schwarz inequality, (19) is again a consequence of the reproducing property (Eq. (2.3) in Micchelli and Pontil (2005)), (20) can be deduced from Assumption 11 and (21) is a consequence of Cauchy-Schwarz inequality as well. Hence, we finally have:

$$
L\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big) \leq L\Big(\mathcal{H}'_{2,t}, \mathbb{R}^{np}\Big) \leq Lt. \qquad (22)
$$

Bounding $\widehat{\mathscr{G}}_n\Big(\mathcal{H}'_{1,s}(S)\Big)$. Consider the assumption below.

Assumption 12. *There exists a constant $K < +\infty$ such that: $\forall x \in \mathcal{X}_0$,*

$$
\mathbf{Tr}\Big(\mathcal{K}_1(x, x)\Big) \leq Kp.
$$

This assumption is mild as well, since the sum of M decomposable kernels $k_m(\cdot, \cdot)A_m$ such that the scalar kernels are bounded by κ_m (as X is supposed to be bounded, any continuous kernel is valid). Indeed, we have: $\forall x \in \mathcal{X}_0$,

$$
\mathbf{Tr}\Big(\mathcal{K}_1(x, x)\Big) = \sum_{m=1}^M k_m(x, x)\ \mathbf{Tr}(A_m) \leq \left(\sum_{m=1}^M \kappa_m \|A_m\|_\infty\right) p.
$$

Let the OVK $\mathcal{K}_1$ satisfy Assumption 12 and be such that $\mathcal{H}_1$ is separable. We then know that there exists $\Phi \in \mathcal{L}(\ell_2(\mathbb{N}), \mathbb{R}^p)$ such that: $\forall (x, x') \in \mathcal{X}_0$, $\mathcal{K}_1(x, x') = \Phi(x)\Phi^*(x')$ and $\forall f \in \mathcal{H}_1, \exists u \in \ell_2(\mathbb{N})$ such that

$f(\cdot) = \Phi(\cdot)u, \quad \|f\|_{\mathcal{H}_1} = \|u\|_{\ell_2}$ (see Micchelli and Pontil (2005)). We have:

$$
\begin{aligned}
n\,\widehat{\mathscr{G}}_n\Big(\mathcal{H}'_{1,s}(S)\Big) &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{f\in\mathcal{H}_{1,s}}\sum_{i=1}^{n}\langle\boldsymbol{\gamma}_i, f(x_i)\rangle_{\mathbb{R}^p}\right], \\
&= \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{\|u\|_{\ell_2}\leq s}\sum_{i=1}^{n}\sum_{k=1}^{p}\gamma_{i,k}\langle\Phi(x_i)u, e_k\rangle_{\mathbb{R}^p}\right], \\
&= \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{\|u\|_{\ell_2}\leq s}\left\langle u, \sum_{i=1}^{n}\sum_{k=1}^{p}\gamma_{i,k}\Phi^*(x_i)e_k\right\rangle_{\ell_2}\right], \\
&\leq s\;\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|\sum_{i=1}^{n}\sum_{k=1}^{p}\gamma_{i,k}\Phi^*(x_i)e_k\right\|_{\ell_2}\right], &(23)\\
&\leq s\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|\sum_{i=1}^{n}\sum_{k=1}^{p}\gamma_{i,k}\Phi^*(x_i)e_k\right\|_{\ell_2}^2\right]}, &(24)\\
&\leq s\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{p}\langle\mathcal{K}(x_i,x_i)e_k, e_k\rangle_{\mathbb{R}^p}}, &(25)\\
&\leq s\sqrt{\sum_{i=1}^{n}\mathbf{Tr}\Big(\mathcal{K}_1(x_i,x_i)\Big)}, &(26)\\
n\,\widehat{\mathscr{G}}_n(\mathcal{H}'_{1,s}(S)) &\leq s\sqrt{nKp}, &(27)
\end{aligned}
$$

where (23) follows from Cauchy-Schwarz inequality, (24) from Jensen's inequality, (25) results from the orthogonality of the Gaussian variables introduced and (27) from Assumption 12. Finally, we have:

$$
\widehat{\mathscr{G}}_n\Big(\mathcal{H}'_{1,s}(S)\Big) \leq s\sqrt{\frac{Kp}{n}}. \tag{28}
$$

Bounding $R\Big(\mathcal{H}'_{2,t}, \mathcal{H}'_{1,s}(S)\Big)$. Consider the following hypothesis.

Assumption 13. *There exists a constant $L < +\infty$ such that: $\forall (y, y') \in \mathbb{R}^p$,*

$$
\mathbf{Tr}\Big(\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\Big) \leq L^2\;\|y - y'\|^2_{\mathbb{R}^p}.
$$

Suppose that the OVK $\mathcal{K}_2$ is the sum of M decomposable kernels $k_m(\cdot,\cdot)A_m$ such that the scalar feature maps ϕ_m are L_m-Lipschitz and the A_m operators are trace class. Then, we have: $\forall (y, y') \in \mathbb{R}^p$,

$$
\mathbf{Tr}\Big(\mathcal{K}_2(y,y) - 2\mathcal{K}_2(y,y') + \mathcal{K}_2(y',y')\Big) = \sum_{m=1}^{M}\|\phi_m(y) - \phi_m(y')\|^2\;\mathbf{Tr}(A_m) \leq \left(\sum_{m=1}^{M}L_m^2\mathbf{Tr}(A_m)\right)\|y - y'\|^2_{\mathbb{R}^p}.
$$

Note also that Assumption 13 is stronger than Assumption 11, since $\|A\|_* \leq \mathbf{Tr}(A)$ for any trace class operator A.

Let the OVK $\mathcal{K}_2$ satisfy Assumption 13 and be such that $\mathcal{H}_2$ is separable. We then know that there exists $\Psi \in \mathcal{L}(\ell_2(\mathbb{N}), \mathcal{X}_0)$ such that $\forall (y, y') \in \mathbb{R}^p$, $\mathcal{K}_2(y, y') = \Psi(y)\Psi^*(y')$ and $\forall g \in \mathcal{H}_2, \exists v \in \ell_2(\mathbb{N})$ such that $g(\cdot) =$

$\Psi(\cdot)v, \quad \|g\|_{\mathcal{H}_2} = \|v\|_{\ell_2}$. We have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{g\in\mathcal{H}_{2,t}} \langle \boldsymbol{\gamma}_i, g(\boldsymbol{y} - g(\boldsymbol{y'})\rangle_{\mathcal{X}_0^n}\right] &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{g\in\mathcal{H}_{2,t}} \sum_{i=1}^{n}\sum_{k=1}^{\infty} \gamma_{i,k} \langle (\Psi(y_i) - \Psi(y_i'))v, e_k\rangle_{\mathcal{X}_0}\right], \\
&= \mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{g\in\mathcal{H}_{2,t}} \left\langle \sum_{i=1}^{n}\sum_{k=1}^{\infty} \gamma_{i,k}(\Psi^*(y_i) - \Psi^*(y_i'))e_k, v\right\rangle_{\ell_2}\right], \\
&\leq t\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}\left\|\sum_{i=1}^{n}\sum_{k=1}^{\infty} \gamma_{i,k}(\Psi^*(y_i) - \Psi^*(y_i'))e_k\right\|_{\ell_2}^2}, \\
&\leq t\sqrt{\sum_{i=1}^{n} \mathbf{Tr}\Big(\mathcal{K}_2(y_i, y_i) - 2\mathcal{K}_2(y_i, y_i') + \mathcal{K}_2(y_i', y_i')\Big)}, \\
\mathbb{E}_{\boldsymbol{\gamma}}\left[\sup_{g\in\mathcal{H}_{2,t}} \langle \boldsymbol{\gamma}_i, g(\boldsymbol{y} - g(\boldsymbol{y'})\rangle_{\mathcal{X}_0^n}\right] &\leq tL\ \|\boldsymbol{y} - \boldsymbol{y'}\|_{\mathbb{R}^{np}},
\end{aligned}$$

where only Assumption 13 and arguments previously involved have been used. Finally, we get:

$$R\Big(\mathcal{H}_{2,t}', \mathcal{H}_{1,s}'(S)\Big) \leq R\Big(\mathcal{H}_{2,t}', \mathbb{R}^{np}\Big) \leq tL. \tag{29}$$

Bounding $D\Big(\mathcal{H}_{1,s}'(S)\Big)$. Consider the assumption below.

Assumption 14. *There exists $\kappa < +\infty$ such that: $\forall x \in S$,*

$$\|\mathcal{K}_1(x, x)\|_* \leq \kappa^2.$$

This assumption is easily fulfilled, since X is almost surely bounded. Indeed, any ov-kernel which is the (finite) sum of decomposable kernels with continuous scalar kernels fulfills it. Note also that it is a weaker assumption than Assumption 12, since one could choose $\kappa = \sqrt{Kp}$.

Let $\mathcal{K}_1$ satisfy Assumption 14 and $(\boldsymbol{y}, \boldsymbol{y'}) \in \mathcal{H}_{1,s}'(S)$. There exists $(f, f') \in \mathcal{H}_{1,s}$ such that $\boldsymbol{y} = (f(x_1), \ldots, f(x_n))$ and $\boldsymbol{y'} = (f'(x_1), \ldots, f'(x_n))$. We have:

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{y'}\|_{\mathbb{R}^{np}}^2 &= \sum_{i=1}^{n} \|f(x_i) - f'(x_i)\|_{\mathbb{R}^p}^2, \\
&\leq \sum_{i=1}^{n} \left(\|f(x_i)\|_{\mathbb{R}^p} + \|f'(x_i)\|_{\mathbb{R}^p}\right)^2, \\
&\leq \sum_{i=1}^{n} \left(\|f\|_{\mathcal{H}_1} \|\mathcal{K}_1(x_i, x_i)\|_*^{1/2} + \|f'\|_{\mathcal{H}_1} \|\mathcal{K}_1(x_i, x_i)\|_*^{1/2}\right)^2, \qquad (30) \\
\|\boldsymbol{y} - \boldsymbol{y'}\|_{\mathbb{R}^{np}}^2 &\leq 4\kappa^2 s^2 n,
\end{aligned}$$

where (30) follows from Eq. (f) of Proposition 2.1 in Micchelli and Pontil (2005). Finally, we get:

$$D\Big(\mathcal{H}_{1,s}', S\Big) \leq 2\kappa s\sqrt{n}. \tag{31}$$

Bounding $G\Big(\mathcal{H}_{2,t}'(0)\Big)$. We introduce the following assumption.

Assumption 15. *$\mathcal{K}_2(0, 0)$ is trace class.*

Then, using the same arguments as for (26), we get:

$$n\ G\Big(\mathcal{H}_{2,t}'(0)\Big) \leq t\sqrt{n\ \mathbf{Tr}\Big(\mathcal{K}_2(0,0)\Big)}, \qquad \text{or} \qquad G\Big(\mathcal{H}_{2,t}'(0)\Big) \leq t\sqrt{\frac{\mathbf{Tr}\big(\mathcal{K}_2(0,0)\big)}{n}}.$$

Rather than shifting the kernel $\widetilde{\mathcal{K}}_2(y,y') = \mathcal{K}_2(y,y') - \mathcal{K}_2(0,0)$, one could consider that Assumption 15 is always satisfied. In addition, we have $\mathbf{Tr}\Big(\widetilde{\mathcal{K}}_2(0,0)\Big) = 0$ and consequently $G\Big(\mathcal{H}'_{2,t}(0)\Big) \leq 0$.

A.2.4 Final Argument

Now, combining inequalities (13), (15), (16), (22), (28), (29), (31) and defining $C_0 \coloneqq 8\sqrt{\pi}(C_1 + 2C_2)$, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\epsilon(\hat{h}_n) - \epsilon^* \leq C_0 LMst\sqrt{\frac{Kp}{n}} + 24M^2\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

□

A.3 Proof of Theorem 6

Lemma 16. *See Theorem 3.1 in Micchelli and Pontil (2005). Let $\mathcal{X}$ be a measurable space, $\mathcal{Y}$ a real Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, $\mathcal{K}: \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{Y})$ an operator-valued kernel, $\mathcal{H} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ the corresponding vv-RKHS, with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. We have the reproducing property : $\langle y, f(x)\rangle_{\mathcal{Y}} = \langle \mathcal{K}_x y, f\rangle_{\mathcal{H}}$, with the notation $\mathcal{K}_x y = \mathcal{K}(\cdot,x)y : \mathcal{X} \to \mathcal{Y}$. Suppose also that the linear functionals $L_{x_i} f = f(x_i), f \in \mathcal{H}, i \in [\![n]\!]$ are linearly independent. Then the unique solution to the variational problem:*

$$\min_{f\in\mathcal{H}} \Big\{ \|f\|^2_{\mathcal{H}} : f(x_i) = y_i,\ i \in [\![n]\!] \Big\},$$

is given by :

$$\hat{f} = \sum_{i=1}^n \mathcal{K}_{x_i} c_i,$$

where $\{c_i,\ i \in [\![n]\!]\} \subset \mathcal{Y}^n$ is the unique solution of the linear system of equations :

$$\sum_{i=1}^n \mathcal{K}(x_k, x_i)c_i = y_k, \qquad k \in [\![n]\!].$$

Proof. Let $f \in \mathcal{H}$ such that $f(x_i) = y_i \ \ \forall\ i \in [\![n]\!]$, and set $g = f - \hat{f}$. We have :

$$\|f\|^2_{\mathcal{H}} = \|\hat{f}\|^2_{\mathcal{H}} + \|g\|^2_{\mathcal{H}} + 2\langle \hat{f}, g\rangle_{\mathcal{H}}.$$

Observe also that :

$$\langle \hat{f}, g\rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^n \mathcal{K}_{x_i} c_i, g \right\rangle_{\mathcal{H}} = \sum_{i=1}^n \langle \mathcal{K}_{x_i} c_i, g\rangle_{\mathcal{H}} = \sum_{i=1}^n \langle c_i, g(x_i)\rangle_{\mathcal{Y}} = 0.$$

Finally, we have :

$$\|f\|^2_{\mathcal{H}} = \|\hat{f}\|^2_{\mathcal{H}} + \|g\|^2_{\mathcal{H}} \geq \|\hat{f}\|^2_{\mathcal{H}}.$$

□

Proof of Theorem 6. We shall use the following shortcut notation:

$$\xi(f_1^*, \ldots, f_{L_0}^*, \mathcal{S}) \coloneqq V\Big((f_{L_0} \circ \ldots \circ f_1)(x_1), \ldots, (f_{L_0} \circ \ldots \circ f_1)(x_n), \|f_1\|_{\mathcal{H}_1}, \ldots, \|f_{L_0}\|_{\mathcal{H}_{L_0}}\Big).$$

Let $l_0 \in [\![L_0]\!]$. Let $g_{l_0} \in \mathcal{H}_{l_0}$ such that :

$$g_{l_0}\left(x_i^{*(l_0-1)}\right) = f_{l_0}^*\left(x_i^{*(l_0-1)}\right), \qquad \forall\ i \in [\![n]\!].$$

By definition, we have :

$$\xi(f_1^*, \ldots, f_{l_0}^*, \ldots, f_{L_0}^*, \mathcal{S}) \leq \xi(f_1^*, \ldots, g_{l_0}, \ldots, f_{L_0}^*, \mathcal{S}),$$

thus we necessarily have :

$$\|f_{l_0}^*\|_{\mathcal{H}_{l_0}}^2 \leq \|g_{l_0}\|_{\mathcal{H}_{l_0}}^2.$$

Therefore $f_{l_0}^*$ is a solution to the problem :

$$\min_{f \in \mathcal{H}_{l_0}} \left\{ \|f\|_{\mathcal{H}_{l_0}}^2 : f\left(x_i^{*(l_0-1)}\right) = f_{l_0}^*\left(x_i^{*(l_0-1)}\right),\ i \in [\![n]\!] \right\}.$$

From Lemma 16, there exists $\left(\varphi_{l_0,1}^*, \dots, \varphi_{l_0,n}^*\right) \in \mathcal{X}_{l_0}^n$, such that :

$$f_{l_0}^*(\cdot) = \sum_{i=1}^n \mathcal{K}_{l_0}\left(\ \cdot\ , x_i^{*(l_0-1)}\right) \varphi_{l_0,i}^*.$$

□

A.4 Non-convexity of the Problem

A.4.1 Functional Setting

We prove that problem (2) is not convex by showing that the objective function $(f,g) \mapsto \hat{\epsilon}_n(g \circ f) + \Omega(f,g)$ is not. We denote this application by $\mathcal{O}$ and suppose it is. If it were convex, one would have :

$$\mathcal{O}\left(\kappa(f,g) + (1-\kappa)(f',g')\right) \ \leq\ \kappa\mathcal{O}(f,g) + (1-\kappa)\mathcal{O}(f',g'), \tag{32}$$

for any $\kappa \in [0,1]$ and any functions $f, f', g, g' \in \mathcal{H}_1^2 \times \mathcal{H}_2^2$. Now, consider the particular case where we want to encode a single point ($n=1$) from $\mathcal{X}_0 = \mathbb{R}$ to $\mathcal{X}_1 = \mathbb{R}$, using one single hidden layer ($L=2$). Let $x_1 = 1$, and assume that both kernels are linear : $\mathcal{K}_1(x,x') = xx'$, $\mathcal{K}_2(y,y') = yy'$. $f : x \mapsto \mathcal{K}_1(x,x_1)\varphi = \varphi x$ and $f' : x \mapsto \mathcal{K}_1(x,x_1)\varphi' = \varphi' x$ are elements of $\mathcal{H}_1$ for any coefficients φ, φ'. In the same way, $g : y \mapsto \mathcal{K}_2(y, f(x_1))\psi = \psi f(1) y$ and $g' : y \mapsto \mathcal{K}_2(y, f'(x_1))\psi' = \psi' f'(1) y$ are elements of $\mathcal{H}_2$ for any $\psi, \psi' \in \mathbb{R}^2$.

Therefore, $\mathcal{O}(f,g)$ depends only on φ and ψ. Let $\mathcal{P}$ denote the application from $\mathbb{R}^2$ to $\mathbb{R}$ such that $\mathcal{O}(f,g) = \mathcal{P}(\varphi,\psi)$. Then, one has also $\mathcal{O}(f',g') = \mathcal{P}(\varphi',\psi')$. And finally, it holds :

$$\begin{aligned}
\mathcal{O}\left(\kappa(f,g) + (1-\kappa)(f',g')\right) &= \mathcal{O}\left(\kappa f + (1-\kappa)f', \kappa g + (1-\kappa)g'\right), \\
&= \mathcal{P}\left(\kappa\varphi + (1-\kappa)\varphi', \kappa\psi + (1-\kappa)\psi'\right), \\
\mathcal{O}\left(\kappa(f,g) + (1-\kappa)(f',g')\right) &= \mathcal{P}\left(\kappa(\varphi,\psi) + (1-\kappa)(\varphi',\psi')\right).
\end{aligned}$$

So if (32) were true, in particular it would be true for the specific f, f', g, g' functions we just defined. Hence, the following would hold for any $\varphi, \varphi', \psi, \psi' \in \mathbb{R}^4$:

$$\mathcal{P}\left(\kappa(\varphi,\psi) + (1-\kappa)(\varphi',\psi')\right) \ \leq\ \kappa\mathcal{P}(\varphi,\psi) + (1-\kappa)\mathcal{P}(\varphi',\psi').$$

This is exactly the convexity of $\mathcal{P}$ in (φ,ψ). So the convexity of the objective function in the functional setting (problem (2)) implies the convexity of the objective function in the parametric setting (obtained after application of Theorem 6). In the following section we show that the latest does not even hold, which allows to conclude that neither problem is convex.

A.4.2 Parametric Setting

As a reminder, we have :

$$\begin{aligned}
f(x) &= \mathcal{K}_1(x,x_1)\varphi = \varphi x, & f(1) &= \varphi, \\
g(y) &= \mathcal{K}_2(y, f(x_1))\psi = \varphi\psi y, & g(f(1)) &= \varphi^2\psi.
\end{aligned}$$

Our problem reads :

$$\min_{\varphi \in \mathbb{R},\ \psi \in \mathbb{R}} \ \mathcal{P}(\varphi,\psi) \overset{def}{=} \left(1 - \varphi^2\psi\right)^2 + \lambda\varphi^2 + \mu\psi^2,$$

or equivalently :

$$\min_{\varphi\in\mathbb{R},\ \psi\in\mathbb{R}} \quad 1+\lambda\varphi^2+\mu\psi^2-2\varphi^2\psi+\varphi^4\psi^2.$$

Let us find the critical points and analyze them. We have :

$$\begin{aligned}
\frac{\partial\mathcal{P}}{\partial\varphi}(\varphi,\psi) &= 2\lambda\varphi-4\varphi\psi+4\varphi^3\psi^2,\\
\frac{\partial\mathcal{P}}{\partial^2\varphi}(\varphi,\psi) &= 2\lambda-4\psi+12\varphi^2\psi^2,\\
\frac{\partial\mathcal{P}}{\partial\psi}(\varphi,\psi) &= 2\mu\psi-2\varphi^2+2\varphi^4\psi,\\
\frac{\partial\mathcal{P}}{\partial^2\psi}(\varphi,\psi) &= 2\mu+2\varphi^4,\\
\frac{\partial\mathcal{P}}{\partial\varphi\partial\psi}(\varphi,\psi) &= -4\varphi+8\varphi^3\psi.
\end{aligned}$$

The two following equivalence relationships hold true:

$$\frac{\partial\mathcal{P}}{\partial\varphi}(\varphi^*,\psi^*)=\left(2\lambda-4\psi^*+4{\varphi^*}^2{\psi^*}^2\right)\varphi^*=0 \qquad\Leftrightarrow\qquad \varphi^*=0 \ \text{ or } \ {\varphi^*}^2=\frac{2\psi^*-\lambda}{2{\psi^*}^2},$$

$$\frac{\partial\mathcal{P}}{\partial\psi}(\varphi^*,\psi^*)=2\mu\psi^*-2{\varphi^*}^2+2{\varphi^*}^4\psi^*=0 \qquad\Leftrightarrow\qquad \psi^*=\frac{{\varphi^*}^2}{{\varphi^*}^4+\mu}.$$

Obviously, the point $(\varphi^*,\psi^*)=(0,0)$ is always critical. Notice that :

$$\mathbf{Hess}_{(0,0)}\mathcal{P}=\begin{pmatrix}2\lambda & 0\\ 0 & 2\mu\end{pmatrix}\succ 0.$$

Thus $(0,0)$ is a local minimum and $\mathcal{P}(0,0)=1$. To prove that it is not a global minimizer, it is enough to find a couple (φ,ψ) such that $\mathcal{P}(\varphi,\psi)<1$. For example $\mathcal{P}(1,1)=\lambda+\mu$. As soon as $\lambda+\mu<1$, the objective $\mathcal{P}$ is not invex, and a fortiori non-convex.

Figure 4 shows the heatmaps of $\mathcal{P}$ with respect to φ and ψ for different regularization settings. Note that in the non-regularized setting ($\lambda=\mu=0$), every point $(0,\psi)$ with $\psi<0$ is a local minimizer but not a global one. They are represented by red crosses. On the other hand, we have also an infinite number of global minima, namely every couple satisfying $\varphi^2\psi=1$. See the black crosses on the top left figure. When the regularization parameters remain small enough, $(0,0)$ is a local minimizer but not a global one (top right figure). Finally, the higher the regularization, the smoother the objective, even if convexity can never be verified (bottom figures).

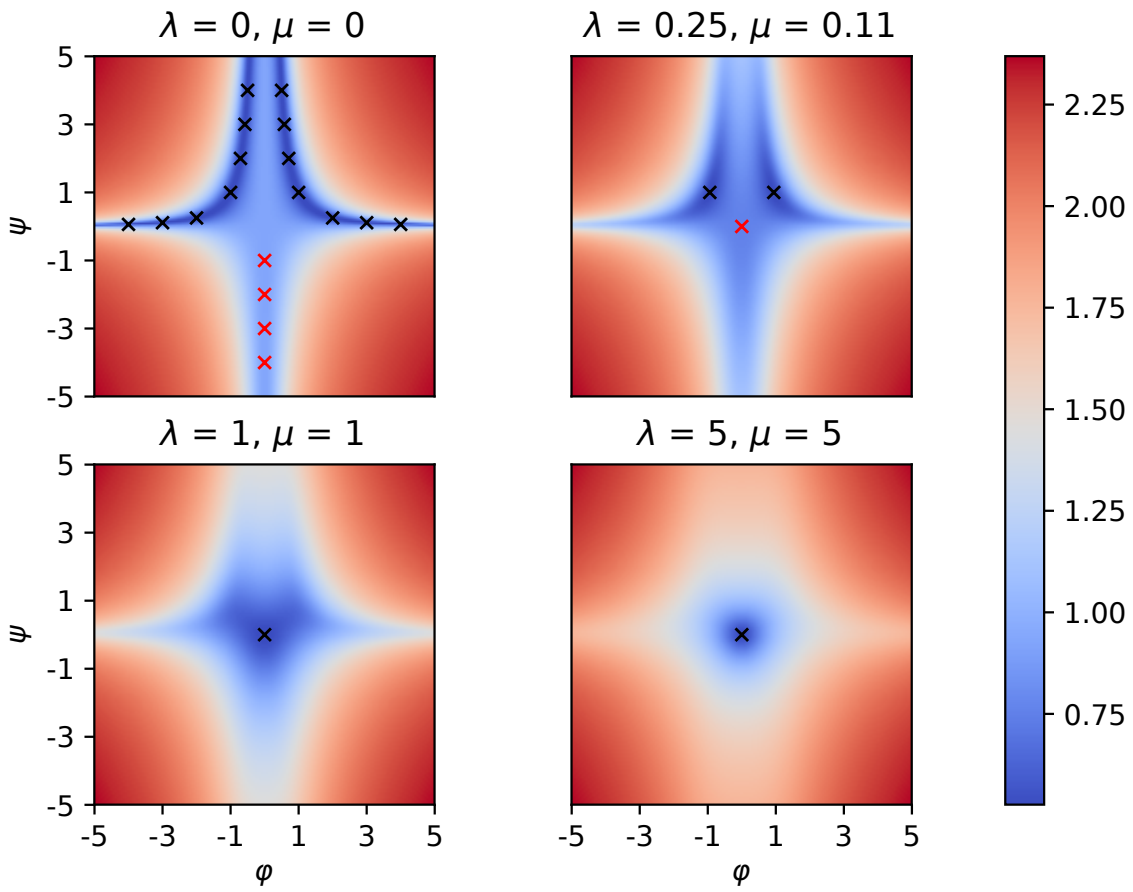


Figure 4: Heatmaps of $\mathcal{P}$ for different values of λ and μ

B Gradient Derivation Details

B.1 Detail of Equation (8)

$$
\begin{aligned}
\|f_l\|^2_{\mathcal{H}_l} &= \langle f_l, f_l \rangle_{\mathcal{H}_l}\,, \\
&= \left\langle \sum_{i=1}^{n} \mathcal{K}_l\left(\,.\,, {x_i}^{(l-1)}\right)\varphi_{l,i}\ ,\ \sum_{i'=1}^{n} \mathcal{K}_l\left(\,.\,, {x_{i'}}^{(l-1)}\right)\varphi_{l,i'} \right\rangle_{\mathcal{H}_l}\,, \\
&= \sum_{i,i'=1}^{n} \left\langle \mathcal{K}_l\left(\,.\,, {x_i}^{(l-1)}\right)\varphi_{l,i}\ ,\ \mathcal{K}_l\left(\,.\,, {x_{i'}}^{(l-1)}\right)\varphi_{l,i'} \right\rangle_{\mathcal{H}_l}\,, \\
&= \sum_{i,i'=1}^{n} \left\langle \varphi_{l,i}\ ,\ \mathcal{K}_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\varphi_{l,i'} \right\rangle_{\mathcal{X}_l}\,, \\
\|f_l\|^2_{\mathcal{H}_l} &= \sum_{i,i'=1}^{n} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\langle \varphi_{l,i}\ ,\ A_l\ \varphi_{l,i'} \rangle_{\mathcal{X}_l}\,.
\end{aligned}
$$

□

B.2 Detail of Equation (10)

$$
\begin{aligned}
\left(\nabla_{\varphi_{l_0,i_0}}\ \|f_l\|^2_{\mathcal{H}_l}\right)^T &= \sum_{i,i'=1}^{n} [N_l]_{i,i'} \left(\nabla_{\varphi_{l_0,i_0}}\ k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T\,, \\
&= \sum_{i,i'=1}^{n} [N_l]_{i,i'} \Bigg[\left(\nabla^{(1)} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}) \\
&\qquad + \left(\nabla^{(2)} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0}) \Bigg], \\
&= \sum_{i,i'=1}^{n} [N_l]_{i,i'} \left(\nabla^{(1)} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}) \\
&\quad + \sum_{i',i=1}^{n} [N_l]_{i',i} \left(\nabla^{(1)} k_l\left({x_{i'}}^{(l-1)}, {x_i}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0}), \\
\left(\nabla_{\varphi_{l_0,i_0}}\ \|f_l\|^2_{\mathcal{H}_l}\right)^T &= 2 \sum_{i,i'=1}^{n} [N_l]_{i,i'} \left(\nabla^{(1)} k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}),
\end{aligned}
$$

where $\nabla^{(1)} k_l\,(x, x')$ (respectively $\nabla^{(2)} k_l\,(x, x')$) denotes the gradient of $k_l(\cdot,\cdot)$ with respect to the 1^{st} (respectively 2^{nd}) coordinate evaluated in (x, x'). □

B.3 Detail of Jacobians Computation

All previously written gradients involve Jacobian matrices. Their computation is to be detailed in this subsection. First note that $\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0})$ only makes sense if $l_0 \leq l$. Indeed, ${x_i}^{(l)}$ is completely independent from φ_{l_0,i_0} otherwise. Let us first detail ${x_i}^{(l)}$ and use the linearity of the Jacobian operator :

$$
\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0}) = \sum_{i'=1}^{n} \mathbf{Jac}_{k_l\left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)}\right) A_l\ \varphi_{l,i'}}(\varphi_{l_0,i_0}).
$$

Just as in the norm gradient case (see Section 4.2), there are two different outputs depending on whether $l = l_0$ (this gives an initialization), or $l > l_0$ (this leads to a recurrence formula).

Own Jacobian ($l = l_0$) :

$$\begin{aligned}
\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l,i_0}) &= \sum_{i'=1}^{n} \mathbf{Jac}_{k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)A_l\ \varphi_{l,i'}}(\varphi_{l,i_0}), \\
&= \sum_{i'=1}^{n} k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)\mathbf{Jac}_{A_l\ \varphi_{l,i'}}(\varphi_{l,i_0}), \\
\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l,i_0}) &= [K_l]_{i,i_0}\ A_l.
\end{aligned}$$

Higher Jacobian ($l > l_0$) :

$$\begin{aligned}
\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0}) &= \sum_{i'=1}^{n} \mathbf{Jac}_{k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)A_l\ \varphi_{l,i'}}(\varphi_{l_0,i_0}), \\
&= \sum_{i'=1}^{n} A_l\ \varphi_{l,i'}\left(\nabla_{\varphi_{l_0,i_0}}\ k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)\right)^T, \\
&= A_l \sum_{i'=1}^{n} \varphi_{l,i'}\Bigg[\ \left(\nabla^{(1)}k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}) \\
&\qquad + \left(\nabla^{(1)}k_l\left({x_{i'}}^{(l-1)},{x_i}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0})\ \Bigg], \\
&= A_l\left[\sum_{i'=1}^{n} \varphi_{l,i'}\left(\nabla^{(1)}k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)\right)^T\right]\mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}) \\
&\quad + A_l\left[\sum_{i'=1}^{n} \varphi_{l,i'}\left(\nabla^{(1)}k_l\left({x_{i'}}^{(l-1)},{x_i}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0})\right], \\
\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0}) &= A_l\Bigg[\Phi_l^T\Delta_l\left({x_i}^{(l-1)}\right)\mathbf{Jac}_{{x_i}^{(l-1)}}(\varphi_{l_0,i_0}) \\
&\qquad + \sum_{i'=1}^{n} \varphi_{l,i'}\left(\nabla^{(1)}k_l\left({x_{i'}}^{(l-1)},{x_i}^{(l-1)}\right)\right)^T \mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0})\Bigg],
\end{aligned}$$

with $\Delta_l(x) \coloneqq \left(\left(\nabla^{(1)}k_l\left(x,{x_1}^{(l-1)}\right)\right)^T,\dots,\left(\nabla^{(1)}k_l\left(x,{x_n}^{(l-1)}\right)\right)^T\right)^T$ the $n \times d_{l-1}$ matrix storing the $\nabla^{(1)}k_l\left(x,{x_i}^{(l-1)}\right)$ in rows. These matrices are computed on Appendix B.4 (especially for $x = {x_i}^{(l-1)}$). Assuming these quantities are known, we have an expression of $\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0})$ that only depends on the $\mathbf{Jac}_{{x_{i'}}^{(l-1)}}(\varphi_{l_0,i_0})$. Thus we can unroll the recurrence until $l = l_0$ and, using the previous subsection, compute $\mathbf{Jac}_{{x_i}^{(l)}}(\varphi_{l_0,i_0})$ for every couple (l, l_0) such that $l > l_0$.

An interesting remark can be made on the two-terms structure of the Jacobians. Indeed, the first term corresponds to the chain rule on ${x_i}^{(l)} = f_l\left({x_i}^{(l-1)}\right)$ assuming that f_l is constant : $\frac{\partial f_l\left({x_i}^{(l-1)}\right)}{\partial\varphi_{l_0,i_0}} = \frac{\partial f_l\left({x_i}^{(l-1)}\right)}{\partial {x_i}^{(l-1)}} \cdot \frac{\partial {x_i}^{(l-1)}}{\partial\varphi_{l_0,i_0}}$ (notation abuse on ∂ in order to preserve understandability). On the contrary, the second term corresponds to a chain rule assuming that ${x_i}^{(l-1)}$ does not vary with φ_{l_0,i_0}, but that f_l does, through the influence of φ_{l_0,i_0} on the supports of f_l, namely the ${x_{i'}}^{(l-1)}$.

B.4 Detail of the Δ_l Matrices Computation

In this section we derive the quantities $\nabla^{(1)}k_l\left({x_i}^{(l-1)},{x_{i'}}^{(l-1)}\right)$ and more specifically the matrices $\Delta_l\left({x_i}^{(l-1)}\right)$ for $l \in [\![L]\!]$ and $i \in [\![n]\!]$. Note that all previously computed quantities are independent from the kernel chosen. Actually, the $\Delta_l\left({x_i}^{(l-1)}\right)$ matrices encapsulate all the kernel specificity of the algorithm. Thus, tailoring a new algorithm by changing the kernels only requires computing the new Δ_l matrices. This flexibility is a key asset of our approach, and more generally a crucial characteristic of kernel methods. In the following, we describe the Δ_l derivation for two popular kernels : the Gaussian and the polynomial ones.

Gaussian kernel :

$$\nabla^{(1)} k_l(x, x') = \nabla_x \Big(\exp\left(-\gamma_l \|x - x'\|^2_{\mathcal{X}_{l-1}}\right) \Big) = -2\gamma_l \; e^{-\gamma_l \|x-x'\|^2_{\mathcal{X}_{l-1}}} \; (x - x').$$

$$\begin{aligned}
\Delta_l \left({x_i}^{(l-1)}\right) &= \left[\left(\nabla^{(1)} k_l \left({x_i}^{(l-1)}, {x_1}^{(l-1)}\right)\right)^T, \ldots, \left(\nabla^{(1)} k_l \left({x_i}^{(l-1)}, {x_n}^{(l-1)}\right)\right)^T \right]^T, \\
&= -2\gamma_l \Big[e^{-\gamma_l \left\| {x_i}^{(l-1)} - {x_1}^{(l-1)} \right\|^2_{\mathcal{X}_{l-1}}} \left({x_i}^{(l-1)} - {x_1}^{(l-1)}\right)^T, \ldots \\
&\qquad \ldots, e^{-\gamma_l \left\| {x_i}^{(l-1)} - {x_n}^{(l-1)} \right\|^2_{\mathcal{X}_{l-1}}} \left({x_i}^{(l-1)} - {x_n}^{(l-1)}\right)^T \Big]^T, \\
\Delta_l \left({x_i}^{(l-1)}\right) &= -2\gamma_l \; \tilde{K}_{l,i} \circ \left(\tilde{X}_i^{(l-1)} - X^{(l-1)}\right),
\end{aligned}$$

where :

- $X^{(l-1)} \coloneqq \left(\left({x_1}^{(l-1)}\right)^T, \ldots, \left({x_n}^{(l-1)}\right)^T\right)^T \in \mathbb{R}^{n \times d_{l-1}}$ stores the level $l-1$ representations of the x_i's in rows
- $\tilde{X}_i^{(l-1)} \coloneqq \left(\left({x_i}^{(l-1)}\right)^T, \ldots, \left({x_i}^{(l-1)}\right)^T\right)^T \in \mathbb{R}^{n \times d_{l-1}}$ stores the level $l-1$ representation of x_i n times in rows
- $\tilde{K}_{l,i} \in \mathbb{R}^{n \times n}$ is the k_l Gram matrix between $X^{(l-1)}$ and $\tilde{X}_i^{(l-1)}$ $\left(i.e.\ [\tilde{K}_{l,i}]_{s,t} = k_l\left({x_i}^{(l-1)}, {x_t}^{(l-1)}\right)\right)$
- $\circ$ denotes the Hadamard (termwise) product for two matrices of the same shape

In practice, it is important to note that computing the Δ_l matrices with the Gaussian kernel needs not new calculations, but only uses already computed quantities : the level $l-1$ representations and their Gram matrix.

Polynomial kernel :

$$\nabla^{(1)} k_l(x, x') = \nabla_x \left(\left(a \langle x, x' \rangle + b\right)^c \right) = ca \Big(\left(a \langle x, x' \rangle + b\right)^{c-1} \Big) x'.$$

$$\begin{aligned}
\Delta_l \left({x_i}^{(l-1)}\right) &= \left[\left(\nabla^{(1)} k_l \left({x_i}^{(l-1)}, {x_1}^{(l-1)}\right)\right)^T, \ldots, \left(\nabla^{(1)} k_l \left({x_i}^{(l-1)}, {x_n}^{(l-1)}\right)\right)^T \right]^T, \\
&= ca \Big[\left(a \left\langle {x_i}^{(l-1)}, {x_1}^{(l-1)} \right\rangle + b\right)^{c-1} \left({x_1}^{(l-1)}\right)^T, \ldots \\
&\qquad \ldots, \left(a \left\langle {x_i}^{(l-1)}, {x_n}^{(l-1)} \right\rangle + b\right)^{c-1} \left({x_n}^{(l-1)}\right)^T \Big]^T, \\
\Delta_l \left({x_i}^{(l-1)}\right) &= ca \; \left(\tilde{K}_{l,i}\right)^{\frac{c-1}{c}} \circ X^{(l-1)},
\end{aligned}$$

where we keep the notations introduced in the Gaussian kernel example for $X^{(l-1)}$, $\tilde{K}_{l,i}$ and $\circ$. Note that the exponent on $\tilde{K}_{l,i}$ must be understood as a termwise power, and not a matrix multiplication power.

In practice, it is important to note that computing the Δ_l matrices with the polynomial kernel only requires a slight and cheap new calculation : putting the - already computed - Gram matrix at layer $l-1$ to the termwise power $(c-1)/c$.

B.5 Detail of N_L Computation

$$\begin{aligned}
\langle x_j, x_{j'} \rangle_{\mathcal{X}_0} &= \Bigg\langle \sum_{i=1}^{n} \left(\mathcal{K}_L \left(x_j^{(L-1)}, x_i^{(L-1)} \right) + n\lambda_L \delta_{ij} \right) \varphi_{L,i} \ , \\
&\qquad \sum_{i'=1}^{n} \left(\mathcal{K}_L \left(x_{j'}^{(L-1)}, x_{i'}^{(L-1)} \right) + n\lambda_L \delta_{i'j'} \right) \varphi_{L,i'} \Bigg\rangle_{\mathcal{X}_0} , \\
&= \sum_{i,i'=1}^{n} \Bigg\langle \left(k_L \left(x_j^{(L-1)}, x_i^{(L-1)} \right) + n\lambda_L \delta_{ij} \right) \varphi_{L,i} \ , \\
&\qquad \left(k_L \left(x_{j'}^{(L-1)}, x_{i'}^{(L-1)} \right) + n\lambda_L \delta_{i'j'} \right) \varphi_{L,i'} \Bigg\rangle_{\mathcal{X}_0} , \\
\langle x_j, x_{j'} \rangle_{\mathcal{X}_0} &= \sum_{i,i'=1}^{n} \left(k_L \left(x_j^{(L-1)}, x_i^{(L-1)} \right) + n\lambda_L \delta_{ij} \right) \\
&\qquad \left(k_L \left(x_{j'}^{(L-1)}, x_{i'}^{(L-1)} \right) + n\lambda_L \delta_{i'j'} \right) \langle \varphi_{L,i}, \varphi_{L,i'} \rangle_{\mathcal{X}_0} . \qquad (33)
\end{aligned}$$

As a reminder, N_L denotes the matrix such that $[N_L]_{i,i'} = \langle \varphi_{L,i}, \varphi_{L,i'} \rangle_{\mathcal{X}_0}$. Let K_{in} denote the input Gram matrix such that $[K_{in}]_{j,j'} = \langle x_j, x_{j'} \rangle_{\mathcal{X}_0}$. Finally, following notations of Section 4.2 for K_L, and denoting I_n the identity matrix on $\mathbb{R}^n$, equation (33) may be rewritten as:

$$[K_{in}]_{j,j'} = \sum_{i,i'=1}^{n} [K_L + n\lambda_L I_n]_{j,i} [N_L]_{i,i'} [K_L + n\lambda_L I_n]_{i',j},$$

or equivalently:

$$K_{in} = (K_L + n\lambda_L I_n) \ N_L \ (K_L + n\lambda_L I_n),$$

so that the computation of the desired linear products $\langle \varphi_{L,i}, \varphi_{L,i'} \rangle_{\mathcal{X}_0}$ becomes straightforward:

$$N_L = (K_L + n\lambda_L I_n)^{-1} \ K_{in} \ (K_L + n\lambda_L I_n)^{-1}. \qquad (34)$$

Since K_L is recursively derived from K_{in} and $\Phi_1, \dots, \Phi_{L-1}$, the optimal matrix N_L (in the sense of the Kernel Ridge Regression) only depends on K_{in}, the coefficient matrices, and the last layer regularization parameter λ_L. Let N_{KRR} be the function that computes N_L of equation (34) from $\Phi_1, \dots, \Phi_{L-1}$, K_{in} and λ_L.

B.6 Detail of Equation (12)

Since $\mathcal{X}_L$ is now infinite dimensional, $\mathbf{Jac}_{x_i^L}(\varphi_{l_0,i_0})$ makes no more sense. Nevertheless, $\varphi_{l,i}$ remains finite dimensional, and the distortion a scalar: a gradient does exist. One is just forced to use the differential of $\|x_i - f_L \circ \dots \circ f_1(x_i)\|^2_{\mathcal{X}_0}$ to make it appear. As a reminder, the chain rule for the differentials reads : $d(g \circ f)(x) = dg(f(x)) \circ df(x)$. Let us apply it with $g(\cdot) = \|\cdot\|^2_{\mathcal{X}_0}$ and $f : \varphi_{l_0,i_0} \mapsto x_i - x_i^{(L)}$. Let $h \in \mathcal{X}_{l_0}$ and $h' \in \mathcal{X}_0$, we have:

$$\Big(dg(y) \Big)(h') = 2 \langle y, h' \rangle_{\mathcal{X}_0} .$$

$$\begin{aligned}
\Big(df(\varphi_{l_0,i_0}) \Big)(h) &= \left(d \left(x_i - \sum_{i'=1}^{n} k_L \left[x_i^{(L-1)}, x_{i'}^{(L-1)} \right] \varphi_{L,i'} \right) (\varphi_{l_0,i_0}) \right) (h), \\
&= - \sum_{i'=1}^{n} \left(d \left(k_L \left[x_i^{(L-1)}, x_{i'}^{(L-1)} \right] \varphi_{L,i'} \right) (\varphi_{l_0,i_0}) \right) (h), \\
&= - \sum_{i'=1}^{n} \left(d \left(k_L \left[x_i^{(L-1)}, x_{i'}^{(L-1)} \right] \right) (\varphi_{l_0,i_0}) \right) (h) \ \varphi_{L,i'}, \\
\Big(df(\varphi_{l_0,i_0}) \Big)(h) &= - \sum_{i'=1}^{n} \left\langle \nabla_{\varphi_{l_0,i_0}} k_L \left(x_i^{(L-1)}, x_{i'}^{(L-1)} \right), h \right\rangle_{\mathcal{X}_{l_0}} \varphi_{L,i'}.
\end{aligned}$$

Combining both expressions with $y = x_i - x_i^{(L)}$ gives:

$$\begin{aligned}
\left(d(\|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0})(\varphi_{l_0,i_0})\right)(h) &= \left(d(g \circ f)(\varphi_{l_0,i_0})\right)(h), \\
&= \left(dg\left(x_i - x_i^{(L)}\right)\right) \circ \left(df(\varphi_{l_0,i_0})\right)(h), \\
&= 2\left\langle x_i - x_i^{(L)}, -\sum_{i'=1}^{n} \left\langle \nabla_{\varphi_{l_0,i_0}} k_L\left(x_i^{(L-1)}, x_{i'}^{(L-1)}\right), h\right\rangle_{\mathcal{X}_{l_0}} \varphi_{L,i'}\right\rangle_{\mathcal{X}_0}, \\
&= -2\sum_{i'=1}^{n} \left\langle \nabla_{\varphi_{l_0,i_0}} k_L\left(x_i^{(L-1)}, x_{i'}^{(L-1)}\right), h\right\rangle_{\mathcal{X}_{l_0}} \left\langle x_i - x_i^{(L)}, \varphi_{L,i'}\right\rangle_{\mathcal{X}_0}, \\
\left(d(\|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0})(\varphi_{l_0,i_0})\right)(h) &= \left\langle -2\sum_{i'=1}^{n} \left\langle x_i - x_i^{(L)}, \varphi_{L,i'}\right\rangle_{\mathcal{X}_0} \nabla_{\varphi_{l_0,i_0}} k_L\left(x_i^{(L-1)}, x_{i'}^{(L-1)}\right), h\right\rangle_{\mathcal{X}_{l_0}}.
\end{aligned}$$

A direct identification leads to equation (12). □

Like in the finite dimensional case, the gradient of the whole criterion is just the (weighted) sum of the gradients of the distortion and the norm penalizations. However, since we assume N_L to be fixed (and known) in order to propagate the gradient, we use the shortcut notation $\nabla_{\Phi_l}(\hat{\epsilon}_n + \Omega \mid N_L)$ in Algorithm 1 to denote the gradient of the whole criterion with respect to Φ_l, assuming that N_L is fixed.

B.7 Solutions to Equations (11) and Test Distortion

Since we have assumed that A_L is the identity operator on $\mathcal{X}_L$, equations (11) simplify to:

$$\forall\, i \in [\![n]\!], \qquad \sum_{i'=1}^{n} W_{i,i'}\ \varphi_{L,i'} = x_i, \tag{35}$$

where $W = K_L + n\lambda_L I_n$. It is then easy to show that the

$$\varphi_{L,i'} = \sum_{i=1}^{n} \left[W^{-1}\right]_{i',i}\ x_i \qquad \forall\, i' \in [\![n]\!]$$

are solutions to equations (35) and therefore to equations (11). Note that using this expansion directly leads to equation (34). But more interestingly, this new writing allows for computing the distortion on a test set. Indeed, let $x \in \mathcal{X}_0$, one has:

$$\begin{aligned}
\|x - f_L \circ \ldots \circ f_1(x)\|^2_{\mathcal{X}_0} &= \left\|x - f_L\left(x^{(L-1)}\right)\right\|^2_{\mathcal{X}_0}, \\
&= \|x\|^2_{\mathcal{X}_0} + \left\|f_L\left(x^{(L-1)}\right)\right\|^2_{\mathcal{X}_0} - 2\left\langle x, f_L\left(x^{(L-1)}\right)\right\rangle_{\mathcal{X}_0}, \\
&= \|x\|^2_{\mathcal{X}_0} + \left\|\sum_{i=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right)\varphi_{L,i}\right\|^2_{\mathcal{X}_0} - 2\left\langle x, \sum_{i=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right)\varphi_{L,i}\right\rangle_{\mathcal{X}_0}, \\
&= \|x\|^2_{\mathcal{X}_0} + \sum_{i,j=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right) k_L\left(x^{(L-1)}, x_j^{(L-1)}\right)\langle\varphi_{L,i}, \varphi_{L,j}\rangle_{\mathcal{X}_0} \\
&\quad - 2\sum_{i=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right)\langle x, \varphi_{L,i}\rangle_{\mathcal{X}_0}, \\
\|x - f_L \circ \ldots \circ f_1(x)\|^2_{\mathcal{X}_0} &= \|x\|^2_{\mathcal{X}_0} + \sum_{i,j=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right) k_L\left(x^{(L-1)}, x_j^{(L-1)}\right)\langle\varphi_{L,i}, \varphi_{L,j}\rangle_{\mathcal{X}_0} \\
&\quad - 2\sum_{i,j=1}^{n} k_L\left(x^{(L-1)}, x_i^{(L-1)}\right)\left[W^{-1}\right]_{i,j}\langle x, x_j\rangle_{\mathcal{X}_0}.
\end{aligned}$$

Just like in Section 4.3 and Appendix B.5, knowing the scalar products in $\mathcal{X}_0$ is the only thing we need to compute the test distortion (all other quantities are finite dimensional and thus computable).

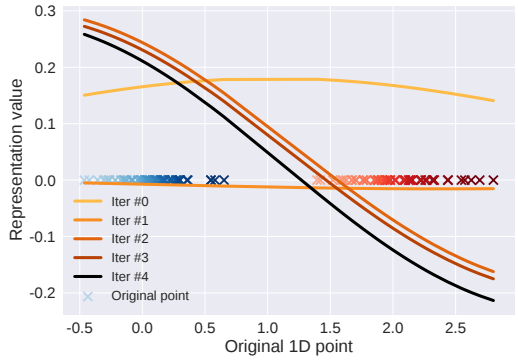
C Additional Experiments

C.1 2D Data

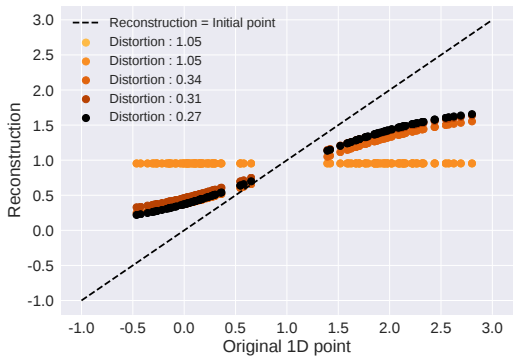
Figure 5 gives a look on the algorithm behavior on 1D data. Results on 1D data are displayed and analyzed here as they are easily understandable. Indeed, one dimension of the plot (the x axis) is used to display the original 1D points (the crosses), while their representations (the $f(x_i)$) vary along the y axis. As soon as the original point or the representation needs more than 1 dimension to be plotted, a 2D plot lacks of dimensions to correctly display the behavior of the algorithm. Original data (to be represented) are sampled from 2 Gaussian distributions, of standard deviation 0.1, and with expected value 0 and 2 respectively.

Figure 5(a) and Figure 5(b) show the evolution of the encoding / decoding functions along the iterations of the algorithm. From the initial yellow representation function, obtained by uniform weights, the algorithm learns the black function, which seems satisfying in two ways. First, the representations of the two clusters are easily separable. Points from the first blue cluster (i.e. drawn from the Gaussian centered at 0) have positive representations, while points from the red one (i.e. drawn from the Gaussian centered at 2) have negative ones. If computed in a clustering purpose, the representation thus gives an easy criterion to distinguish the two clusters. Second, in order to be able to reconstruct any point, one must observe variability within each cluster. This way, the reconstruction function can easily reassign every point. On the contrary, the yellow representation function represents all points by almost the same value, which leads necessarily to a uniform (and bad) reconstruction.

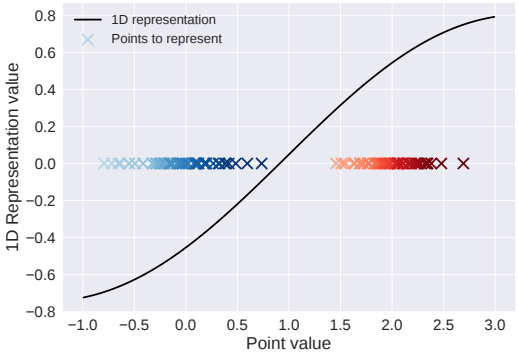
Figure 5(c) shows another 1D representation of the two clusters, while Figure 5(d) shows a 2D encoding of these points. Interestingly, the two components of the 2D representation are highly correlated. This can be interpreted as the fact that a 2D descriptor is over-parameterizing a 1D point.



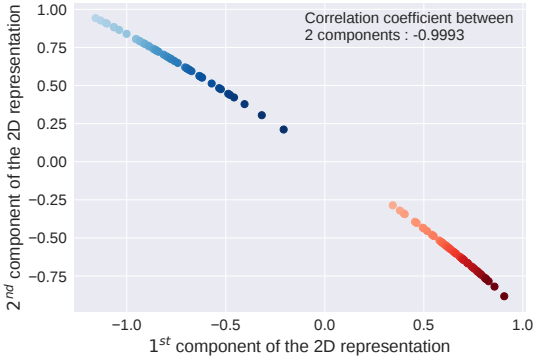
(a) Encoding evolution during fitting



(b) Decoding evolution during fitting



(c) 1D Gaussian clusters and 1D representation

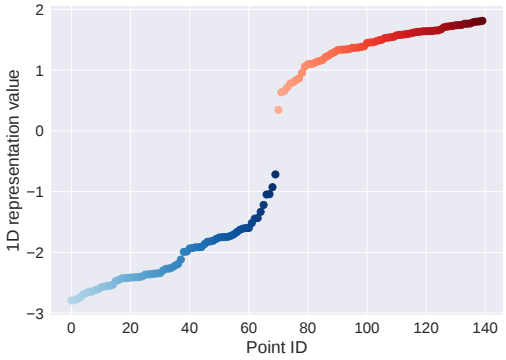


(d) 2D representation of the clusters

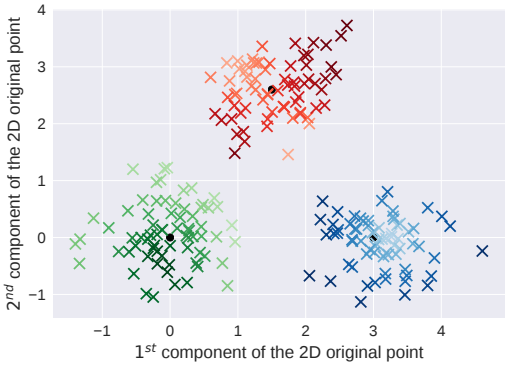
Figure 5: Algorithm behavior on 1D data

Figure 6 shows the algorithm's behavior on Gaussian clusters. Whenever original points and their representations cannot be displayed on the same graph (*i.e.* when whether the original data or its representation is of dimension more than 2), the colormap helps linking them. In Figure 6(a), the original 2D data are plotted, while Figure 6(b) shows their 1D representations. The colormap has been established according to the value of this representation. First, the two clusters remain well separated in the representation space (positive/negative representations). But what is really interesting is how they are separated. The lighter the blue points are, the most negative representation they have, or in other terms, the *most certain* they are to be in the blue cluster. Similarly, the darker the red points are, the most positive representation they have. When looking at these points on Figure 6(a), one sees that it matches the distribution: light blue points are the most distant from the red cluster, and conversely for the dark red ones. The algorithm has found the direction that discriminates the two clusters. Similar results are shown for 3 Gaussian clusters on Figure 6(c) and Figure 6(d).

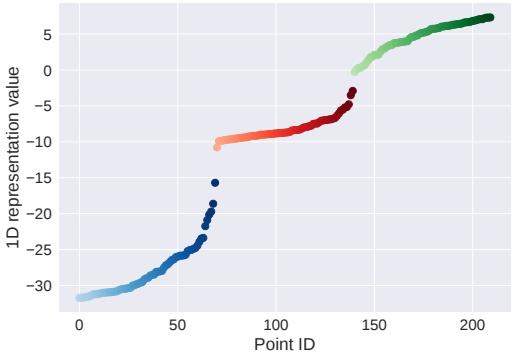
(a) 2D Gaussian clusters



(b) 1D representation of the clusters



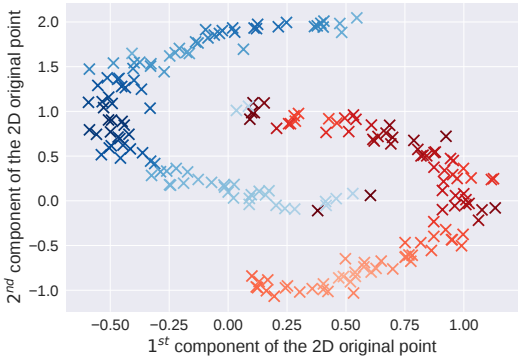
(c) 2D Gaussian clusters (3)



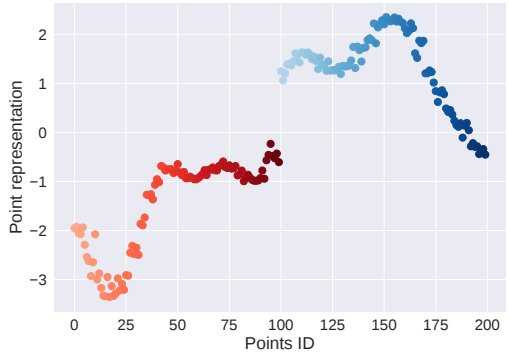
(d) 1D representation of the clusters

Figure 6: Algorithm behavior on Gaussian clusters

Finally, Figure 7 shows the algorithm's behavior on the so called *two moons dataset*. 2D original points (Figure 7(a) and Figure 7(c), colored differently according to the representation on their right) are first mapped to a 1D representation (Figure 7(b)). Just as for the 3 concentric circles example, this 1D representation is discriminative, also with intra-cluster variability in order to reconstruct properly. The 2D re-representation on Figure 7(d) shows again the disentangling properties of the KAE.

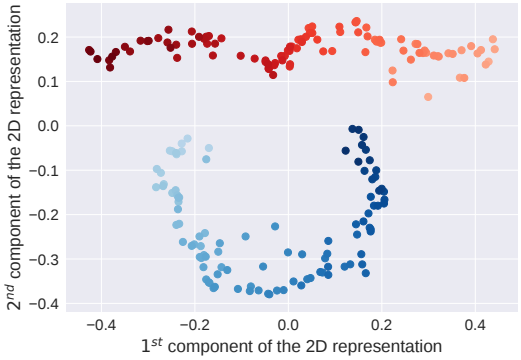


(a) Two moons dataset, colored w.r.t. its 1D representation



(b) 1D representation of the 2 moons

(c) Two moons dataset, colored w.r.t. its 2D representation



(d) 2D representation of the 2 moons

Figure 7: Algorithm behavior on the 2 moons dataset

C.2 NCI Data

C.2.1 All Strategies on 8 Cancers Graph

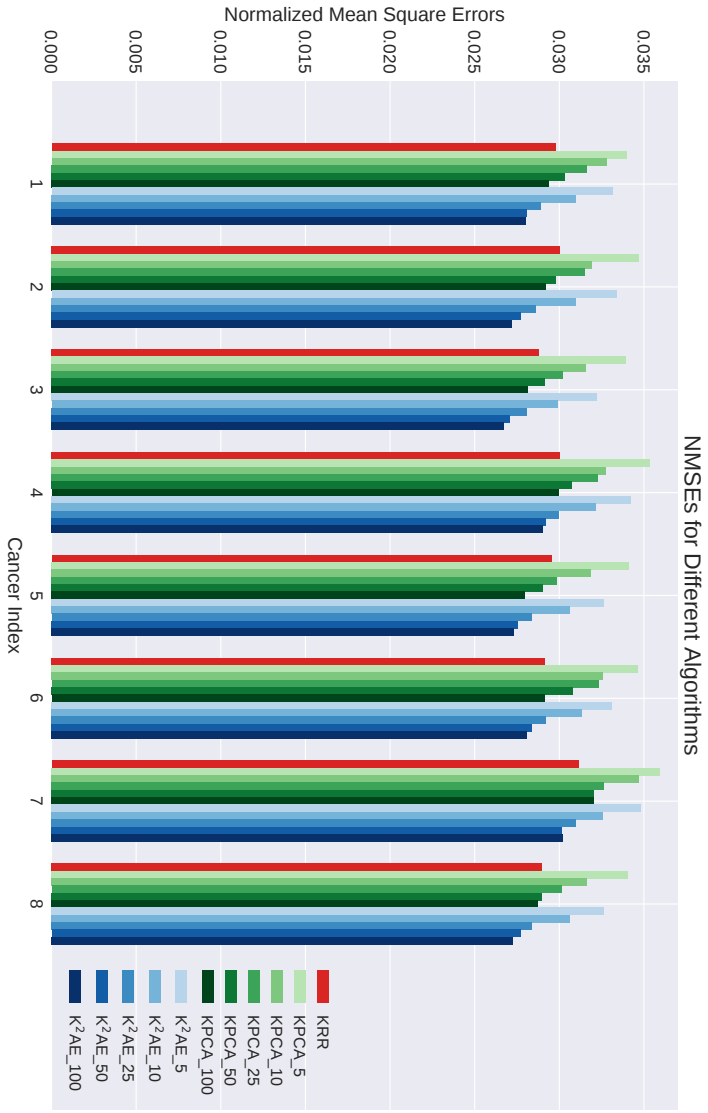


Figure 8: Performance of the Different Strategies on 8 Cancers

As expected, the greater the dimension of the extracted representations, the better the prediction performance by the RF, for both K^2AE and KPCA. However, it is worth noticing that for cancer 7, the prediction error increases between the 50 and the 100-long representations. This might be the beginning of an overfitting phenomenon (seen on 8 of the 59 cancer types, always between the 50 and the 100-dimensional representations), as the extracted components may become less relevant, thus misleading the RF in its predictions.

C.3 5 strategies on 59 cancers table

Table 3: NMSEs on Molecular Activity for Different Types of Cancer

	KRR	KPCA 10 + RF	KPCA 50 + RF	K^2AE 10 + RF	K^2AE 50 + RF
Cancer 01	0.02978	0.03279	0.03035	0.03097	**0.02808**
Cancer 02	0.03004	0.03194	0.02978	0.03099	**0.02775**
Cancer 03	0.02878	0.03155	0.02914	0.02989	**0.02709**
Cancer 04	0.03003	0.03274	0.03074	0.03218	**0.02924**
Cancer 05	0.02954	0.03185	0.02903	0.03065	**0.02754**
Cancer 06	0.02914	0.03258	0.03083	0.03134	**0.02838**
Cancer 07	0.03113	0.03468	0.03207	0.03257	**0.03018**
Cancer 08	0.02899	0.03162	0.02898	0.03065	**0.02770**
Cancer 09	0.02860	0.02992	0.02804	0.02872	**0.02627**
Cancer 10	0.02987	0.03291	0.03111	0.03170	**0.02910**
Cancer 11	0.03035	0.03258	0.03095	0.03188	**0.02900**
Cancer 12	0.03178	0.03461	0.03153	0.03253	**0.02983**
Cancer 13	0.03069	0.03338	0.03104	0.03162	**0.02857**
Cancer 14	0.03046	0.03340	0.03102	0.03135	**0.02862**
Cancer 15	0.02910	0.03221	0.03066	0.03131	**0.02806**
Cancer 16	0.02956	0.03220	0.02958	0.03060	**0.02779**
Cancer 17	0.03004	0.03413	0.03140	0.03145	**0.02869**
Cancer 18	0.02954	0.03195	0.03005	0.03108	**0.02805**
Cancer 19	0.03003	0.03211	0.03079	0.03178	**0.02832**
Cancer 20	0.02911	0.03179	0.03041	0.03085	**0.02769**
Cancer 21	0.02963	0.03275	0.03023	0.03152	**0.02837**
Cancer 22	0.03075	0.03391	0.03089	0.03263	**0.02958**
Cancer 23	0.03006	0.03286	0.02983	0.03109	**0.02760**
Cancer 24	0.03075	0.03398	0.03112	0.03242	**0.02894**
Cancer 25	0.02977	0.03307	0.03054	0.03159	**0.02824**
Cancer 26	0.03083	0.03358	0.03132	0.03206	**0.02959**
Cancer 27	0.03083	0.03347	0.03116	0.03230	**0.02974**
Cancer 28	0.03061	0.03256	0.03116	0.03185	**0.02918**
Cancer 29	0.03056	0.03360	0.03147	0.03181	**0.02892**
Cancer 30	0.03099	0.03288	0.03100	0.03181	**0.02906**
Cancer 31	0.03082	0.03361	0.03161	0.03242	**0.02986**
Cancer 32	0.03233	0.03562	0.03300	0.03422	**0.03158**
Cancer 33	0.03065	0.03208	0.03045	0.03142	**0.02909**
Cancer 34	0.03326	0.03668	0.03423	0.03486	**0.03183**
Cancer 35	0.03292	0.03587	0.03393	0.03450	**0.03146**
Cancer 36	0.03068	0.03389	0.03122	0.03249	**0.02925**
Cancer 37	0.03023	0.03310	0.03061	0.03130	**0.02878**
Cancer 38	0.03100	0.03487	0.03156	0.03327	**0.02974**
Cancer 39	0.02989	0.03288	0.03149	0.03148	**0.02865**
Cancer 40	0.03166	0.03525	0.03201	0.03352	**0.03010**
Cancer 41	0.03139	0.03501	0.03203	0.03316	**0.03025**
Cancer 42	0.03010	0.03251	0.03013	0.03072	**0.02807**
Cancer 43	0.03042	0.03324	0.03062	0.03144	**0.02806**
Cancer 44	0.02838	0.03045	0.02821	0.02927	**0.02679**
Cancer 45	0.02910	0.03085	0.02895	0.02970	**0.02651**
Cancer 46	0.02969	0.03258	0.02996	0.03111	**0.02834**
Cancer 47	0.03148	0.03438	0.03346	0.03286	**0.03056**
Cancer 48	0.03272	0.03640	0.03397	0.03425	**0.03197**
Cancer 49	0.03305	0.03392	0.03329	0.03334	**0.03148**
Cancer 50	0.03229	0.03637	0.03300	0.03404	**0.03155**
Cancer 51	0.02943	0.03188	0.03028	0.03072	**0.02857**
Cancer 52	0.03309	0.03420	0.03252	0.03335	**0.03130**
Cancer 53	0.03170	0.03340	0.03105	0.03170	**0.02843**
Cancer 54	0.03189	0.03439	0.03164	0.03345	**0.03036**
Cancer 55	0.03082	0.03339	0.03146	0.03207	**0.02892**
Cancer 56	0.03026	0.03327	0.03041	0.03185	**0.02901**
Cancer 57	0.02962	0.03237	0.02990	0.03162	**0.02855**
Cancer 58	0.02883	0.03200	0.02978	0.03058	**0.02783**
Cancer 59	0.02936	0.03208	0.02914	0.03032	**0.02750**