# 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 $\lambda_l$'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 $\ell$ 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 $\ell$ 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 $1 - \delta$:

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

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:

$$\begin{aligned}
\widehat{\mathscr{R}}_n\Big(\big(\ell \circ (\mathrm{id} - \mathcal{H}_{s,t})\big)(S)\Big) &\leq 2\sqrt{2}M\ \widehat{\mathscr{R}}_n\Big(\big(\mathrm{id} - \mathcal{H}_{s,t}\big)(S)\Big), \qquad (14)\\
&\leq 2\sqrt{2}M\ \left[\widehat{\mathscr{R}}_n\Big(\{\mathrm{id}\}(S)\Big) + \widehat{\mathscr{R}}_n\Big(\mathcal{H}_{s,t}(S)\Big)\right],\\
&\leq 2\sqrt{2}M\ \widehat{\mathscr{R}}_n\Big(\mathcal{H}_{s,t}(S)\Big),\\
\widehat{\mathscr{R}}_n\Big(\big(\ell \circ (\mathrm{id} - \mathcal{H}_{s,t})\big)(S)\Big) &\leq 2\sqrt{\pi}M\ \widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S)\Big), \qquad (15)
\end{aligned}$$

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 $\gamma$ 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:*

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

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

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 $S = (x_1, \dots, x_n)$ and return $(f(x_1), \dots, f(x_n))$, $f \in \mathcal{H}_{1,s}$. 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 $(y_1, \dots, y_n)$ and return $(g(y_1), \dots, g(y_n))$, $g \in \mathcal{H}_{2,t}$. Finally, let $y_0 = (0_{\mathbb{R}^p}, \dots, 0_{\mathbb{R}^p})$ (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

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

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: $\forall (y, y') \in \mathbb{R}^p$,*

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

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 $\phi_m$ 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 $\sigma_m$. Indeed, we would have then: $\forall z \in \mathcal{X}_0$,

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

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

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

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:

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

**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 $\kappa_m$ (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:

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

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:

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

**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:* $\forall (y, y') \in \mathbb{R}^p$,

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

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 $\phi_m$ are $L_m$-Lipschitz and the $A_m$ operators are trace class. Then, we have: $\forall (y, y') \in \mathbb{R}^p$,

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

Note also that Assumption 13 is stronger than Assumption 11, since $\|A\|_* \le \mathbf{Tr}(A)$ 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:

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

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:

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

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 $C_0 := 8\sqrt{\pi}(C_1 + 2C_2)$, 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 $f(x_i) = y_i \quad \forall\, i \in [\![n]\!]$, 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:

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

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 $\left(\varphi_{l_0,1}^*, \ldots, \varphi_{l_0,n}^*\right) \in \mathcal{X}_{l_0}^n$, 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 $\varphi, \varphi'$. 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 $\varphi$ and $\psi$. 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 $(\varphi,\psi)$. 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 :

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

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 $(\varphi,\psi)$ 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 $\varphi$ and $\psi$ 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).

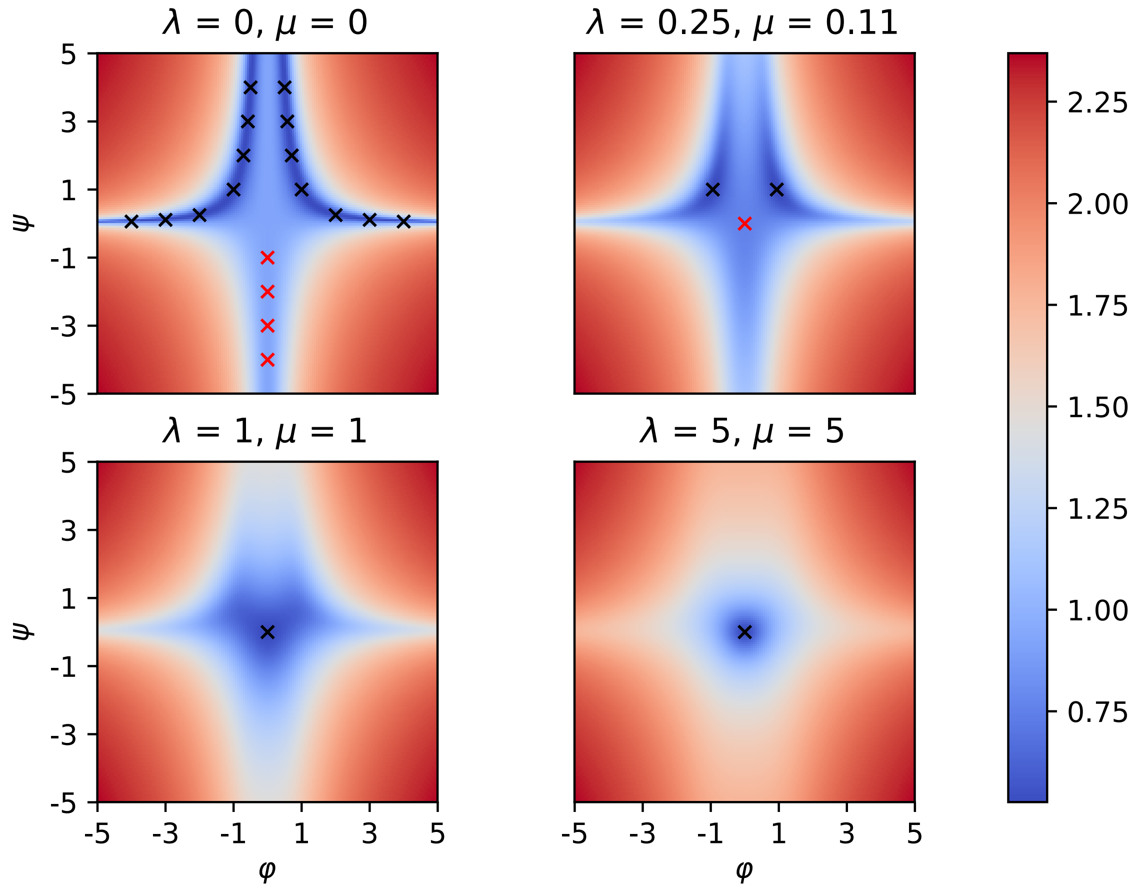


Figure 4: Heatmaps of $\mathcal{P}$ for different values of $\lambda$ and $\mu$

# 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)

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

where $\nabla^{(1)} k_l\left(x, x'\right)$ (respectively $\nabla^{(2)} k_l\left(x, x'\right)$) 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 $\varphi_{l_0,i_0}$ 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 $(l > l_0)$ :

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

with $\Delta_l(x) \coloneqq \left(\left(\nabla^{(1)} k_l\left(x, x_1{}^{(l-1)}\right)\right)^T, \ldots, \left(\nabla^{(1)} k_l\left(x, x_n{}^{(l-1)}\right)\right)^T\right)^T$ 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 $\partial$ 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 $\varphi_{l_0,i_0}$, but that $f_l$ does, through the influence of $\varphi_{l_0,i_0}$ on the supports of $f_l$, namely the $x_{i'}{}^{(l-1)}$.

### B.4 Detail of the $\Delta_l$ 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 $\Delta_l$ 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 $\Delta_l$ derivation for two popular kernels : the Gaussian and the polynomial ones.

Gaussian kernel :

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

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

where :

- $X^{(l-1)} := \left( \left({x_1}^{(l-1)}\right)^T, \dots, \left({x_n}^{(l-1)}\right)^T \right)^T \in \mathbb{R}^{n \times d_{l-1}}$ stores the level $l-1$ representations of the $x_i$'s in rows
- $\tilde{X}_i^{(l-1)} := \left( \left({x_i}^{(l-1)}\right)^T, \dots, \left({x_i}^{(l-1)}\right)^T \right)^T \in \mathbb{R}^{n \times d_{l-1}}$ 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 $\Delta_l$ 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 :

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

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

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 $\Delta_l$ 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

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

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 $\Phi_1, \ldots, \Phi_{L-1}$, 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 $\lambda_L$. Let $N_{\text{KRR}}$ be the function that computes $N_L$ of equation (34) from $\Phi_1, \ldots, \Phi_{L-1}$, $K_{in}$ and $\lambda_L$.

### 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 $\|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0}$ 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:

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

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 $\Phi_l$, 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).

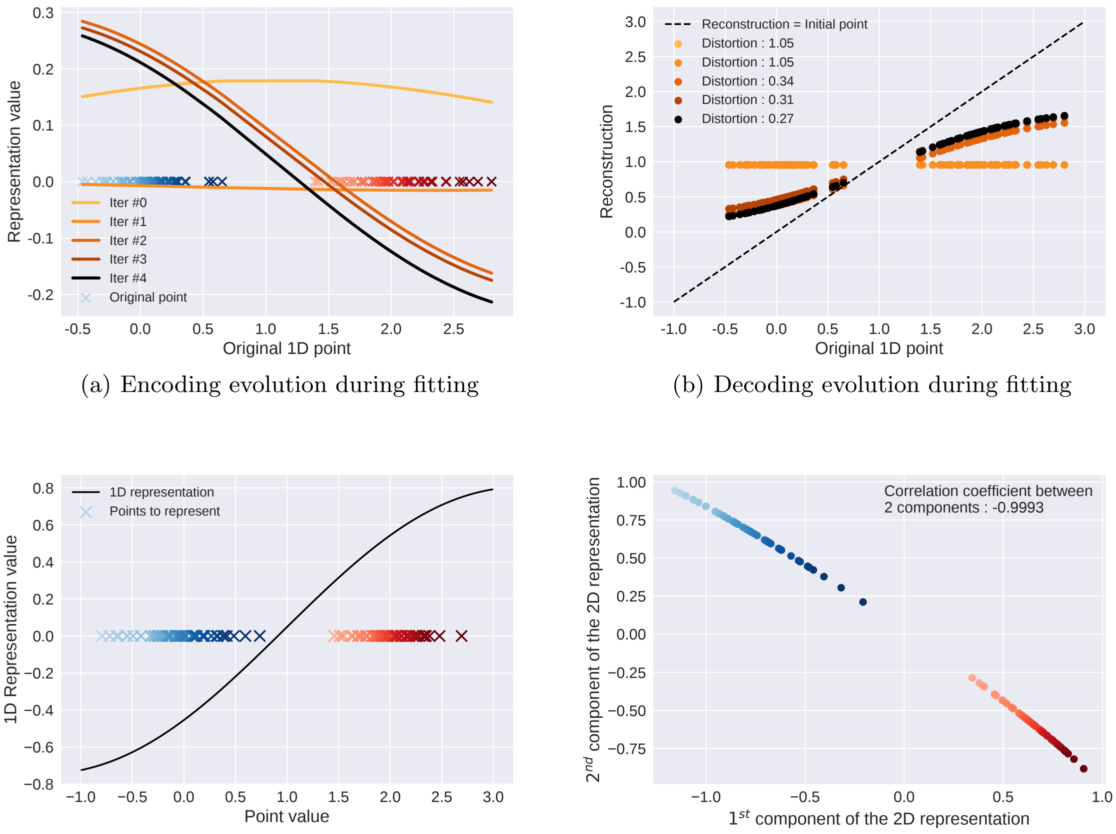
# C Additional Experiments

## 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

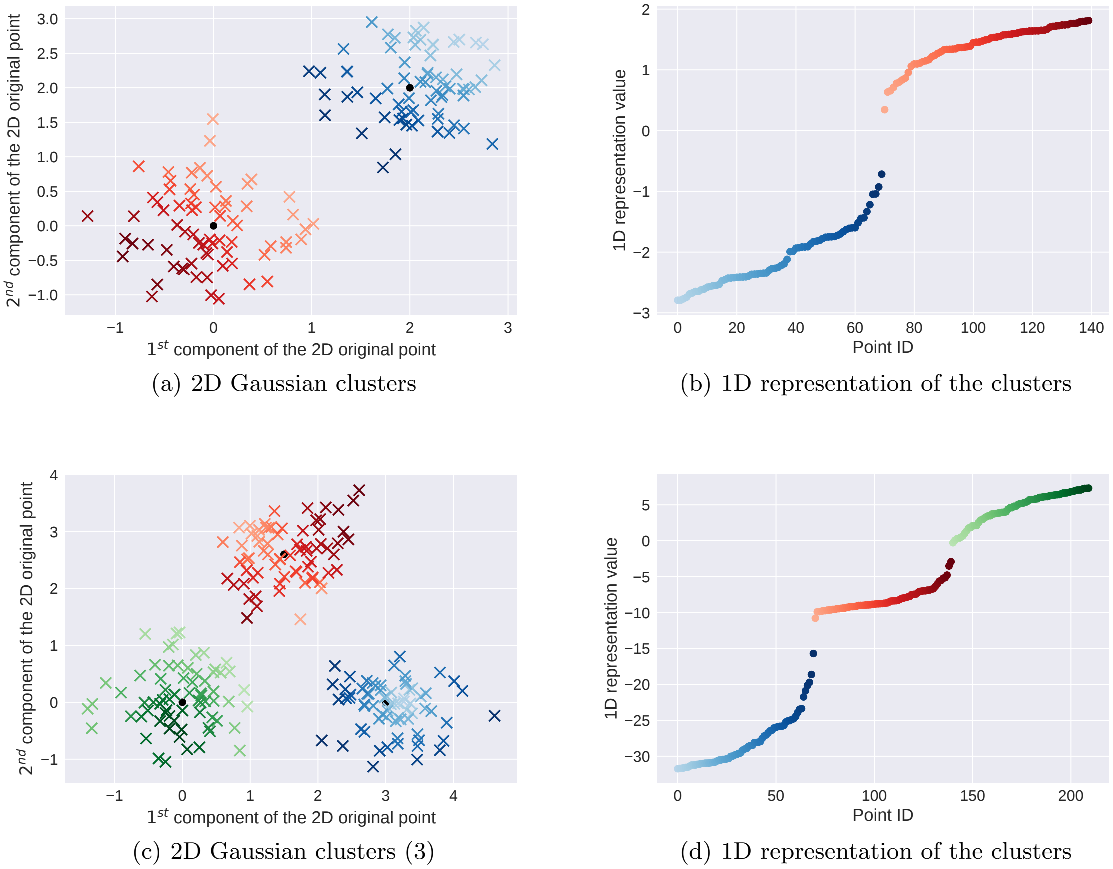
(b) Decoding evolution during fitting

(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

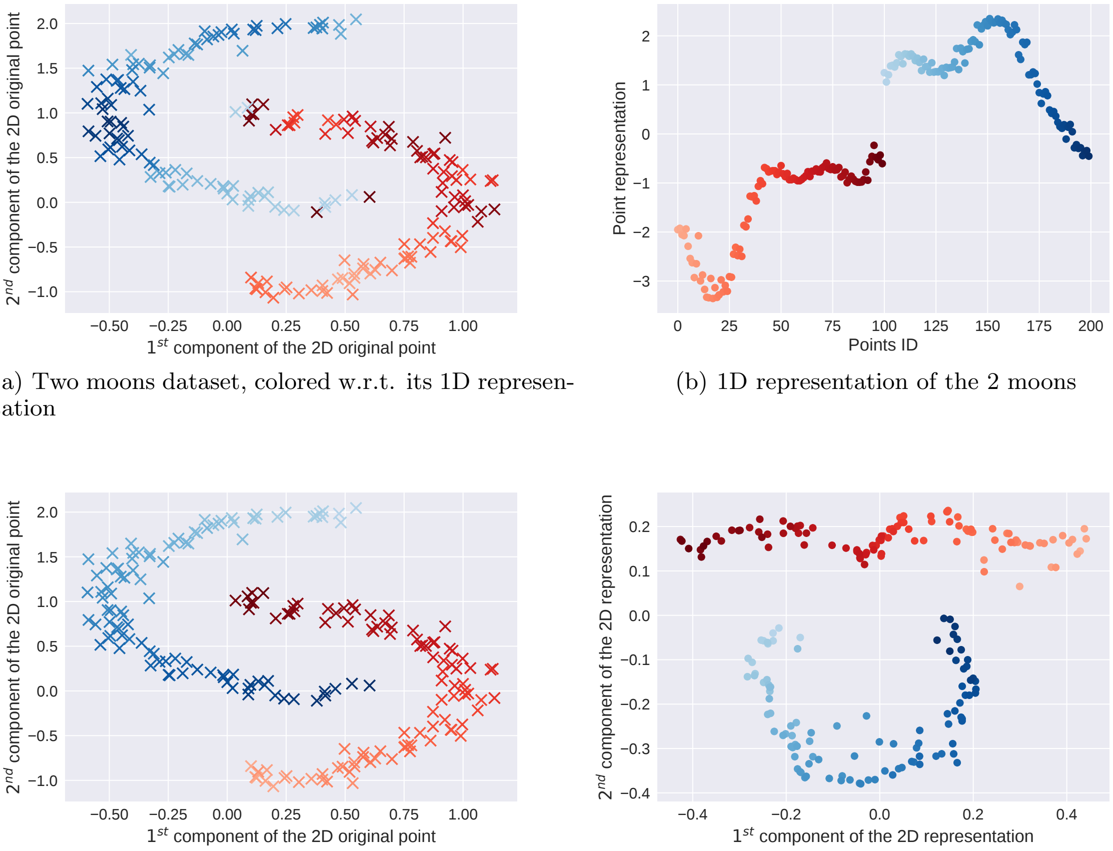
(b) 1D representation of the clusters

(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

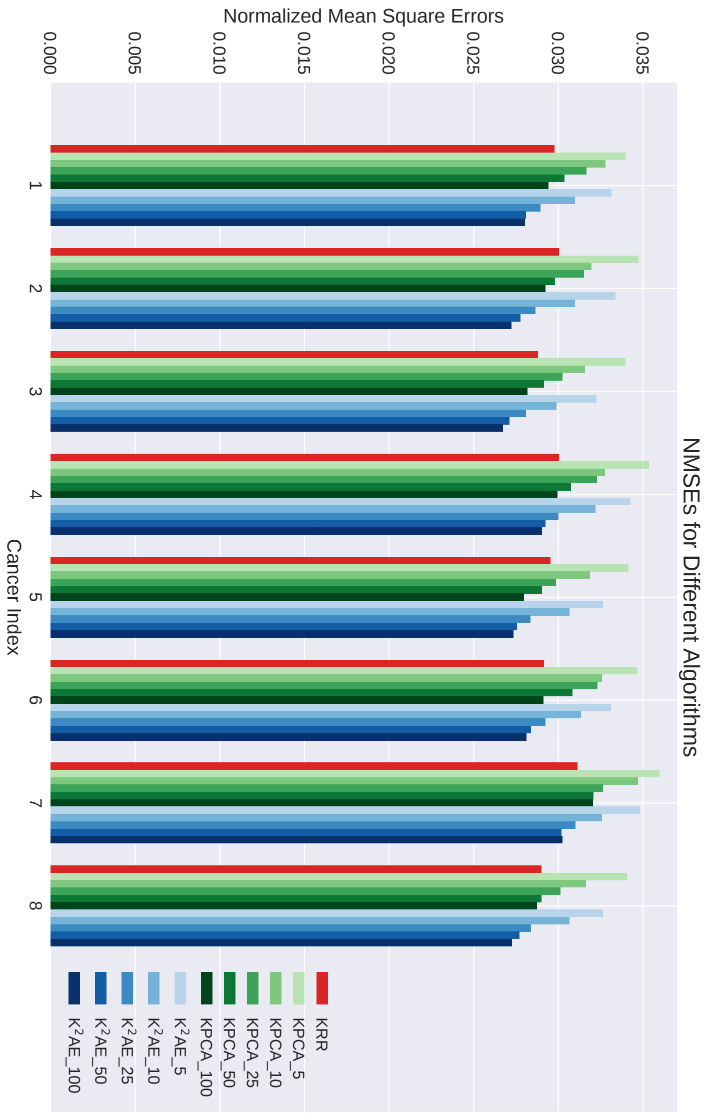


Figure 8: Performance of the Different Strategies on 8 Cancers

As expected, the greater the dimension of the extracted representations, the better the prediction performance by the RF, for both K$^2$AE 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 | $K^2$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** |
| 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** |