Measuring dissimilarity with diffeomorphism invariance

Théophile Cantelobre [1 2] **Carlo Ciliberto** [3] **Benjamin Guedj** [2 3 4] **Alessandro Rudi** [1]

Abstract

Measures of similarity (or dissimilarity) are a key ingredient to many machine learning algorithms. We introduce DID, a pairwise dissimilarity measure applicable to a wide range of data spaces, which leverages the data's internal structure to be invariant to diffeomorphisms. We prove that DID enjoys properties which make it relevant for theoretical study and practical use. By representing each datum as a function, DID is defined as the solution to an optimization problem in a Reproducing Kernel Hilbert Space and can be expressed in closed-form. In practice, it can be efficiently approximated via Nyström sampling. Empirical experiments support the merits of DID.

1. Introduction

One of the overarching goals of most machine learning algorithms is to generalize to unseen data. Ensuring and quantifying generalization is of course challenging, especially in the high-dimensional setting. One way of reducing the hardness of a learning problem is to study the invariances that may exist with respect to the distribution of data, effectively reducing its dimension. Handling invariances in data has attracted considerable attention over time in machine learning and applied mathematics more broadly. Two notable examples are image registration (De Castro and Morandi, 1987; Reddy and Chatterji, 1996) and time series alignment (Sakoe and Chiba, 1978; Cuturi and Blondel, 2017; Vayer et al., 2020; Blondel et al., 2021; Cohen et al., 2021).

In practice, data augmentation is a central tool in the machine learning practitioner's toolbox. In computer vision for instance, images are randomly cropped, color spaces are changed, and artifacts are added. Such heuristics enforce some form of invariance to transformations that are chosen by hand, often parametrically, and comes at the cost of a more demanding learning step (*e.g.*, more data to store and process, more epochs, to name but a few).

Prior work. Learning under invariances has also spurred significant theoretical interest. DeCoste and Schölkopf (2002), Haasdonk and Burkhardt (2007), Kondor (2008) and Mroueh et al. (2015) algebraically studied how kernels invariant to group actions behave for learning. Bruna and Mallat (2013) takes inspiration in signal processing (with wavelet spaces) to build scattering networks, which can present good properties with respect to invariances. Focusing on neural networks, Mairal et al. (2014) and Bietti and Mairal (2019) introduced and analyzed a model aiming to mimic Convolutional Neural Networks (CNNs). More recently, Bietti et al. (2021) studied the sample complexity of learning in the presence of invariances, with invariant kernels. Conversely, the hypothesis that invariance can be a proxy for neural network performance has been put to test empirically by Petrini et al. (2021).

Contributions. We introduce a dissimilarity called DID (standing for Diffeomorphism Invariant Dissimilarity) which is invariant to smooth diffeomorphisms present between data points. Although DID is somewhat less sophisticated than most of the models presented above, it is considerably more generic and can be seen as a building block for devising practical machine learning algorithms in the presence of invariances.

In order to exploit the internal structure of a data point (*e.g.* of an image), we cast them as functions between two spaces (*e.g.* coordinate and color spaces). This unlocks the potential of using the change of variable formula to eliminate diffeomorphisms between functions (*i.e.* transformations between data points) when they exist. DID is based on a generic method for identifying if a function g is of the form $f \circ Q$, without any parametric model over Q, founded on the change of variable formula. This makes DID a promising and flexible tool for image processing (*e.g.*, image registration), time-series analysis (*e.g.*, dynamic time warping) and

[1]DI ENS, Ecole normale supérieure, Université PSL, CNRS, Inria, Paris, France [2]Inria London, UK [3]Centre for AI, Department of Computer Science, University College London, UK [4]Inria, Lille - Nord Europe Research Centre, Lille, France. Correspondence to: Théophile Cantelobre <theophile.cantelobre@inria.fr>, Carlo Ciliberto <c.ciliberto@ucl.ac.uk>, Benjamin Guedj <benjamin.guedj@inria.fr>, Alessandro Rudi <alessandro.rudi@inria.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

machine learning (*e.g.*, nearest neighbors).

DID is defined as an optimisation problem in a Reproducing Kernel Hilbert Space (RKHS), of which we present a closed form solution (Thm. 4.2). We then show how it can be approximated in practice, using Nyström sampling techniques (Lemma 4.3, Thm. 4.4). By relying on standard matrix linear algebra, this approximation can be efficiently implemented with batch techniques, and accelerated hardware.

A key aspect is that DID has very few "hyper-parameters" which can easily be chosen by a domain expert: the kernels on the input and output space and a regularization parameter. We provide guidance on choosing the regularization parameter in Sec. 4.2.

Using tools from functional analysis, we prove that DID behaves as expected when comparing f and $f \circ Q$ (*i.e.*, that it considers these two functions to be close) in the limit of vanishing regularization (Thm. 3.2). We support our theoretical claims with numerical experiments on images.

Outline. We introduce our new dissimilarity in Sec. 2. In Sec. 3, we prove that the intuition leading to the definition in Sec. 2 is well founded and discuss the theoretical properties of the dissimilarity. We then show how to compute the dissimilarity in Sec. 4: we first show that it has a closed-form expression; we then present and justify an approximation scheme based on Nyström sampling. We illustrate the behavior of the dissimilarity with experiments on images in Sec. 5.

2. The dissimilarity

2.1. Informal derivation of the dissimilarity

The dissimilarity we describe in this work relies on the internal structure of the objects it compares. An efficient way of encoding this structure is to, whenever possible, view objects as maps between an input space and an output space. In this way, we can consider both the values taken by the function and the locations at which these values were taken.

Consider f and g two maps between $\mathbb{R}^d$ and $\mathbb{R}^p$. Our goal is to determine whether there exists a diffeomorphism $Q : \mathbb{R}^d \to \mathbb{R}^d$ such that $g = f \circ Q$. In practice, such transformations could be rigid-body transformations, a non-singular projective transformation or more generally a mild distorsion of the space (such as warping). The goal is thus to find a measure of dissimilarity between f and g that is robust to such diffeomorphisms. We derive such a measure informally in this section around three key ideas.

Change of variable formula. Integrals offer a natural way to "eliminate" a diffeomorphism from a function, via the change of variable formula

$$\int f(Q(x))|\nabla Q(x)|\mathrm{d}x = \int f(x)\mathrm{d}x.$$

As $|\nabla Q|$ (the determinant of the Jacobian of Q) is unknown, we can approximate the above formula with:

$$\min_q \left| \int g(x)q(x)\mathrm{d}x - \int f(x)\mathrm{d}x \right|, \tag{1}$$

where q lies in a space of functions.

Indeed, when $g = f \circ Q$ for some diffeomorphism Q, choosing $q = |\nabla Q|$ minimizes (1). However, this solution is not unique. Indeed there exist trivial solutions as $q = f/g$, that are irrespective of the existence of a Q such that $g = f \circ Q$ or not.

Range of statistics. One way of reducing the class of solutions to ones that are relevant to our original question is to study not only how well the weighted integral of g can approximate the integral of f, but also require that the same weight approximate a wide class of transformations of f. A natural example with inspiration in probability theory is to be able to approximate all moments of f, *i.e.*, the integrals of the moments $v_1(f) = f, v_2(f) = f^2, \ldots$, or more general statistics. The function $q = f/g$ may match the integral of $v_1(g)$ with that of $v_1(f)$, but cannot work for v_2. However, if $g = f \circ Q$, $q = |\nabla Q(x)|$ (the solution we are seeking), satisfies that for any continuous function $v : \mathbb{R}^p \to \mathbb{R}$,

$$\int v(f(Q(x)))|\nabla Q(x)|\mathrm{d}x = \int v(f(x))\mathrm{d}x.$$

Problem (1) is thus replaced by the following one, which also has q as a solution when $g = f \circ Q$:

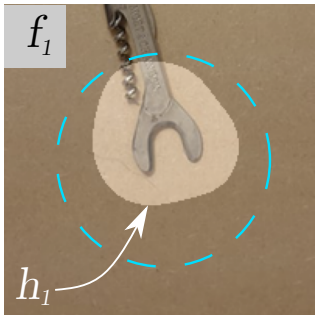
$$\min_q \max_{v \in V} \left| \int v(g(x))q(x)\mathrm{d}x - \int v(f(x))\mathrm{d}x \right|, \tag{2}$$

where V is a rich set of statistics, *e.g.*, continuous integrable functions on $\mathbb{R}^p$.

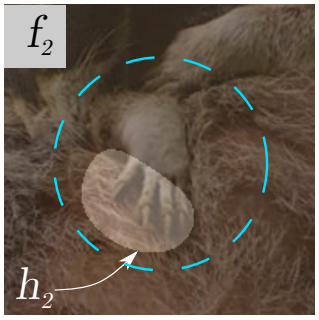
Uniformity over regions. Problem (2) averages the statistics uniformly over the whole space irrespectively of the fact that a witness of $g \neq f \circ Q$ could live in a lower dimensional region. For instance g and f might be non-zero only on a small region of the space, consequently yielding to a relatively small value for Eq. (2). To enhance such regions, we choose to integrate with respect to a smooth function h, that is chosen adversarially to maximize the dissimilarity between f and g. In other words, we arrive at the following optimization problem:

$$\max_{h \in \mathcal{H}_1} \min_q \max_{v \in V} \left| \int v(g(x))q(x)\mathrm{d}x - \int v(f(x))h(x)\mathrm{d}x \right|,$$

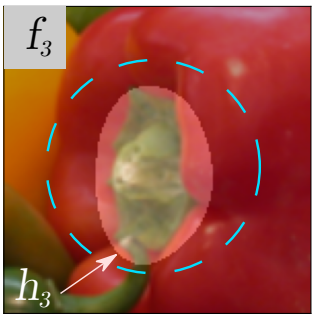
where $\mathcal{H}_1$ is a suitable set of smooth functions. Again, if $g = f \circ Q$, the above is solved by $q = h \circ Q\,|\nabla Q|$.



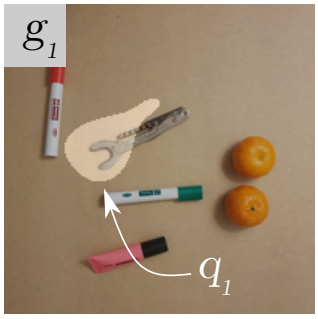
(a) f_1 (close-up)



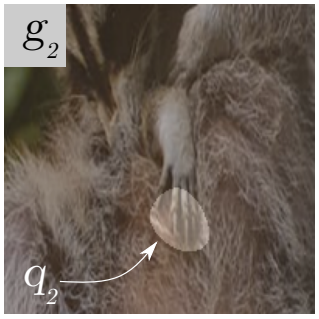
(b) f_2 (raccoon's paw)



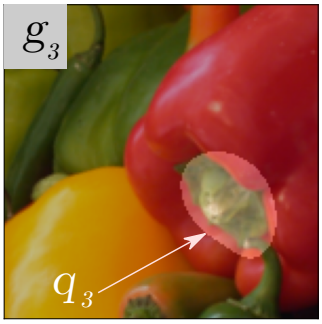
(c) f_3 (pepper close-up)



(d) g_1 (scene with multiple objects)



(e) g_2 (raccoon's abdomen, with natural keystone deformation.)



(f) g_3 (peppers among vegetables)

Figure 1. Illustration of $\widehat{D}_\lambda$ on images. We compute $\widehat{D}_\lambda(f_i, g_i)$ and materialize the optimal h_i and q_i selected by DID (with thresholding for visualisation purposes), for f_i and g_i images taken with a smartphone. The mask μ is illustrated by the dashed circle (in light blue), see Sec. 5.1 for details. The images are taken from different views so as to provide different angles and lighting. Figs. 1a and 1d show a scene of `objects` and a close-up on one of them (bottle opener). Figs. 1b and 1e are taken from images of a raccoon (`raccoon`). Figs. 1c and 1f are sub-patches from `peppers`, a scene with vegetables. Notice that q_i visually matches the area highlighted by h_i, despite the perspective and scale changes. Additional details are gathered in Sec. 5.2.

Note that the smoothness of h and q is crucial. On the one hand, the smoothness of h ensures that the considered regions are of interest with respect to the underlying metric on $\mathbb{R}^d$, i.e. they cannot be too close to diracs on pathological sets. On the other, the smoothness of q ensures that the transformations are not matched by "cherry-picking" dispersed points on the domain such that the integrals match.

2.2. Definition of the dissimilarity

Now that we have given the motivation as well as the intuition behind our method, we can formally introduce it.

Let $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^p$ where $d, p \geq 1$. In this paper, the objects of interest are maps from X to Y. Note that this is quite flexible and can reflect many of the rich types of data considered in image processing, time-series modelling and machine learning. Indeed, an image can be seen as a map from $\mathbb{R}^2$ (the coordinates of the pixels) to $\mathbb{R}^3$ (the color space, RGB for instance). A time-series can be seen as a map from $\mathbb{R}^1$ (time) to $\mathbb{R}^n$ (the space of the sample values).

Let k_X be a reproducing kernel on X with Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ (Aronszajn, 1950) and k_Y be a reproducing kernel on Y with RKHS $\mathcal{F}$. In particular, we assume that they are bounded.

To conclude, the formalization below allows naturally for the presence of a mask function μ, with the role of focusing the matching process on a subregion of interest of f. The role of the mask will be discussed after the definition.

Definition 2.1 (Dissimilarity $D_\lambda(f, g)$)**.** Let $\lambda \geq 0$ and bounded integrable $\mu : \mathcal{X} \to \mathbb{R}$. For any $f, g : X \to Y$, we define the dissimilarity

$$D_\lambda(f, g) := \max_{\|h\|_{\mathcal{H}} \leq 1} \min_{q \in \mathcal{H}} \Delta_{f,g}(h, q) + \lambda \|q\|^2_{\mathcal{H}} \quad (3)$$

where $\Delta_{f,g}(h, q)$ is defined as follows,

$$\Delta_{f,g}(h, q) := \max_{\|v\|_{\mathcal{F}} \leq 1} \left| \int_X v(g(x))q(x)\mathrm{d}x - \int_X v(f(x))\mu(x)h(x)\mathrm{d}x \right|^2.$$

When clear from context, we write D instead of D_λ for the sake of conciseness.

The reader should easily recognize the construction described in Sec. 2.1. We have made the space in which we search for q and h precise: $\mathcal{H}$, the RHKS of k_X. Regularity of h is enforced by searching in the unit ball, while regularity of q is enforced with Tikhonov regularization. This choice allows at the same time to compute the dissimilarity in closed form (see Sec. 4), while not sacrificing its expressivity (see Sec. 4.2). In particular, to allow a very rich set of statistics that is also manageable from a computational viewpoint, we choose V to be the unit ball of $\mathcal{F}$. For example, if Y is a bounded subset of $\mathbb{R}^p$ and k_Y is chosen as the Laplace kernel $k_Y(y, y') = \exp(-\|y - y'\|)$, then a rescaled version of any infinitely smooth function belongs to V (including in particular all polynomials, smooth probabilities, Fourier basis – see Appendix A for more details). For the same reason we choose $\mathcal{H}_1$ to be the unit ball of $\mathcal{H}$ and we choose also $q \in \mathcal{H}$. Fig. 1 shows a few examples of the role played by the h and q optimizing the problem in Eq. (3).

The role of the mask μ. We introduced a function $\mu : X \to \mathbb{R}$ which applies to the term depending on f and h. This function is meant to be a mask which focuses the distance on a certain region of f, discounting other regions. Such presence is useful in practice, since typically the space X is given by the problem. For example, if we want to use the dissimilarity to check if the content of a given image (f) is contained in an image (g), the shape X is typically rectangular, while the region of interest is the interior of the image. In this case the mask is useful to avoid the artifacts introduced by the corners. Notice that this addition further breaks the symmetry between f and g: f becomes a *reference*, and we search in g for matching statistics. In the experiments, we relied for example on a Blackman Window, a classical window function in signal and image processing (see Sec. 5.1) to reduce the impact of the corners.

3. Robustness to diffeomorphisms

The dissimilarity D is designed (consistently with the derivation in Sec. 2.1) to be small when f and g are equal up to a diffeomorphism. The ideal result would be something along the lines of the following *informal* theorem:

Theorem 3.1 (Ideal)**.** *For any $f : X \to Y$ and any Q diffeomorphism over X, $D(f, f \circ Q) \approx 0$.*

This is of course too much to ask. Indeed, the regularization over the choice of q (which in turn controls the regularity of the jacobian of a hypothetical Q) introduces a bias: even if $g = f \circ Q$, the dissimilarity is not 0. This bias vanishes if and only if $q^* = 0$ (by definiteness of the norm).

However, when the RKHS is assumed to be rich enough and Q and μ are regular enough, then we have the following result. Before stating it, we recall that the Laplace kernel $k_X(x, x') = \exp(-\|x - x'\|)$ belongs to the more general family of Sobolev kernels (Wendland, 2004). In particular, it corresponds to the Sobolev kernel of smoothness m, with $m = (d + 1)/2$, where d is the dimension of $X \subset \mathbb{R}^d$.

Theorem 3.2. *Let $X \subset \mathbb{R}^d$ be an open bounded set with Lipschitz boundary. Let $\mu \in C^\infty(\mathbb{R}^d)$ with compact support $\Omega \subset X$. Choose k_X to be a Sobolev kernel of smoothness m, with $m > d/2$. Then for any C^{m+1} diffeomorphism Q on $\mathbb{R}^d$ satisfying $Q^{-1}(\Omega) \subset X$, we have that*

$$D_\lambda(f, f \circ Q) \;\leq\; \lambda\, C_\mu^2 C_Q^2 \qquad \forall f \text{ measurable},$$

where $C_\mu = \|\mu\|_{H^m(\mathbb{R}^d)}$, and C_Q is defined in Eq. (19) in Appendix B.2 and depends only on Ω, X, Q, d, m.

The theorem above is a special case of Thm. B.3, presented in Appendix B.2. There we prove the more general result: $D_\lambda(f, g) \leq \lambda C_\mu^2 C_Q^2$, for all measurable f, g that satisfy $g(x) = (f \circ Q)(x)$, only in the region not canceled by the mask, *i.e.*, $\forall x \in Q^{-1}(\Omega)$.

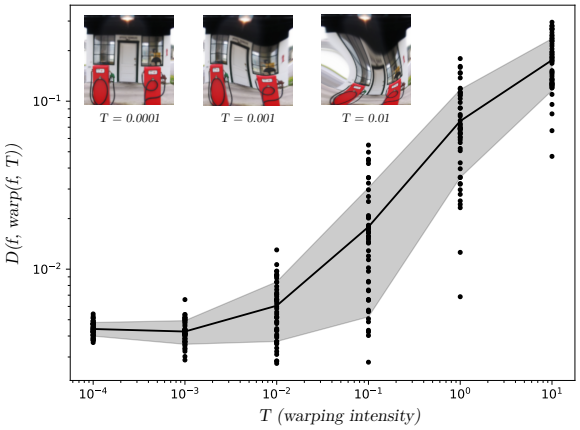
A first consequence of Thm. 3.2 is that the regularization parameter λ controls the threshold to decide whenever $g \approx f \circ Q$. In particular, we easily see that $D_\lambda(f, f \circ Q) \to 0$, when $\lambda \to 0$. This result confirms that in the limit where the regularization vanishes, so does the bias and we have the result we would have imagined à la Thm. 3.1. We will see in Sec. 4.2 that λ has also an important role in controlling the approximation error of $D(f, g)$. This shows that λ controls a similar bias-variance trade-off as in classical kernel supervised learning (Shawe-Taylor and Cristianini, 2004).

To show concretely the dependence of C_Q with respect to Q, in the following example we show C_Q explicitly for an interesting class of diffeomorphisms.

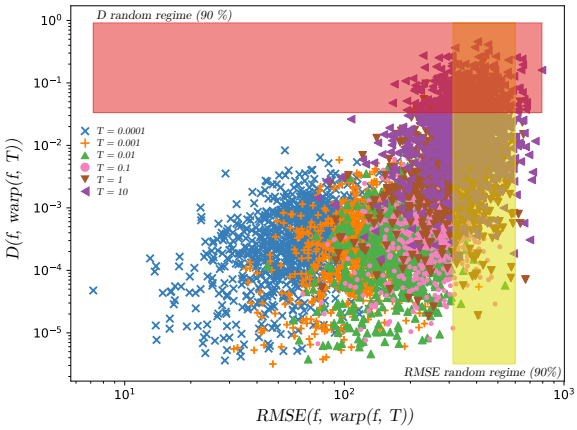
Example 3.3 (Magnitude of C_Q for rigid transforms). Let X be the unit ball in $\mathbb{R}^d$ and let μ be a mask supported on Ω, the ball of radius $r < 1$. We consider the diffeomorphisms $Q(x) = \alpha R x$, with R a unitary matrix and $r < \alpha < 1/r$. We use the Laplace kernel $k_X(x, x') = \exp(-\|x - x'\|)$ for k_X (analogously for k_Y). Then, we compute explicitly the bound in Eq. (19), since the Laplace kernel corresponds to Sobolev kernel with exponent $m = (d + 1)/2$, obtaining

$$C_Q \;\leq\; C_0 \left(\frac{\alpha}{\min(\alpha, 1) - r} \right)^{m + d/2} \alpha^d (1 + \alpha + \alpha^m).$$

Remark 3.4. By seeing data points as functions, we are able to efficiently reason about them. Hence, since data points are continuous objects, $d = \dim(X)$, the dimensions of X, and not the number of pixels, which we denote N below. Indeed, X is low-dimensional by nature : for images, $d = 2$ since images are 2-d objects, while N could be very large. This way, the dimensional dependence of C_Q is very reasonable. In the extreme case where we consider time series of point



(a) $\widehat{D}_\lambda(f, warp(f,T))$ *as a function of* T. We repeat the warp 50 times ($+$ markers) on the same image and represent average values $\pm$ standard deviation (in grey).



(b) *RMSE*$(f, warp(f,T))$ *against* $\widehat{D}_\lambda(f, warp(f,T))$ *for various values of* T *and images* f. 1000 images are each warped once for each value of T. We represent the random regime for each metric.

Figure 2. Invariance to general diffeomorphisms (warping). Warping is randomly generated, its intensity controlled by a temperature parameter T (higher T produces, on average, warps with higher displacement norm). In (a): $\widehat{D}_\lambda$ stays constant (i.e. invariant to the warps) as long as their norm is not too strong (small T), while RMSE increases exponentially. When T becomes large the transformations become intense (indeed they are non-diffeomorphic) and $\widehat{D}_\lambda$ grows to reflect this fact. In (b): we see that $\widehat{D}_\lambda$ stays invariant to warps as long as $T \leq 0.1$ (far from the random regime interval), while the Euclidean distance increases exponentially with T, even for small T. See Sec. 5.3 for more details.

clouds (for example, brain scans through time), $d = 4$ (time $+$ 3 spatial coordinates). If we consider a scale change range of 50%, then C_Q is of the order of $1.5^4 \approx 5$. We emphasize that C_Q is not related to the number of samples (for images, the number of pixels) used to compute DID in practice.

3.1. Discussion on the discriminatory power of D_λ

In Thm. 3.2, we proved that when $g = f \circ Q$ for some diffeomorphism Q, then $D(f,g)$ is small, i.e. that the dissimilarity is essentially invariant to the diffeomorphisms. However, to fully characterize the properties of the proposed dissimilarity it would be interesting to study also its discriminatory power, i.e. the fact that $D(f,g)$ is small *only if* there exists a diffeomorphism Q such that $g = f \circ Q$. Fig. 2 investigates this question from the empirical perspective. The details of these experiments are reported in Sec. 5 (and further explored in Appendix G). They show that, DID is very robust to significant transformations $f \circ Q$ of the original signal f. Additionally we observe that D_λ is very discriminative, in contrast to less diffeomorphism invariant metrics such as the euclidean distance, when comparing $D_\lambda(f, f \circ Q)$ with $D_\lambda(f,g)$ for a random signal g. We care to point out however, that the theoretical analysis of DID 's the discriminative abilities is beyond the scope of this work (whose aim is to introduce the discrepancy and study its invariance properties) and we postpone it to future research.

4. Computing the dissimilarity

Before deriving the closed form solution for D_λ, we need to recall some basic properties of kernels. Reproducing kernels and RKHSs satisfy the so called *reproducing property*, i.e. There exists a map $\psi : X \to \mathcal{H}$ such that, for any $f \in \mathcal{H}$ and $x \in X$, it holds that $f(x) = \langle f, \psi(x) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the scalar product associated to the RKHS $\mathcal{H}$. Moreover $k_X(x,x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}}$, for all $x, x' \in X$. The same holds for k_Y and $\mathcal{F}$. i.e., there exists $\Phi : Y \to \mathcal{F}$ such that $v(y) = \langle v, \Phi(y) \rangle_{\mathcal{F}}$ for all $v \in \mathcal{F}, y \in Y$ and $k_Y(y,y') = \langle \Phi(y), \Phi(y') \rangle$ for all $y, y' \in Y$, where $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is the inner product associated to $\mathcal{F}$. In particular, note that, since we assumed that k_X, k_Y are bounded kernels, then there exist two constants κ_X, κ_Y such that $\sup_{x \in X} \|\psi(x)\|_{\mathcal{H}} \leq \kappa_X$ and analogously $\sup_{y \in Y} \|\Phi(y)\|_{\mathcal{F}} \leq \kappa_Y$.

4.1. Closed form solution

In Definition 2.1, we define $D_\lambda(f,g)$ as an optimization problem in $\mathcal{H} \times \mathcal{H}$. In fact, this optimization problem has a closed-form solution, as the solution of an eigenvalue problem of an operator between $\mathcal{H}$ and $\mathcal{F}$ as derived in Eq. (7). We introduce the relevant objects and prove Thm. 4.2.

Definition 4.1 (Operators F_μ, G)**.** Given $f, g : X \to Y$, the feature map $\Phi : Y \to \mathcal{F}$ and the mask function $\mu : X \to \mathbb{R}$,

define the linear operators $F_\mu, G : \mathcal{H} \to \mathcal{F}$ as follows:

$$F_\mu = \int_X \Phi(f(x)) \otimes \psi(x)\mu(x)\mathrm{d}x, \tag{4}$$

$$G = \int_X \Phi(g(x)) \otimes \psi(x)\mathrm{d}x. \tag{5}$$

Definition 4.1 is a compact notation for the the integral operators in the RKHS. Noticing that we can rewrite $\Delta_{f,g}$ (see below) as a function of F_μ and G is key to deriving the closed-form expression and the finite-dimensional approximation in Eqs. (10) and (11), that can be computed in practice.

When X is a bounded set, the two operators above are trace class and, by the representer property $\langle v, F_\mu h\rangle_{\mathcal{F}} = \int_X v(f(x))h(x)\mu(x)dx$ and also $\langle v, Gq\rangle_{\mathcal{F}} = \int_X v(f(x))q(x)dx$ (see Lemma C.1 in Appendix C for a detailed proof). Using this result and considering the linearity of $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ and the variational characterization of the Hilbert norm (i.e. $\|u\|_{\mathcal{F}} = \max_{\|v\|_{\mathcal{F}}\leq 1} |\langle v, u\rangle_{\mathcal{F}}|$), we have

$$\Delta_{f,g}(h, q) = \|F_\mu h - Gq\|^2_{\mathcal{F}},$$

for any $h, q \in \mathcal{H}$. From which we characterize D_λ as

$$D_\lambda(f, g) = \max_{\|h\|_{\mathcal{H}}\leq 1} \min_{q\in\mathcal{H}} \|F_\mu h - Gq\|^2_{\mathcal{F}} + \lambda \|q\|^2_{\mathcal{H}}. \tag{6}$$

To conclude, note that the optimization problem in the equation above has a closed-form expression in terms of the operatorial norm of an operator depending on F_μ and G. All the reasoning above is formalized below.

Theorem 4.2 (Closed-form solution)**.** *Let $X \subset \mathbb{R}^d$ be an open bounded set. Using the notations above, we have:*

$$D_\lambda(f, g) = \lambda\|(GG^* + \lambda I)^{-1/2}F_\mu\|^2_{op}. \tag{7}$$

The proof of Thm. 4.2 comes from identifying the inner optimization problem as a linear regression problem, and the the outer maximization problem as an eigenproblem. The complete proof is presented in Appendix C.1.

4.2. Approximate computation

Although Thm. 4.2 gives a closed-form expression of $D_\lambda(f, g)$, F_μ and G are defined as integral operators between infinite dimensional Hilbert spaces. In practice, we only have access to a discretization of f and g (e.g. an image is a discretized spatial signal represented by N pixels). A first natural approximation is thus to replace the integral with an empirical counterpart. This estimate is then a sum of rank-one operators between $\mathcal{H}$ and $\mathcal{F}$. To reduce the computational cost, while keeping good accuracy, we can then further approximate it using Nyström methods for kernels. The resulting estimator is $\widehat{D}_\lambda$ presented in Eq. (12), its convergence to D_λ is studied in Thm. 4.4.

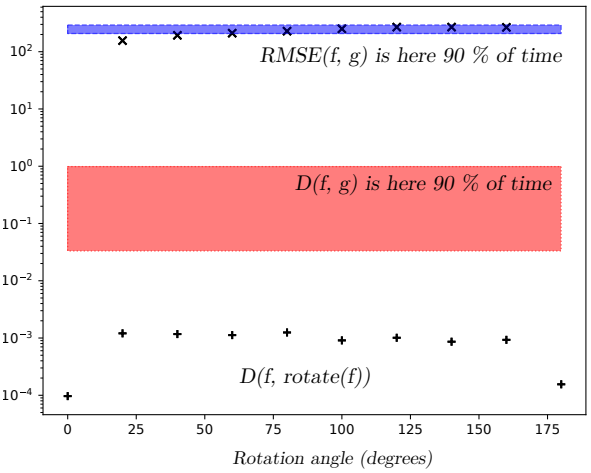


Figure 3. Invariance to rotation. We consider a patch f (size 100×100) of a larger scene (`peppers.jpeg`) and compare it to its rotated versions rotate(f, α), where α is an angle. $D(f, \text{rotate}(f, \alpha))$ is represented with $+$ symbols and RMSE$(f, \text{rotate}(f, \alpha))$ with $\times$ symbols. Random regimes are represented by shaded areas (see text). Both $\widehat{D}_\lambda$ and the RMSE seem constant as a function of α (although with $\alpha = 0$ or $\alpha = 180$ a smaller value is achieved). However, the RMSE of the rotated patches falls in (or close to) the confidence interval, making them indistinguishable from random patches from the same image. $\widehat{D}_\lambda$ takes values that are over $10\times$ smaller for rotated patches that for random patches. Hence, DID is invariant to rotation, whereas RMSE is constant (for $\alpha > 0$). Here $\lambda = 10^{-6}$. See Sec. 5.4 for more details.

Quadrature approximation. We replace F_μ with an estimator $F_{\mu,N}$ and G with an estimator G_N:

$$F_{\mu,N} = \frac{v_X}{N}\sum_{i=1}^{N} \Phi(f(x_i)) \otimes \psi(x_i)\mu(x_i) \tag{8}$$

$$G_N = \frac{v_X}{N}\sum_{i=1}^{N} \Phi(g(x_i)) \otimes \psi(x_i), \tag{9}$$

where $v_X := \int_X \mathrm{d}x$ is the volume of the domain X. Note that the set of $\{x_1, \ldots, x_N\}$ can be chosen at random or arbitrarily to best approximate the integrals. F and G can be approximated using different points. In practice, they are often given as the positions of pixels of images, sample times of a time series.

Nyström approximation. From the previous section, it is clear that $\text{rank}\,(F_{\mu,N}) \leq N$ and $\text{rank}\,(G_N) \leq N$. This justifies using the low-rank approximations we introduce in this section. It is possible to further reduce the rank of the matrices, while keeping a good accuracy, by using the so-called *Nyström approximation* (Williams and Seeger, 2001; Drineas et al., 2005; Rudi et al., 2015). Let $M_X, M_Y \in \mathbb{N}$ and choose the set of points $\widetilde{X} = \{\widetilde{x}_1, \ldots, \widetilde{x}_{M_X}\} \subset X$ and $\widetilde{Y} = \{\widetilde{y}_1, \ldots, \widetilde{y}_{M_Y}\} \subset Y$. The Nyström approximation of

a vector $v \in \mathcal{H}$ consists in the projected vector $P_{\tilde{X}} v$, where $P_{\tilde{X}} : \mathcal{H} \to \mathcal{H}$ is the projection operator with range corresponding to span$\{\psi(\tilde{x}_1), \dots, \psi(\tilde{x}_{M_X})\}$. Note, in particular, that $P_{\tilde{X}}$ has rank M_X when k_X is universal. $P_{\tilde{Y}} : \mathcal{F} \to \mathcal{F}$ on $\tilde{Y}$ in defined analogously.

Combining the two approximations. Assume in this section, that the kernels of choice are universal (as, e.g., the Laplacian or the Gaussian kernel). Let $P_{\tilde{X}} : \mathcal{H} \to \mathcal{H}$ be the projection operator associated to the Nyström points on X and $P_{\tilde{Y}}$ the one associated to the Nyström point on Y. Combining the two approximations, we define the following estimator for D_λ,

$$\widehat{D}_\lambda(f,g) := \lambda \| (P_{\tilde{Y}} G_N P_{\tilde{X}} G_N^* P_{\tilde{Y}} + \lambda)^{-\frac{1}{2}} P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}} \|_{op}^2.$$

Note, however, that $\widehat{D}_\lambda(f,g)$ has a finite dimensional characterization that we are going to derive now. Let $K_{\widetilde{Y}f} \in \mathbb{R}^{M_Y \times N}$ the matrix defined by $(K_{\widetilde{Y}f})_{i,j} = k_Y(\widetilde{y}_i, f(x_j))$, $K_{\widetilde{Y}g}$ in an analogous way, and finally, $K_{X\widetilde{X}} \in \mathbb{R}^{N \times M_X}$ the matrix defined by $(K_{X\widetilde{X}})_{i,j} = K_X(x_i, \widetilde{x}_j)$. Let $\widehat{\mu} = [\mu(x_1), \dots, \mu(x_{M_X})]$ and $R_{\widetilde{X}} \in \mathbb{R}^{M_X \times M_X}$ be the upper-triangular Cholesky decomposition of $K_{\widetilde{X}\widetilde{X}}$ defined by $(K_{\widetilde{X}\widetilde{X}})_{ij} = k_X(\widetilde{x}_i, \widetilde{x}_j)$. Analogously, define $K_{\widetilde{Y}\widetilde{Y}}$ and define $R_{\widetilde{Y}} \in \mathbb{R}^{M_Y \times M_Y}$ its Cholesky decomposition. Note that the decomposition exists since the kernel k_X is universal and so the kernel matrix $K_{\tilde{X},\tilde{X}}$ is invertible.

We introduce the following operators $\widehat{A}, \widehat{B} \in \mathbb{R}^{M_Y \times M_X}$, which are the finite dimensional representations in appropriate spaces $\mathbb{R}^{M_X}, \mathbb{R}^{M_Y}$ of, respectively, $\tilde{P}_Y F_{\mu,N} \tilde{P}_X$ and $\tilde{P}_Y G_N \tilde{P}_X$:

$$\widehat{A} = \frac{v_X}{N} R_{\widetilde{Y}}^{-T} K_{\widetilde{Y}f} \operatorname{diag}(\widehat{\mu}) K_{X\widetilde{X}} R_{\widetilde{X}}^{-1}, \tag{10}$$

$$\widehat{B} = \frac{v_X}{N} R_{\widetilde{Y}}^{-T} K_{\widetilde{Y}g} K_{X\widetilde{X}} R_{\widetilde{X}}^{-1}. \tag{11}$$

In particular, we have the following characterization for $\widehat{D}_\lambda$.

Lemma 4.3. *With the notation above,*

$$\widehat{D}_\lambda(f,g) = \lambda \| \widehat{A}^* (\widehat{B}\widehat{B}^* + \lambda I)^{-1} \widehat{A} \|_{op}. \tag{12}$$

Note that, in practice, $\widetilde{X}$ and $\widetilde{Y}$ can be chosen either deterministically, e.g. on a grid, or randomly. Now, we provide a bound on the approximation error associated to $\widehat{D}_\lambda(f,g)$. We assume that the N points in $x_1, \dots, x_N$ are sampled independently and uniformly at random in X and, moreover, that the M_X points in $\widetilde{X}$ and the M_Y points in $\widetilde{Y}$ are sampled independently and uniformly at random in, respectively, X, Y (similar result can be derived for a grid).

Theorem 4.4. *Let $\delta \in (0,1)$. Let $X \subset \mathbb{R}^d, Y \subset \mathbb{R}^p$ be bounded sets and k_X, k_Y be Sobolev kernels with smoothness, respectively, $s + d/2$ and $z + p/2$, for some $s, z > 0$. There exists two constants c_1, c_2 s. t., when $M_X \geq c_1$ and $M_Y \geq c_2$, then the following holds with probability $1 - \delta$,*

$$|\widehat{D}_\lambda(f,g) - D_\lambda(f,g)| \leq c\Big(\frac{\log \frac{1}{\delta}}{\lambda\sqrt{N}} + \frac{(\log \frac{M_X}{\delta})^\alpha}{\lambda M_X^{s/d}} + \frac{(\log \frac{M_Y}{\delta})^\beta}{\lambda M_Y^{z/p}}\Big),$$

for any measurable $f, g : X \to Y$, where $\widehat{D}_\lambda(f,g)$ is defined as in Eq. (12). Here c_1, c_2, c depend only on X, Y, μ, s, z, d, p, while $\alpha = s/d + 1/2$, $\beta = z/p + 1/2$.

The theorem above shows that the estimation error of $\widehat{D}_\lambda$ with respect to D_λ goes to 0 when $N, M_X, M_Y \to \infty$. On the contrary, the error diverges in λ. This is in accordance with the fact that the error is of variance type and shows that λ plays the role of a regularization parameter. The bound shows also that, when $s \gg d$ and $z \gg p$, i.e. when we are choosing very smooth Sobolev kernels, the decay rate of the error in M_X and M_Y is faster. For example, if we choose $s = rd$, $z = rp$, for some $r > 0$, then choosing $M_X = M_Y = O(N^{r/2})$ leads to the rate

$$|\widehat{D}_\lambda(f,g) - D_\lambda(f,g)| = O\left(\frac{1}{\lambda\sqrt{N}}\right).$$

On the choice of λ. To conclude, a choice of λ as $\lambda = N^{-1/4}$ guarantees a final convergence rate of $\widehat{D}_\lambda$ to D_λ in the order of $N^{-1/4}$ and, together with Thm. 3.2 a level of invariance to diffeomorphism for $\widehat{D}_\lambda$ of the order

$$\widehat{D}_\lambda(f, f \circ Q) = O(N^{-1/4} C_\mu^2 C_Q^2),$$

which can become a statistically significant threshold to decide if, in practice $f \approx g$ up to diffeomorphism. Clearly the choice of r, s while reducing the number of Nyström points required in the approximation (with important computational implications that we see below) increases the constant C_Q as shown, e.g., in Example 3.3, where $m = s + d/2$.

Algorithm and computational complexity. The final form of the empirical estimator $\widehat{D}_\lambda$ is Eq. (12). An efficient algorithm to compute it consists in (1) first computing the matrices $\widehat{A}$ and $\widehat{B}$, then the inverse $\widehat{C} = (\widehat{B}\widehat{B}^* + \lambda)^{-1}$ and finally compute the largest eigenvalue of $\widehat{A}^* \widehat{C} \widehat{A}$ via a power iteration method (Trefethen and Bau III, 1997).

Assuming that (a) the cost of one kernel computation in $\mathbb{R}^d$ is $O(d)$ (as in the case of any translation invariant kernel as the Laplace kernel) (b) $M_X \leq N$ and $M_Y \leq N$ (which is reasonable in light of Thm. 4.4), then the cost of computing $\widehat{D}_\lambda$ with the algorithm above is $O(dNM_X + pNM_Y + M_X M_Y^2 + M_X^3 + M_Y^3)$. Choosing the parameters, as in the discussion after Thm. 4.4, with $r = 1$, would lead to a total computational cost of $O(N^{1.5}(p + d))$.

Computing DID between different pairs of data-points can easily be parallelized as each computation relies on matrix-vector products and matrix inversions.

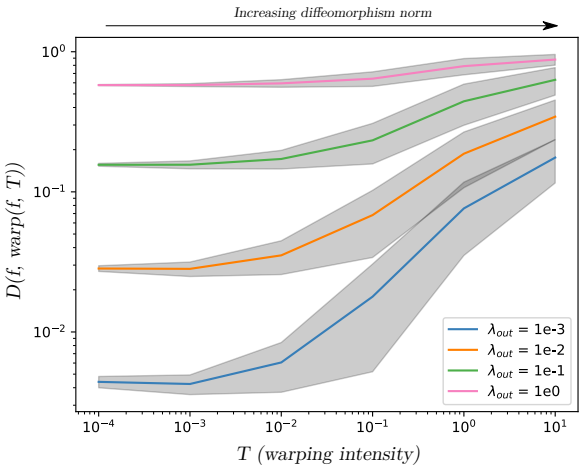


Figure 4. Effect of regularization. We consider the same image and setting as in Fig. 2 and vary T and λ (shaded areas are std deviation over 50 warps). Observe: (1) $\widehat{D}_\lambda$ increases as a function of T (as the norm of the diffeomorphism increases); (2) $\widehat{D}_\lambda$ is proportional to λ. Both of these phenomena are predicted by Thm. 3.2. See Sec. 5.5 for more details.

Choice of kernel hyperparameters. As in any kernel-based method, selecting kernel hyper-parameters is important. The usual considerations apply: we advise cross-validation should be used to choose hyper-parameters (such as the bandwith). We do not discuss this further for space considerations.

In the theoretical analysis above, the rates in N hold for any parameter of the kernel, but constants depend on the chosen parameters. This is typical of the Sobolev analysis technique, where the important parameter is the kernel regularity m, which corresponds to the differentiability class of the diffeormorphisms one aims to capture.

Finally, the method introduced is empirically quite robust to the choice of parameters.

5. Experiments

This section investigates the empirical performance of DID. We observe that, in line with our results in Thm. 3.2, when $g = f \circ Q$ the resulting D_λ is small, while it is consistently large for signals that are not diffeomorphic versions of f.

5.1. Implementation details

We implemented $\widehat{D}_\lambda$ as in Eq. (12) as described in the end of Sec. 4.2 using standard linear algebra routines such as matrix and matrix-vector products, matrix inversions, and eigendecompositions. The Python source code used for the experiments presented here is freely available at https://github.com/theophilec/diffy, depends on Numpy and Pytorch and supports using GPUs.

Normalization. In practice we normalize $\widehat{A}$ and $\widehat{B}$ by their operator norms $\|\widehat{A}\|_{op}$ and $\|\widehat{B}\|_{op}$. This normalizes D_λ between 0 and 1 and makes interpretation easier.

Choice of mask. We choose μ to be a Blackman window, a standard windowing function in image processing. In 1-D, the Blackman window is defined for any $0 \leq t \leq 1$ as: $\mu(t) = 0.42 - 0.5\cos(2\pi t) + 0.08\cos(4\pi t)$ (Oppenheim et al., 1999). We generalize it to higher-dimension by considering its tensor-product over dimensions.

Choice of kernel. Because we work with images in the experiments we present, $X = \mathbb{R}^2$ (coordinate space) and $Y = \mathbb{R}^3$ (color space). We consider the Gaussian kernel defined as $k(x, x') = \exp(-\|x-x'\|^2/(2\sigma^2))$ on X and the Laplace kernel defined as $k(y, y') = \exp(-a\|y-y'\|)$ on Y. For the experiments presented in this paper, unless otherwise stated, DID has parameters: $M_X = 100$, $M_Y = 16^3$, $\sigma = 1/6$ and $a = 5$. **Datasets.** We rely on images from Imagenet (more precisely from the `Imagenette` subset), example images from the Matlab software (`peppers`), and finally images taken with our personal devices for illustrations (`raccon`, `flowers`, `objects`). All images are made available with the source code.

Diffeomorphism generation. Diffeomorphism are obtained either by affine transformations or by generating warpings. In particular, for warpings, we use the code from (Petrini et al., 2021). We generate random transformations of images by displacing each of its coordinates independently (while enforcing zero displacement at the edge of the grid) then interpolating the colors. We choose the standard deviation of the displacements of each pixel so as to obtain a transformation of given (average) displacement norm. This can be controlled by two parameters T (a temperature, between 10^{-4} and 10) and c (a cut-off parameter, taking $c = 2$). We denote warp(f, T) such a (random) warp. Examples of warps for various T parameters (and samples) are provided in Appendix H.

5.2. Illustrative examples

In Fig. 1, we show how the h and q that optimize DID concentrate on related regions on for different scenes which which are related by a diffeomorphism, in particular rigid-body and perspective for `objects` and keystone deformation for `raccoon`. We present supplementary illustrative examples in Appendix G.

5.3. Invariance to warping

Diffeomorphisms (even infinitely regular) are a much wider class of transformations than rigid body transformations such as scale, rotation and translation (or combinations thereof). In this experiment (Fig. 2), we evaluating DID's

behavior against a wide family of transformations (we call warps). Consider f an image from `Imagenette`. For $T = 10^k$ for $-4 \leq k \leq 2$ and various images, we evaluate $\widehat{D}_\lambda(f, \text{warp}(f, T))$ as well as $\text{RMSE}(f, \text{warp}(f, T))$. We compare the values observed to $\widehat{D}_\lambda(f, g)$ and $\text{RMSE}(f, g)$ for random images f and g. We call this the *random regime* (90% confidence interval). Finally, we look at the performance of DID on a fixed image (gas station), with repeating warps. Fig. 2 shows that DID is invariant to diffeomorphic warping for $T \leq 10^{-1}$ whereas the Euclidean distance increases exponentially with T (making $\text{warp}(f, T)$ indistinguishable from g, a random image).

5.4. Invariance to rotation

The `peppers` image is often used to demonstrate image registration techniques. In this experiment (see Fig. 3) we show that DID is invariant to rotation using patches taken from it. Consider f a patch from `peppers` and $g = \text{rotate}(f, \alpha)$, rotated version of f by angle α (in practice, we rotate a larger patch then crop to avoid artifacts). We then compare $D(f, \text{rotate}(f, \alpha))$ and $\text{RMSE}(f, \text{rotate}(f, \alpha))$. We show that while the Euclidean distance is *constant* for $\alpha \neq 0$, DID is *invariant*. We compare DID's and the Euclidean distance values with their values for random patches from the image. As before, we call this the *random regime* (90% confidence interval). This shows that the Euclidean distance is not able to distinguish between a random patch and a rotated version of the same patch, while DID can. See Fig. 3 for the results of the experiments.

5.5. Effect of regularization

In order to understand the effect of regularization, we reuse the setup from Sec. 5.3 with a single image, with varying λ. In Fig. 4, we observe two phenomena: (1) as T increases, so does D; (2) as λ decreases, so does D. This shows that DID behaves close to what is predicted by Thm. 3.2. Indeed, D seems proportional to λ. Also, as T increases, so does the norm of the transformation between f and $\text{warp}(f, T)$. This makes the upper bound of Thm. 3.2 increase in turn.

Acknowledgements

T.C. gratefully acknowledges support from the French National Agency for Research, grant ANR-18-CE40-0016-01. C.C. acknowledges the support of the Royal Society (grant SPREM RGS\R1\201149) and Amazon.com Inc. (Amazon Research Award – ARA). B.G. acknowledges partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1; B.G. also acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02. A.R. acknowledges partial support from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and support from the European Research Council (grant REAL 947908).

References

Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

Charalambos D Aliprantis and Owen Burkinshaw. *Principles of real analysis*. Gulf Professional Publishing, 1998.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20(25):1–49, 2019.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Gérard Bourdaud and Winfried Sickel. Composition operators on function spaces with fractional order of smoothness. *Harmonic Analysis and Nonlinear Partial Differential Equations*, 26:93–132, 2011.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

Martins Bruveris. Regularity of maps between sobolev spaces. *Annals of Global Analysis and Geometry*, 52(1): 11–24, 2017.

Samuel Cohen, Giulia Luise, Alexander Terenin, Brandon Amos, and Marc Deisenroth. Aligning time series on incomparable spaces. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 894–903. PMLR, 06–11 Aug 2017.

E. De Castro and C. Morandi. Registration of translated and rotated images using finite Fourier transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):700–703, 1987.

Dennis DeCoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1):161–190, 2002.

Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005.

Bernard Haasdonk and Hans Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine learning*, 68(1):35–61, 2007.

Imre Risi Kondor. Group theoretical methods in machine learning. *PhD thesis*, 2008.

Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Francis Narcowich, Joseph Ward, and Holger Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250):743–763, 2005.

Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. Prentice-hall Englewood Cliffs, second edition, 1999.

Leonardo Petrini, Alessandro Favero, Mario Geiger, and Matthieu Wyart. Relative stability toward diffeomorphisms indicates performance in deep nets. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

B.S. Reddy and B.N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8): 1266–1271, 1996.

Alessandro Rudi and Carlo Ciliberto. PSD representations for effective probability models. *Advances in Neural Information Processing Systems*, 34, 2021.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *NIPS*, pages 1657–1665, 2015.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *arXiv preprint arXiv:2012.11978*, 2020.

Thomas Runst and Winfried Sickel. *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations*. de Gruyter, 2011.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26 (1):43–49, 1978.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. *arXiv preprint arXiv:2101.05380*, 2021.

Titouan Vayer, Laetitia Chapel, Nicolas Courty, Rémi Flamary, Yann Soullard, and Romain Tavenard. Time series alignment with global invariances. preprint, February 2020.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems*, pages 682–688, 2001.

Appendix

A. Background

We recall here the classical version of change of variable theorem for $\mathbb{R}^d$, see e.g. Thm. 40.7 of Aliprantis and Burkinshaw (1998).

Theorem A.1 (Change of Variables)**.** *Let V be an open set in $\mathbb{R}^d$ and $Q : \mathbb{R}^d \to \mathbb{R}^d$ be an injective continuosly differentiable map. Let $f \in L^1(V)$. Denote by $\nabla Q(x) \in \mathbb{R}^{d\times d}$ the gradient of Q for any $x \in \mathbb{R}^d$. Then, we have*

$$\int_{Q^{-1}(V)} f(Q(x))|\nabla Q(x)|dx = \int_V f(y)dy.$$

A.1. Sobolev spaces

We recall some basic properties of Sobolev spaces. The Sobolev space $H^m(Z)$ for $m > 0$ and an open set $Z \subseteq \mathbb{R}^d$ is defined as follows (Adams and Fournier, 2003)

$$H^m(Z) = \left\{f \in L^2(\mathbb{R}^d) \;\middle|\; \|f\|_{H^m(Z)}\right\} < \infty\}, \quad \|f\|^2_{H^m(Z)} = \sum_{\alpha_1+\cdots+\alpha_d\le m} \int_Z \left|\frac{\partial^{\alpha_1+\cdots+\alpha_d} f(x)}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}}\right|^2 dx.$$

Analogously, we define the space $H^m(Z, \mathbb{R}^t)$ for functions with output in $\mathbb{R}^t$, $t \ge 1$, with norm $\|f\|^2_{H^m(Z,\mathbb{R}^t)} = \sum_{j=1}^t \|f_j\|^2_{H^m(Z)}$. In the following theorem we collect some important properties of $H^m(Z)$ that will be useful for the proof of Thm. 3.2.

Theorem A.2. *Let $m > d/2$ and Z be an open set with locally Lipschitz continuous boundary. The following properties hold*

(a) *$H^m(Z)$ is a Reproducing Kernel Hilbert space. Moreover, $H^m(Z) \subset C(Z)$, and, when Z is bounded we have $f|_Z \in H^m(Z)$, for any $f \in C^m(\mathbb{R}^d)$.*

(b) Restriction and extension. *For any $u \in H^m(\mathbb{R}^d)$, it holds that $u|_Z \in H^m(Z)$ and $\|u|_Z\|_{H^m(Z)} \le c_1\|u\|_{H^m(\mathbb{R}^d)}$. Moreover, for any $u \in H^m(Z)$ there exists a function $E_Z[u] \in \mathcal{H}(\mathbb{R}^d)$ such that $(E_Z[u])|_Z = u$ and $\|E_Z[u]\|_{H^m(\mathbb{R}^d)} \le c_2\|u\|_{H^m(Z)}$. The constants c_1, c_2 depend only on Z, m, d.*

(c) Pointwise product. *For any $u, v \in H^m(Z)$ we have $u \cdot v \in H^m$. In particular, there exists $c_0 > 0$ depending only on Z, m, d such that $\|u \cdot v\|_{H^m(Z)} \le c_0\|u\|_{H^m(Z)}\|v\|_{H^m(Z)}$.*

Proof. The space $H^m(\mathbb{R}^d)$ is a RKHS due to the characterization of the norm with respect to the Fourier transform (Wendland, 2004). The space $H^m(Z)$ is a RKHS since any restriction of a RKHS to a subset of the set of definition is still a RKHS (Aronszajn, 1950). The inclusions derive directly by the definition of the H^m norm (see Adams and Fournier, 2003, for more embeddings of Sobolev spaces).

The second point is a classical result on Sobolev space and is derived in Adams and Fournier (2003). The third point is equivalent to showing that the Sobolev spaces with $m > d/2$ are Banach algebras and it is derived in Adams and Fournier (2003). For the case $Z = \mathbb{R}^d$ an explicit derivation based on the Fourier transform is done in Lemma 10 of Rudi et al. (2020). □

B. Proof of the more general version of Theorem 3.2

In the next subsection we introduce some preliminary results that will be necessary to prove the more general version of Theorem 3.2, which is in the subsection Appendix B.2.

B.1. Preliminary result on composition of Sobolev functions

The following theorem quantifies the fact that a Sobolev space with $m > d/2$ is closed with respect to composition with diffeomorphisms. There exists many abstract results about the closure of composition in Sobolev space in the literature (see

e.g. Bruveris, 2017). Here, however, we want a quantitative bound. We base our result on the explicit bound in Bourdaud and Sickel (2011). To obtain a final readable form we have to do a bit of slalom between restriction and extension between Z and $\mathbb{R}^d$. Indeed, we need functions that are equivalent to the functions of interest on Z, but whose norm does not diverge when going to $\mathbb{R}^d$. For example, constant functions don't belong to $H^m(\mathbb{R}^d)$, but for us it is enough to have a function that is equal to a constant on Z and that goes to zero at infinity fast enough to have finite $H^m(Z)$ norm.

Theorem B.1 (Smooth composition)**.** *Let $m > d/2$. Let $Q : \mathbb{R}^d \to \mathbb{R}^d$ be an invertible m-times differentiable map whose inverse has continuous Lipschitz derivative. Let Z be open bounded set with Lipschitz boundary and a compact Ω such that $\Omega \subset Z$ and $Q^{-1}(\Omega) \subset Z$. Assume also, without loss of generality, that 0 is in the interior of Ω. For any $h \in H^m(\mathbb{R}^d)$ supported on Ω the following holds:*

1. *there exists $b \in H^m(\mathbb{R}^d)$ satisfying $b(x) = h(Q(x))$ for any $x \in Q^{-1}(\Omega)$.*
2. *if $h(x) = 0$ for any $x \in \mathbb{R}^d \setminus \Omega$, then $b(x) = 0$ for any $x \in \mathbb{R}^d \setminus Q^{-1}(\Omega)$,*
3. *the norm of b is controlled by*

$$\|b\|_{H^m(\mathbb{R}^d)} \;\leq\; C\,\|h\|_{H^m(Z)}\, d_{\Omega,Q}^{-m-d/2}\,(1 + \|Q\|_{H^m(Z,\mathbb{R}^d)} + D^m + L^m), \tag{13}$$

where $d_{\Omega,Q} := \min[1,\ d_H(Q^{-1}(\Omega), Q^{-1}(Z) \cap Z)]$, $d_H(A, B)$ is the Haussdorff distance between two sets A, B and $D :=$ diam(Z), $L = \max_{x\in Z} \|\nabla Q(x)\|$, while C depends only on d, m, Z.

Proof. The fact

To apply Bourdaud and Sickel (2011) we need a function such that $h(0) = 0$. With this aim, we rewrite $h \circ Q$ as

$$h \circ Q = h(0) + ((h - h(0)) \circ Q).$$

Step 1. Construction of b. Denote by U the open set $U = Q^{-1}(Z) \cap Z$. Note that the set is not empty since, by construction, $Q^{-1}(\Omega) \subset U$. Define $\tilde{Q} = E_Z[Q|_Z]$, i.e. the extension to the whole $\mathbb{R}^d$ of the restriction of Q on the set Z (restriction and extension done componentwise). By the first point we have that $Q|_Z$ belongs to $H^m(Z, \mathbb{R}^d)$ and, by the second point, that $\tilde{Q}$ belongs to $H^m(\mathbb{R}^d, \mathbb{R}^d)$ and moreover $\tilde{Q}(x) = Q(x)$ for any $x \in Z$. Denote by $s = h(0) \in C^\infty(\mathbb{R}^d)$ the constant function equal to $h(0)$ everywhere on $\mathbb{R}^d$. In particular, note that $s|_Z \in H^m(Z)$ via Thm. A.2(b). Denote by τ, the extension of $s|_Z$ to $\mathbb{R}^d$, i.e., $\tau = E_Z[s|_Z]$. Denote by u, the function $u = h|_Z - s|_Z$ and by $\tilde{u} \in H^m(\mathbb{R}^d)$ the function $\tilde{u} = E_Z[h|_Z - s|_Z]$. Note that $\tilde{u}(x) = h(x) - h(0)$ for any $x \in Z$, and, in particular for any $x \in \Omega$. Define by ρ the $C^\infty(\mathbb{R}^d)$ function that is 1 on $Q^{-1}(\Omega)$ and 0 on $\mathbb{R}^d \subseteq U$.

Now define,

$$b \;=\; \rho \cdot (\tau \,+\, \tilde{u} \circ \tilde{Q}).$$

Denote by $\tilde{b}$ the function $\tilde{b} = \tau + \tilde{u} \circ \tilde{Q}$. We have that $\tilde{b}(x) = h(Q(x))$ for any $x \in U$. Since $Q(U) \subseteq Z$ and that $\tilde{h}(x) = h(x), \tilde{Q}(x) = Q(x)$ for any $x \in Z$. In particular, since $\tilde{b}(x) = 0$ for all $x \in U \setminus Q^{-1}(\Omega)$ and by definition of ρ, we have: (a) $\rho \cdot \tilde{b} = \tilde{b} = h(Q(x))$ on $Q^{-1}(\Omega)$, (b) $\rho \cdot \tilde{b} = 0$ on $UQ^{-1}(\Omega)$; (c) $\rho \cdot \tilde{b} = 0$ and 0 on $\mathbb{R}^d \setminus U$. Then $b = \rho \cdot \tilde{b} = h(Q(x))$ for any $x \in \mathbb{R}^d$.

Step 2. Bound of b. Now, let us bound the norm of b. By applying Thm. A.2(c) we have

$$\|b\|_{H^m(\mathbb{R}^d)} \leq c\|\rho\|_{H^m(\mathbb{R}^d)}(\|\tau\|_{H^m(\mathbb{R}^d)} \;+\; \|\tilde{u} \circ \tilde{Q}\|_{H^m(\mathbb{R}^d)}). \tag{14}$$

The bound for τ is obtained applying Thm. A.2(b) and the definition of $H^m(Z)$ norm, as follows,

$$\|\tau\|_{H^m(\mathbb{R}^d)} \leq c_2(Z)\|s|_Z\|_{H^m(Z)} = c_2 h(0)\text{vol}(Z)^{1/2}.$$

Now we bound the norm $\|\tilde{u} \circ \tilde{Q}\|_{H^m(\mathbb{R}^d)}$ with respect to the norms of $\tilde{u}$ and $\tilde{Q}$. We use a result on the composition of Sobolev functions (Bourdaud and Sickel, 2011, Theorem 27) that is highly technical, but allows to highlight the quantities of interests for us. For a more extensive treatment of the topic see for example Runst and Sickel (2011). Since $\tilde{u}(0) = 0$ by

construction, by applying Theorem 27 of Bourdaud and Sickel (2011) with $p = 2$ and considering that their norm $\|\cdot\|_{\dot{W}^m_{Ep}}$ is bounded by $\|\cdot\|_{H^m(\mathbb{R}^d)}$ and analogously $\|\cdot\|_{\dot{W}^1_{mp}} \leq \|\cdot\|_{\dot{W}^1_\infty} \leq \|\cdot\|_{W^1_\infty(\mathbb{R}^d)}$,

$$\|\tilde{u} \circ \tilde{Q}\|_{H^m(\mathbb{R}^d)} \leq c(\|\tilde{u}\|_{H^m(\mathbb{R}^d)} + \|\tilde{u}\|_{\dot{W}^1_\infty(\mathbb{R}^d)})(\|\tilde{Q}\|_{H^m(\mathbb{R}^d,\mathbb{R}^d)} + \|\tilde{Q}\|^m_{\dot{W}^1_\infty(\mathbb{R}^d,\mathbb{R}^d)}) \tag{15}$$

Here $\|z\|_{\dot{W}^1_\infty(A,\mathbb{R}^d)} = \sup_{x\in A, j\in\{1,\dots,t\}} \|\nabla z_j\|$ for any differentiable $z : A \to \mathbb{R}$ and open set A. Now to conclude, note that

$$\|\tilde{u}\|_{H^m(\mathbb{R}^d)} = \|E_Z[h|_Z - s|_Z]\|_{H^m(\mathbb{R}^d)} \leq c_2\|h|_Z - s|_Z\|_{H^m(Z)} \leq c_2\|h\|_{H^m(Z)} + c_2 h(0)\text{vol}(Z)^{1/2},$$

moreover

$$\|\tilde{Q}\|_{H^m(\mathbb{R}^d,\mathbb{R}^d)} = \|E_Z(Q|_Z)\|_{H^m(\mathbb{R}^d,\mathbb{R}^d)} \leq c_2\|Q|_Z\|_{H^m(Z,\mathbb{R}^d)}.$$

Note also that for the Sobolev space W^1_∞ there exists the same type of result as Thm. A.2(b) and for the same extension operator defined in Thm. A.2(b) (which is a *total extension operator*, see e.g. the Stein extension theorem, Thm 5.4 page 154 of Adams and Fournier, 2003), so, as above, we have

$$\|\tilde{u}\|_{W^1_\infty(\mathbb{R}^d)} \leq c_2\|h\|_{W^1_\infty(Z)} + c_2 h(0),$$

and also

$$\|\tilde{Q}\|_{W^1_\infty(\mathbb{R}^d,\mathbb{R}^d)} \leq c_2\|Q|_Z\|_{W^1_\infty(Z,\mathbb{R}^d)} = c_2 \sup_{x\in Z} \max(\|Q(x)\|, \|\nabla Q(x)\|) \leq c_2\text{diam}(Z) + c_2 \sup_{x\in Z} \|\nabla Q(x)\|.$$

Substituting the six bounds above in Eq. (14), we obtain

$$\|b\|_{H^m(\mathbb{R}^d)} \leq c\|\rho\|_{H^m(\mathbb{R}^d)}(c_2 h(0)\text{vol}(Z)^{1/2} \; + \; c'(\|h\|_{H^m(W)} + \|h\|_{W^1_\infty(Z)} + h(0)c'')(\|Q|_Z\|_{H^m(Z,\mathbb{R}^d)} + D^m + L^m)),$$

where $D = \text{diam}(Z)^m$, $L = \sup_{x\in Z} \|\nabla Q(x)\|$ and $c' = cc_2^2(1 + 2^m c_2^{m-1})$, $c'' = 1 + \text{vol}(Z)^{1/2}$, with c, c_2 depending only on d, m, Z. The final result is obtained considering that $xA + d(B + yA)R \leq (1 + d + dy)(B + A)(1 + R)$ for any $x, y, d, A, B, R \geq 0$, and applying this result with $x = c_2(Z)\text{vol}(Z)^{1/2}$, $A = h(0)$, $d = c'$, $B = \|h\|_{H^m(Z)} + \|h\|_{W^1_\infty(Z)}$, $y = c''$, $R = \|Q|_Z\|_{H^m(Z,\mathbb{R}^d)} + D^m + L^m$. In particular, in the final result, the constant C corresponds to $C = 3(1+d+dy)$ and we used the fact that $h(0) \leq \|h\|_{W^1_\infty(Z)} \leq \|h\|_{H^m(Z)}$, then $\|h\|_{H^m(Z)} + \|h\|_{W^1_\infty(Z)} + h(0) \leq 3\|h\|_{H^m(Z)}$.

Step 3. The norm of ρ . Let A_t be the set $A_t = \{x \mid \min_{y\in Q^{-1}(\Omega)} \|x - y\| \leq t\}$. Let $\eta = d_H(Q^{-1}(\Omega), \overline{U})$, i.e., the Haussdorff distance between the sets $Q^{-1}(\Omega)$ and $\overline{U}$, corresponding to the largest η for which $A_\eta \subseteq U$. Note that $\eta > 0$ since $Q^{-1}(\Omega)$ is compact, while U is open and $Q^{-1}(\Omega) \subset U$. The fact that $\eta > 0$ implies that for any $\eta' < \eta$ it holds that $A_{\eta'} \subset U$.

The function ρ can be obtained by the convolution of the indicator function $1_{A_{\eta/2}}$ with the bump function $\psi_{\eta/2}(x) = (\eta/2)^{-d}\psi(x/(\eta/2))/S$ where $S = \int_{\mathbb{R}^d} \psi(x)dx$ and ψ is an infinitely smooth non-zero non-negative function that is 0 on $\|x\| \geq 1$ as for example $\psi(x) = \exp(-1/(1 - \|x\|^2)_+)$ for any $x \in \mathbb{R}^d$, where $(z)_+ = \max(0, z)$. In particular, since (a) $\|f(y - \cdot)\|_{H^m(\mathbb{R}^d)} = \|f\|_{H^m(\mathbb{R}^d)}$ for any $y \in \mathbb{R}^d$, by construction of the norm $H^m(\mathbb{R}^d)$, and (b) $\|t^{-d} f(\cdot/t)\|_{H^m(\mathbb{R}^d)} \leq c_6 t^{-d/2} \max(1, t^{-m})\|f\|_{H^m(\mathbb{R}^d)}$ (see, e.g., Proposition 3 of Runst and Sickel, 2011) we have

$$\begin{aligned}\|\rho\|_{H^m(\mathbb{R}^d)} &\leq \int_{A_{\eta/2}} \|\psi_\eta(y - \cdot)\|_{H^m(\mathbb{R}^d)} dy \leq \text{vol}(A_{\eta/2})\|\psi_\eta\|_{H^m(\mathbb{R}^d)} \\ &\leq c_6\|\psi\|_{H^m(\mathbb{R}^d)} \text{ vol}(A_{\eta/2}) \, (\eta/2)^{-d/2} \max(1, (\eta/2)^{-m}).\end{aligned}$$

To conclude note that $\text{vol}(A_\eta) \leq \text{vol}(Z)$, since $A_{\eta/2} \subset U \subseteq Z$. □

Lemma B.2 (Existence and norm of $\tilde{q} \in H^m(\mathbb{R}^d)$)**.** *Let μ be an infinitely differentiable function, with compact support $\Omega \subset \mathbb{R}^d$. Let $Q : \mathbb{R}^d \to \mathbb{R}^d$ be a C^{m+1} diffeomorphism. Let Z be an open bounded set with Lipschitz boundary and such that $\Omega \subset Z$ and $Q^{-1}(\Omega) \subset Z$. Moreover, assume without loss of generality that 0 is in the interior of Ω. For any $h \in H^m(\mathbb{R}^d)$, there exists a function $\tilde{q} \in H^m(\mathbb{R}^d)$ satisfying*

$$\tilde{q}(y) = \begin{cases} h(Q(y))\mu(Q(y))|\nabla Q(y)| & y \in Q^{-1}(\Omega) \\ 0 & y \in \mathbb{R}^d \setminus Q^{-1}(\Omega) \end{cases}. \tag{16}$$

In particular, let $b \in H^m(\mathbb{R}^d)$ be defined according to Thm. B.1, then

$$\|q\|_{H^m(\mathbb{R}^d)} \leq C' \|h\|_{H^m(\mathbb{R}^d)} \|\mu\|_{H^m(\mathbb{R}^d)} d_{\Omega,Q}^{-m-d/2} \|\nabla Q\|^d_{H^m(Z,\mathbb{R}^d)} (1 + \|Q\|_{H^m(Z,\mathbb{R}^d)} + D^m + L^m),$$

the constant C depends only on d, m, Z, while $d_{\Omega,Q} := \min[1, d_H(Q^{-1}(\Omega), Q^{-1}(Z) \cap Z)$ and $d_H(A, B)$ is the Haussdorff distance between two sets A, B.

Proof. Let $h \in H^m(\mathbb{R}^d)$ and $\tilde{\mu}$ to be the extension of μ on $\mathbb{R}^d$ (which corresponds to $\tilde{\mu}(x) = \mu(x)$ for any $x \in \Omega$ and $\tilde{\mu}(x) = 0$ for $x \in \mathbb{R}^d \setminus \Omega$). Define the function $s(x) = h \cdot \tilde{\mu}$ and note that $s \in H^m(\mathbb{R}^d)$ since it is the product of two functions in $H^m(\mathbb{R}^d)$ (see Thm. A.2(c)).

Second, note that the function $r = (|\nabla Q|)|_Z \in H^m(Z)$ since (a) the map ∇Q belongs to $C^{m+1}(\mathbb{R}^d, \mathbb{R}^d)$ so its entries $(\nabla Q)_{i,j}|_Z$ belong to $H^m(Z)$ (see Thm. A.2(a)); and (b) the determinant of a matrix, by the Leibniz formula, is defined in terms of sums and products of its entries and $H^m(Z)$ is closed with respect to multiplication, by Thm. A.1. To quantify its norm let's write explicitly the Leibniz formula (Trefethen and Bau III, 1997),

$$|\nabla Q| = \sum_{\sigma \in S_d} sgn(\sigma) \prod_{i=1}^{d} e_{\sigma_i}^\top \frac{\partial Q(x)}{\partial x_i},$$

where $sgn(\sigma) \in \{-1, 1\}$, S_n is the set of permutations of of d elements and $e_1, \ldots, e_d$ is the canonical basis of $\mathbb{R}^d$. Note, in particular, that, by the equation above, since $|S_d| = d!$ and that $\|e_{\sigma_i}^\top \frac{\partial Q(x)}{\partial x_i}\|_{H^m(Z)} \leq \|\nabla Q\|_{H^m(Z,\mathbb{R}^d)}$, we have that

$$\|r\|_{H^m(Z)} \leq d! \|\nabla Q\|^d_{H^m(Z,\mathbb{R}^d)}.$$

Now consider the function $b \in H^m(\mathbb{R}^d)$ defined according to Thm. B.1 we have that $b(x) = s(Q(x))$ for any $x \in \mathbb{R}^d$. Now, define $\tilde{q}$ as follows

$$\tilde{q} = b \cdot E_Z[r].$$

The function $\tilde{q}$ is in $H^m(Z)$, since it is the product of two functions in $H^m(Z)$ (see Thm. A.2(c) and $r \in H^m(Z)$ (see Thm. A.2(b)). Note, in particular, that Eq. (16) holds, by construction. To conclude, note that, by applying Thm. A.2(c) and Thm. A.2(b), we have

$$\|q\|_{H^m(\mathbb{R}^d)} \leq c_0 \|E_Z[r]\|_{H^m(\mathbb{R}^d)} \|b\|_{H^m(\mathbb{R}^d)} \leq c_0 c_2 \|r\|_{H^m(Z)} \|b\|_{H^m(\mathbb{R}^d)} \leq d! c_0 c_2 \|\nabla Q\|^d_{H^m(Z,\mathbb{R}^d)} \|b\|_{H^m(\mathbb{R}^d)}.$$

To conclude, we bound b according with Thm. B.1, obtaining

$$\|b\|_{H^m(\mathbb{R}^d)} \leq C \, \|s\|_{H^m(W)} \, d_{\Omega,Q}^{-m-d/2} (1 + \|Q\|_{H^m(Z,\mathbb{R}^d)} + D^m + L^m),$$

and $\|s\|_{H^m(W)} \leq \|s\|_{H^m(\mathbb{R}^d)} \leq c_0 \|h\|_{H^m(\mathbb{R}^d)} \|\mu\|_{H^m(\mathbb{R}^d)}$, by Thm. A.2(c). □

B.2. Proof of the general version of Theorem 3.2

Thm. 3.2 is a particular case of the following theorem

Theorem B.3. *Let $X \subset \mathbb{R}^d$ be an open bounded set with Lipschitz boundary. Let $\mu \in C^\infty(\mathbb{R}^d)$ with compact support $\Omega \subset X$. Let the RKHS $\mathcal{H}$ be $\mathcal{H} = H^m(\mathbb{R}^d)$, i.e. the Sobolev space of smoothness m, with $m > d/2$. Denote by $\mathcal{F}$ the RKHS induced by the kernel k_Y on Y, that we assume uniformly bounded. Then for any C^{m+1} diffeomorphism Q on $\mathbb{R}^d$ satisfying $Q^{-1}(\Omega) \subset X$, we have that for all f, g measurable functions,*

$$g(x) = (f \circ Q)(x), \;\; \forall x \in Q^{-1}(\Omega) \qquad \textit{implies} \qquad D(f, g) \;\; \leq \;\; \lambda \, C_\mu C_Q,$$

where $C_\mu = \|\mu\|_{H^m(\mathbb{R}^d)}$, and C_Q is defined in Eq. (19) below and depends only on Ω, X, Q, d, m.

Proof. Now we have all the elements to prove the main theorem of the paper. Let $\tilde{q}$ as defined in Lemma B.2, with $Z = X$. Moreover, let $v \in \mathcal{F}$, where $\mathcal{F}$ is the reproducing kernel Hilbert space associated to the kernel k_Y which is bounded by assumption. Then for any continuous f, we have that $v \circ f$ is continuous and bounded. Denote by $\Theta_{f,g}(h, v, q)$ the quantity

$$\Theta_{f,g}(h, v, q) := \int_X v(f(x)) h(x) \mu(x) dx - \int_X v(g(y)) q(y) dy.$$

Step 1. Simplifying $\Theta_{f,g}(h,v,\tilde{q})$. Since μ is supported on $\Omega \subseteq X$, by assumptions, we have

$$\int_X v(f(x))h(x)\mu(x)dx = \int_\Omega v(f(x))h(x)\mu(x)dx.$$

Moreover, by expanding the characterization of $\tilde{q}$ in Eq. (16), we have that $\tilde{q}(x) = 0$ for any $x \in \mathbb{R}^d \setminus Q^{-1}(\Omega)$, since $Q^{-1}(\Omega) \subseteq X$, we have

$$\begin{aligned}\int_X v(g(y))\tilde{q}(y)dy &= \int_{Q^{-1}(\Omega)} v(g(y))h(Q(y))\mu(Q(y))|\nabla Q(y)|dy,\\ &= \int_{Q^{-1}(\Omega)} v(f(Q(y)))h(Q(y))\mu(Q(y))|\nabla Q(y)|dy,\end{aligned}$$

where we used in the last step that $g(y) = f(Q(y))$ for any $y \in Q^{-1}(\Omega)$, which is now the domain of integration.

Step 2. Applying the Change of Variable theorem. Note that the function $\tilde{q}$ is continuous since $H^m(\mathbb{R}^d)$ is subset of continuous functions. By applying the change of variable theorem we have

$$\int_{Q^{-1}(\Omega)} v(f(Q(x)))h(Q(x))\mu(Q(x))|\nabla Q(x)|dx = \int_\Omega v(f(y))h(y)\mu(y)dy.$$

Then, by using the characterizations in Step 1, we have

$$\begin{aligned}\Theta_{f,g}(h,v,\tilde{q}) &= \int_X v(f(x))h(x)\mu(x)dx - \int_X v(g(y))\tilde{q}(y)dy\\ &= \int_\Omega v(f(x))h(x)\mu(x)dx - \int_{Q^{-1}(\Omega)} v(f(Q(y)))h(Q(y))\mu(Q(y))|\nabla Q(y)|dy\\ &= 0.\end{aligned}$$

Step 3. Bound on $D_\lambda(f,g)$. Now, denote by $\mathcal{H}$ the set $H^m(\mathbb{R}^d)$. Since $\tilde{q} \in \mathcal{H}$, and $\Delta_{f,g}(h,v,\tilde{q}) = 0$, we have that

$$\begin{aligned}D_\lambda(f,g) = \max_{\|h\|_{\mathcal{H}}\le 1}\min_{q\in\mathcal{H}}\max_{\|v\|_{\mathcal{F}}\le 1} |\Theta_{f,g}(h,v,q)|^2 + \lambda\|q\|^2_{\mathcal{H}} &\le \max_{\|h\|_{\mathcal{H}}\le 1}\max_{\|v\|_{\mathcal{F}}\le 1} |\Theta_{f,g}(h,v,\tilde{q})|^2 + \lambda\|\tilde{q}\|^2_{\mathcal{H}} &(17)\\ &= \max_{\|h\|_{\mathcal{H}}\le 1} \lambda\|\tilde{q}\|^2_{\mathcal{H}}. &(18)\end{aligned}$$

Step 4. Simplifying $\|\tilde{q}\|_{H^m(\mathbb{R}^d)}$. We bound $\|\tilde{q}\|_{H^m(\mathbb{R}^d)}$ as in Lemma B.2, where b is bounded according to Thm. B.1,

$$\|\tilde{q}\|_{H^m(\mathbb{R}^d)} \le C'\|h\|_{H^m(\mathbb{R}^d)} C_\mu C_Q,$$

where $C_\mu := \|\mu\|_{H^m(\mathbb{R}^d)}$ and

$$C_Q \;:=\; d_{\Omega,Q}^{-m-d/2}\,\|\nabla Q\|^d_{H^m(X,\mathbb{R}^d)}(1 + \|Q\|_{H^m(X,\mathbb{R}^d)} + D^m + L^m), \tag{19}$$

where $D := \operatorname{diam}(X)$, $L = \max_{x\in X}\|\nabla Q(x)\|$ and $d_{\Omega,Q} := \min[1, d_H(Q^{-1}(\Omega), Q^{-1}(X)\cap X)]$ and $d_H(A,B)$ is the Haussdorff distance between two sets A, B.

The final result is obtained by considering that we are optimizing with the constraint $\|h\|_{H^m(\mathbb{R}^d)} \le 1$. □

C. Proof of closed form of D_λ and computational results

In the next lemma, we prove some important properties useful for the characterization of D_λ in terms of F_μ, G.

Lemma C.1. *Let $X \subset \mathbb{R}^d$ be an open bounded set. The linear operators F_μ, G defined above are compact and trace class. Moreover, for any $h, q \in \mathcal{H}$ and $v \in \mathcal{F}$, we have*

$$\begin{aligned}\langle v, F_\mu h\rangle_{\mathcal{F}} &= \int_X v(f(x))h(x)\mu(x)dx,\\ \langle v, Gq\rangle_{\mathcal{F}} &= \int_X v(f(x))q(x)dx.\end{aligned}$$

Proof. Since k_Y is bounded, and f, g are measurable, then $\Phi(f(\cdot)) : X \to \mathcal{F}$ is bounded and measurable. Since $\mu \in C^\infty(\mathbb{R}^d)$ by assumption X is bounded and h is bounded and continuous since it belongs to $\mathcal{H}$ and k_X is bounded and continuous, so $J := \int_X \|\Phi(f(x))\|_\mathcal{F} \|\psi(x)\|_\mathcal{H} |\mu(x)| \mathrm{d}x < \infty$. This guarantees the existence of the following Bochner integral $F_\mu := \int_X \Phi(f(x)) \otimes \psi(x)\mu(x)\mathrm{d}x \quad \in \mathcal{F} \otimes \mathcal{H}$. In particular, denoting by $\|\cdot\|_*$ the trace norm, i.e. $\|A\|_* = \mathrm{Tr}(\sqrt{A^*A})$, and recalling that $\|u \otimes v\|_* = \mathrm{Tr}(\sqrt{(u \otimes v)^*(u \otimes v)}) = \|u\|_F \|v\|_\mathcal{H}$ for any $u \in \mathcal{U}, v \in \mathcal{V}$ and any two separable Hilbert spaces $\mathcal{U}, \mathcal{V}$, we have

$$\|F_\mu\|_* \leq \int_X \|\Phi(f(x)) \otimes \psi(x)\|_* |\mu(x)| dx = \int_X \|\Phi(f(x))\|_\mathcal{F} \|\psi(x)\|_\mathcal{H} |\mu(x)| dx =: J < \infty.$$

Then F_μ is trace class. The same reasoning hold for G, considering that $\int_X \|\Phi(f(x))\|_\mathcal{F} \|\psi(x)\|_H \mathrm{d}x < \infty$, since X is compact. □

C.1. Proof of Thm. 4.2

Proof. We have seen in Lemma C.1, that since X is a bounded set, then F_μ and G are trace class and, by the representer property

$$\langle v, F_\mu h \rangle_\mathcal{F} = \int_X v(f(x))h(x)\mu(x)dx,$$
$$\langle v, Gq \rangle_\mathcal{F} = \int_X v(f(x))q(x)dx$$

Using this result and considering the linearity of the inner product and the variational characterization of the Hilbert norm (i.e. $\|u\|_\mathcal{F} = \max_{\|v\|_\mathcal{F} \leq 1} |\langle v, u \rangle|_\mathcal{F}$ for any $u \in \mathcal{F}$), we have

$$\begin{aligned}\Delta_{f,g}(h, q) &= \max_{\|v\|_\mathcal{F} \leq 1} |\langle v, F_\mu h \rangle_\mathcal{F} - \langle v, Gq \rangle_\mathcal{F}|^2 \\ &= \max_{\|v\|_\mathcal{F} \leq 1} |\langle v, F_\mu h - Gq \rangle_\mathcal{F}|^2 \\ &= \|F_\mu h - Gq\|_\mathcal{F}^2,\end{aligned}$$

for any $h, q \in \mathcal{H}$. From which we characterize D_λ as

$$D_\lambda(f, g) = \max_{\|h\|_\mathcal{H} \leq 1} \min_{q \in \mathcal{H}} \|F_\mu h - Gq\|_\mathcal{F}^2 + \lambda \|q\|_\mathcal{H}^2 . \tag{20}$$

Now, we prove that the problem above has a characterization in terms of the operatorial norm of a given operator. First, notice that $q \mapsto \|Fh - Gq\|_\mathcal{F}^2 + \lambda \|q\|_\mathcal{H}^2$ is 2λ-strongly convex. It therefor has a unique global minimizer $q^*(h)$ which is also a critical point. This leads to $q^*(h) = ZFh$ where $Z = (GG^* + \lambda I)^{-1} G^*$. Here $GG^* + \lambda I$ is a positive linear operator, and therefor invertible.

So far, we have shown that:

$$D(f, g) = \max_{\|h\|_\mathcal{H} \leq 1} \|Fh - GZFh\|_\mathcal{F}^2 + \lambda \|ZFh\|_\mathcal{H}^2 . \tag{21}$$

Rewriting both squared norms as scalar products in $\mathcal{F}$ and $\mathcal{H}$ and using the adjoint operators, we have that: $\|Fh - GZFh\|_\mathcal{F}^2 + \lambda \|ZFh\|_\mathcal{H}^2 = \langle h, Th \rangle_\mathcal{H}$ with $T = F^*(I - Z^*G^*)(I - GZ)F + \lambda F^* Z^* ZF$. Thus, we have rewritten D as the operator norm of T:

$$D(f, g) = \max_{\|h\|_\mathcal{H} \leq 1} \langle h, Th \rangle_\mathcal{H} = \|T\|_{op}. \tag{22}$$

We can now simplify T. Recall that for any bounded operator A, $A(A^*A + \lambda I)^{-1} = (AA^* + \lambda I)^{-1} A$ and $(A + \lambda I)^{-1} A = I - \lambda(A + \lambda I)^{-1}$.

Thus, $GZ = Z^*G^* = G(G^*G + \lambda I)^{-1} G^* = (GG^* + \lambda I)^{-1} GG^* = I - \lambda(GG^* + \lambda I)^{-1}$. Similarly,

$$Z^*Z = G(G^*G + \lambda I)^{-1}(G^*G + \lambda I)^{-1} G^* = (GG^* + \lambda I)^{-1} GG^* (GG^* + \lambda I)^{-1} \tag{23}$$
$$= (I - \lambda(GG^* + \lambda I)^{-1}(GG^* + \lambda I)^{-1} = (GG^* + \lambda I)^{-1} - \lambda(GG^* + \lambda I)^{-2}. \tag{24}$$

Replacing these expression in T, we obtain:

$$T = \lambda^2 F^*(GG^* + \lambda I)^{-1})(GG^* + \lambda I)^{-1})F + \lambda F^* \left[(GG^* + \lambda I)^{-1} - \lambda(GG^* + \lambda I)^{-2}\right] F \tag{25}$$

$$= \lambda F^*(GG^* + \lambda I)^{-1} F. \tag{26}$$

Finally,

$$D_\lambda(f, g) = \|T\|_{op} = \lambda \|F^*(G^* + \lambda I)^{-1} F\|_{op} = \lambda \|(GG^* + \lambda I)^{-1/2} F\|_{op}^2. \tag{27}$$

□

C.2. Proof of Lemma 4.3

Before proceeding with the proof of Lemma 4.3, we introduce some operators, that will be useful also in the rest of the paper. We recall that for this set of results we are assuming that the kernel k_X is universal. This implies that the kernel matrix $K_{\tilde{X},\tilde{X}}$ is invertible and so $R_{\tilde{X}}$ exists and is invertible. The same holds for $R_{\tilde{Y}}$.

Definition C.2 (The operators $S, V : \mathcal{H} \to \mathbb{R}^{M_X}$ and $Z, U : \mathcal{F} \to \mathbb{R}^{M_Y}$)**.** First define $S : \mathcal{H} \to \mathbb{R}^{M_X}$ as

$$Su = (\langle \psi(\tilde{x}_1), u \rangle_{\mathcal{H}}, \dots, \langle \psi(\tilde{x}_{M_X}), u \rangle_{\mathcal{H}}) \in \mathbb{R}^{M_X}, \quad S^*\alpha = \sum_{i=1}^{M_X} \alpha_i \psi(\tilde{x}_i),$$

for all $u \in \mathcal{H}$ and $\alpha \in \mathbb{R}^{M_X}$. Analogously define $Z : \mathcal{F} \to \mathbb{R}^{M_Y}$ as

$$Zv = (\langle \Phi(\tilde{y}_1), v \rangle_{\mathcal{F}}, \dots, \langle \Phi(\tilde{y}_{M_Y}), v \rangle_{\mathcal{F}}) \in \mathbb{R}^{M_Y}, \quad Z^*\beta = \sum_{i=1}^{M_Y} \beta_i \Phi(\tilde{y}_i),$$

for all $v \in \mathcal{F}$ and $\beta \in \mathbb{R}^{M_Y}$. Moreover, define V, U as

$$V = R_{\tilde{X}}^{-\top} S, \qquad V = R_{\tilde{Y}}^{-\top} Z.$$

Remark C.3. We recall the following basic facts about the operator above, together with a short proof, when needed.

1. The range of S^* is span$(\psi(x_1), \dots, \psi(x_{M_X})$,
2. S is full rank and $SS^* = K_{\tilde{X}\tilde{X}}$,
3. $R_{\tilde{X}}^\top R_{\tilde{X}} = K_{\tilde{X}\tilde{X}}$, since $R_{\tilde{X}}$ is the upper-triangular Cholesky of $K_{\tilde{X}\tilde{X}}$.
4. $VV^* = I$, indeed, $VV^* = R_{\tilde{X}}^{-\top} SS^* R_{\tilde{X}}^{-1} = R_{\tilde{X}}^{-\top} K_{\tilde{X}\tilde{X}} R_{\tilde{X}}^{-1} = R_{\tilde{X}}^{-\top} R_{\tilde{X}}^\top R_{\tilde{X}} R_{\tilde{X}}^{-1} = I$.
5. V is a partial isometry, since $VV^* = I$ and it is full rank, since it is the product of two full rank operators.
6. $P_{\tilde{X}} = V^*V$, indeed, V^*V is a projector and the range of V^* is the range of S^* that is span$(\psi(x_1), \dots, \psi(x_{M_X})$.

For the same reasons, we have that: (a) The range of Z^* is span$(\Phi(\tilde{y}_1), \dots, \Phi(\tilde{y}_{M_Y})$ (b) Z is full rank and $ZZ^* = K_{\tilde{Y}\tilde{Y}}$ (c) $R_{\tilde{Y}}^\top R_{\tilde{Y}} = K_{\tilde{Y}\tilde{Y}}$ (d) $UU^* = I$ (e) U is a partial isometry (f) $P_{\tilde{Y}} = U^*U$.

Now we are ready to state the proof of Lemma 4.3.

Proof. First, note that since $\|A\|_{op}^2 = \|A^*A\|_{op}$ for any bounded linear operator A, we have

$$\widehat{D}_\lambda = \lambda \|P_{\tilde{X}} F_{\mu,N}^* P_{\tilde{Y}} (P_{\tilde{Y}} G_N P_{\tilde{X}} G_N^* P_{\tilde{Y}} + \lambda)^{-1} P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}}\|_{op}.$$

From Remark C.3 we recall that $P_{\tilde{X}} = V^*V$ where V is a partial isometry defined in Definition C.2. Analogously $P_{\tilde{Y}} = U^*U$ where V is a partial isometry defined in Definition C.2. Now, we have $U(U^*BU + \lambda I)^{-1}U^* = (B + \lambda I)^{-1}$ for any positive semidefinite operator $B \in \mathbb{R}^{M_Y \times M_Y}$, since since $UU^* = I$ and so $U^*UU^* = U^*$, indeed

$$\begin{aligned} U(U^*BU + \lambda I)^{-1}U^*(B + \lambda I) &= U(U^*BU + \lambda I)^{-1}U^*(B + \lambda I)UU^* \\ &= U(U^*BU + \lambda I)^{-1}(U^*BU + \lambda U^*U)U^* \\ &= U(U^*BU + \lambda I)^{-1}(U^*BU + \lambda I)U^* - \lambda U(U^*BU + \lambda I)^{-1}(I - U^*U)U^* \\ &= UU^* - \lambda U(U^*BU + \lambda I)^{-1}(U^* - U^*UU^*) = I. \end{aligned}$$

In particular, we will use now the result above. Let $C = UG_N P_{\tilde{X}} G_N^* U^*$. So,

$$\begin{aligned} \widehat{D}_\lambda &= \lambda \|V^*VF_{\mu,N}^* U^* U(U^*CU + \lambda I)^{-1}U^*UF_{\mu,N}V^*V\|_{op} \\ &= \lambda \|V^*A^*(C + \lambda I)^{-1}AV\|_{op} \\ &= \lambda \|A^*(C + \lambda I)^{-1}A\|_{op} \end{aligned}$$

where $A = UF_{\mu,N}V^*$ and we used the fact that $\|V^*TU\|_{op} = \|T\|_{op}$ for any couple of partial isometries such that $VV^* = I$ and $UU^* = I$. By applying the definition of U, V and $F_{\mu,N}$, we see that $A = \widehat{A}$ as in Eq. (10). Indeed, by expanding the definitions of U, V from Definition C.2 and denoting by $c_i \in \mathbb{R}^{M_Y}$ and $d_i \in \mathbb{R}^{M_X}$ respectively the vectors $c_i = Z\,\Phi(f(x_i)) = (k_Y(\tilde{y}_1, f(x_i)), \dots, k_Y(\tilde{y}_{M_Y}, f(x_i)))$ and $d_i = S\psi(x_i) = (k_X(\tilde{x}_1, x_i), \dots, k_X(\tilde{x}_{M_X}, x_i))$, we have

$$\begin{aligned} UF_{\mu,N}V^* &= \frac{v_X}{N}\sum_{i=1}^N (U\Phi(f(x_i))) \otimes (V\psi(x_i))\mu(x_i) \\ &= \frac{v_X}{N}\sum_{i=1}^N R_{\tilde{Y}}^{-\top}\left((Z\Phi(f(x_i))) \otimes (S\psi(x_i))\right) R_{\tilde{X}}^{-1}\mu(x_i) \\ &= \frac{v_X}{N}\sum_{i=1}^N R_{\tilde{Y}}^{-\top}\left(c_i d_i^\top\right) R_{\tilde{X}}^{-1}\mu(x_i). \end{aligned}$$

Note now, that by construction c_i is the i-th column of $K_{\tilde{Y},f}$ while d_i is the i-th row of the matrix $K_{X,\tilde{X}}$ for $i = 1, \dots, N$. Denoting by $\operatorname{diag}(\hat{\mu})$ the diagonal matrix whose i-th element of diagonal is $\mu(x_i)$, we have

$$\sum_{i=1}^N \mu(x_i)\, c_i d_i^\top = K_{\tilde{Y},f}\operatorname{diag}(\hat{\mu})\, K_{X,\tilde{X}}.$$

From which have

$$UF_{\mu,N}V^* = \frac{1}{N}R_{\tilde{Y}}^{-\top}K_{\tilde{Y},f}\operatorname{diag}(\hat{\mu})\, K_{X,\tilde{X}}R_{\tilde{X}}^{-1} = \widehat{A}.$$

To conclude, note that

$$C = UG_N P_{\tilde{X}} G_N^* U^* = (UG_N V^*)(VG_N^* U^*) = (UG_N V^*)\,(UG_N V^*)^*,$$

Analogously as we proved that $A = \widehat{A}$, we have that $UG_N V^* = \widehat{B}$, where $\widehat{B}$ is defined in Eq. (11). Then

$$\widehat{D}_\lambda = \lambda \|A^*(C + \lambda I)^{-1}A\|_{op} = \lambda \|\widehat{A}^*(\widehat{B}\widehat{B}^* + \lambda I)^{-1}\widehat{A}\|_{op}.$$

□

D. Proof of Thm. 4.4

Before proving the theorem, we need some preliminary lemmas

Lemma D.1. *Let $\delta \in (0, 1)$. Let $X \subseteq \mathbb{R}^d$ be an open bounded set with locally Lipschitz boundary. Let k be a Sobolev kernel of smoothness m, with $m > d/2$ on X and denote by $\mathcal{H}$ and $\psi : X \to \mathcal{H}$ the associated RKHS and canonical feature*

map. Let $\tilde{X} = \{\tilde{x}_1, \dots, \tilde{x}_M\} \subset X$ be M points sampled independently and uniformly at random in X. Denote by $P_{\tilde{X}}$ the projection operator whose range corresponds to span$\{\psi(\tilde{x}_1), \dots, \psi(\tilde{x}_M)\}$. There exists M_0 such that for all $M \geq M_0$, the following holds with probability at least $1 - \delta$:

$$\sup_{x \in X} \|(I - P_{\tilde{X}})\psi(x)\|_{\mathcal{H}} \leq C M^{-m/d+1/2} (\log \tfrac{C'M}{\rho})^{m/d},$$

where C, C' are constants depending only on X, m, d.

Proof. To prove this result we use the same reasoning of Theorem C.3 of Rudi and Ciliberto (2021), but applied to the Sobolev kernel. First, by applying, first Lemma C.2 of Rudi and Ciliberto (2021) we have that

$$\sup_{x \in X} \|(I - P_{\tilde{X}})\psi(x)\|_{\mathcal{H}} \leq \sup_{\|f\|_{\mathcal{H}} \leq 1} \|f - P_{\tilde{X}} f\|_{L^\infty(X)}.$$

Now, denote by η the so called *fill distance* (Narcowich et al., 2005) defined as $\eta = \sup_{x \in X} \min_{i \in 1,\dots,M} \|x - \tilde{x}_i\|$. By applying Proposition 3.2 of Narcowich et al. (2005) with $\alpha = 0, q = \infty, \tau = m_X$, we have that there exists an η_0 such that when $\eta \geq \eta_0$ then

$$\|f - P_{\tilde{X}} f\|_{L^\infty(X)} \leq C \eta^{-m+d/2} \|f\|_{\mathcal{H}}, \quad \forall\, f \in \mathcal{H},$$

where η_0 and C are constants depending only on d, X, m. To conclude, note that, by using Lemma 11 and 12 of Vacher et al. (2021),

$$\eta \leq (C_1 M^{-1} \log(C_2 M/\delta))^{1/d},$$

with probability $1 - \delta$, where C_1, C_2 depend only on X, d. The result is obtained by combining the three inequalities above and selecting M_0 as the minimum integer satisfying $(C_1 M_0^{-1} \log(C_2 M_0/\rho))^{1/d} \leq \eta_0$. □

Now we are ready to prove Thm. 4.4.

Proof of Theorem 4.4.

Proof. Let $\rho = \delta/4$.Denote by κ_X and κ_Y the constants bounding the kernel k_X, k_Y (which are Sobolev kernels of smoothness s and z, see Wendland (2004) for the explicit definition of such kernel). Note that κ_X, κ_Y are constants depending, respectively, only on s, d and on z, p. We recall here that $v_X := \mathrm{vol}(X) = \int_X dx$.

Step 1. We recall that $x_1, \dots, x_N$ are independently and uniformly distributed with uniform measure over X. Define the random variable $\zeta_i \in \mathcal{F} \otimes \mathcal{H}$ as

$$\zeta_i = v_X \Phi(f(x_i)) \otimes \psi(x_i) \mu(x_i)$$

for $i = 1, \dots, n$. Note now, that

$$F_{\mu,N} = \frac{1}{N} \sum_{i=1}^{n} \zeta_i, \quad F_\mu = \mathbb{E}\zeta_1.$$

Note moreover that

$$\|\zeta_i\| \;\leq\; v_X \kappa_X \kappa_Y \|\mu\|_{L^\infty} \;=:\; L.$$

Denote by $\langle A, B \rangle_{HS}$ the Hilbert-Schmidt inner product defined as $\langle A, B \rangle_{HS} = \mathrm{Tr}(A^* B)$. We recall that the space of $HS(\mathcal{H}, \mathcal{F})$ with finite HS norm, is a separable Hilbert space. We recall also that $\|\cdot\|_{op} \leq \|\cdot\|_{HS} \leq \|\cdot\|_*$ where the last is the trace norm and that $F_{\mu,N}$ has finite trace norm. By applying the Bernstein inequality for random vectors (see, e.g. Prop. 11 of Rudi et al. (2015) and references therein), we have that the following holds with probability $1 - \rho$

$$\|F_{\mu,N} - F_\mu\|_{HS} \leq \frac{4L}{\sqrt{N}} \log \frac{2}{\rho}. \tag{28}$$

Applying the same reasoning for G, we obtain

$$\|G_N - G\|_{HS} \leq \frac{4L'}{\sqrt{N}} \log \frac{2}{\rho}, \tag{29}$$

with probability $1 - \rho$, where $L' := \mathrm{vol}(X) \kappa_X \kappa_Y$.

Step 2. Now, recall that $P_{\tilde{X}}$ is a projection operator whose range is span$\{\psi(\tilde{x}_1), \dots, \psi(\tilde{x}_{M_X})\}$, X is bounded with Lipschitz boundary and k_X is a Sobolev kernel of smoothness m_X. By applying Lemma D.1, we have that there exists C_0 such that, when $M_X \geq C_0$, then with probability $1-\rho$

$$\sup_{x \in X} \|(I - P_{\tilde{X}})\psi(x)\|_{\mathcal{H}} \leq C_1 M_X^{-s/d} (\log \tfrac{C_2 M_X}{\rho})^{s/d+1/2}, \tag{30}$$

where C_0, C_1, C_2 depend only on X, s, d. Applying the same reasoning on $P_{\tilde{Y}}$, we have that there exists C_0' such that, when $M_Y \geq C_0'$, then with probability $1-\rho$

$$\sup_{y \in Y} \|(I - P_{\tilde{Y}})\psi(y)\|_{\mathcal{F}} \leq C_1' M_Y^{-z/p} (\log \tfrac{C_2' M_Y}{\rho})^{z/p+1/2}, \tag{31}$$

where C_0', C_1', C_2' depend only on Y, z, p.

Step 3. Now we can estimate the distance between F_μ and $P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}}$ and, analogously between G and $P_{\tilde{Y}} G_N P_{\tilde{X}}$. In particular, we can rewrite $F_\mu - P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}}$ as

$$F_\mu - P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}} \leq (F_\mu - F_{\mu,N}) + (I - P_{\tilde{Y}}) F_{\mu,N} + P_{\tilde{Y}} F_{\mu,N} (I - P_{\tilde{X}}),$$

from which, using $\|\cdot\|_{op} \leq \|\cdot\|_{HS}$, we derive

$$\|F_\mu - P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}}\|_{op} \leq \|F_\mu - F_{\mu,N}\|_{HS} + \|(I - P_{\tilde{Y}}) F_{\mu,N}\|_{op} + \|P_{\tilde{Y}}\|_{op} \|F_{\mu,N}(I - P_{\tilde{X}})\|_{op}.$$

Now, the term $\|F_\mu - F_{\mu,N}\|_{HS}$ is already studied in Eq. (28). For the second term, note that by expanding the definition of $F_{\mu,N}$ and using Eq. (30) we obtain,

$$\begin{aligned}
\|(I - P_{\tilde{Y}}) F_{\mu,N}\|_{op} &\leq \frac{v_X}{N} \sum_{i=1}^{N} \|(I - P_{\tilde{Y}})(\Phi(f(x_i)) \otimes \psi(x_i))\|_{op} |\mu(x_i)| \leq \\
&\frac{v_X}{N} \sum_{i=1}^{N} \|(I - P_{\tilde{Y}})\Phi(f(x_i))\|_{\mathcal{F}} \|\psi(x_i)\|_{\mathcal{H}} |\mu(x_i)| \\
&\leq \kappa_X v_X \|\mu\|_{L^\infty} C_1' M_Y^{-z/p} (\log \tfrac{C_2' M_Y}{\rho})^{z/p+1/2},
\end{aligned}$$

with probability $1-\rho$. Applying the same reasoning to the third term, we obtain

$$\|F_{\mu,N}(I - P_{\tilde{X}})\|_{op} \leq \kappa_Y v_X \|\mu\|_{L^\infty} C_1 M_X^{-s/d} (\log \tfrac{C_2 M_X}{\rho})^{s/d+1/2},$$

with probability $1-\rho$. Combining all the terms and considering that $\|P_{\tilde{X}}\|_{op} = 1$ since it is a projection, we have

$$\|F_\mu - P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}}\|_{op} \leq \beta \tag{32}$$

with

$$\beta := \frac{4L}{\sqrt{N}} \log \frac{2}{\rho} + \kappa_X \|\mu\|_{L^\infty} C_1' M_Y^{-z/p} (\log \tfrac{C_2' M_Y}{\rho})^{z/p+1/2} + \kappa_Y \|\mu\|_{L^\infty} C_1 M_X^{-s/d} (\log \tfrac{C_2 M_X}{\rho})^{s/d+1/2}.$$

To conclude, note that

$$\|F_\mu\|_{op} \leq \int \|\Phi(f(x))\|_{op} \|\psi(x)\|_{op} |\mu(x)| dx \leq v_X \kappa_X \kappa_Y \|\mu\|_{L^\infty(X)} = L,$$

and with the same reasoning we have $\|F_{\mu,N}\| \leq L$. Then, by considering that $\|AA^* - \hat{A}\hat{A}^*\| \leq (\|A\|_{op} + \|\hat{A}\|_{op}) \|A - \hat{A}\|_{op}$ for any bounded operators A, A^* between the same two Hilbert spaces, we have

$$\|F_\mu F_\mu^* - P_{\tilde{Y}} F_{\mu,N} P_{\tilde{X}} F_{\mu,N}^* P_{\tilde{Y}}\|_{op} \leq 2L\beta.$$

Repeating the same reasoning of the beginning of Step 3 for G and G_N we obtain

$$\|GG^* - P_{\tilde{Y}} G P_{\tilde{X}} G^* P_{\tilde{Y}}\|_{op} \leq 2L'\beta', \tag{33}$$

where

$$\beta' := \frac{4L'}{\sqrt{N}} \log \frac{2}{\rho} + \kappa_X \mathrm{vol}(X) C_1' M_Y^{-z/p} (\log \tfrac{C_2' M_Y}{\rho})^{z/p+1/2} + \kappa_Y \mathrm{vol}(X) C_1 M_X^{-s/d} (\log \tfrac{C_2 M_X}{\rho})^{s/d+1/2}.$$

Step 4. Before deriving the final result, we need an algebraic inequality between bounded operators. Let $B, \hat{B}, Q, \hat{Q}$ be bounded operators and assume that $Q, \hat{Q}$ are also symmetric and invertible. We recall that $\|A\|_{op}^2 = \|A^*A\|_{op} = \|AA^*\|_{op}$ for any bounded operator A. We have

$$\|Q^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}\hat{B}\|_{op}^2 = (\|Q^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}B\|_{op}^2) + (\|\hat{Q}^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}\hat{B}\|_{op}^2).$$

Then, using the equality $A^{-1} - B^{-1} = B^{-1}(A - B)A^{-1}$, valid for any bounded and invertible operator, we have

$$\begin{aligned}|\|Q^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}B\|_{op}^2| &= |\|B^*Q^{-1}B\|_{op} - \|B^*\hat{Q}^{-1}B\|_{op}| \\ &\leq \|B^*(Q^{-1} - \hat{Q}^{-1})B\|_{op} = \|B^*\hat{Q}^{-1}(Q - \hat{Q})Q^{-1}B\|_{op} \\ &\leq \|B^*\|_{op}\|\hat{Q}^{-1}\|_{op}\|Q - \hat{Q}\|_{op}\|Q^{-1}\|_{op}\|B\|_{op}.\end{aligned}$$

moreover, we have

$$\begin{aligned}|\|\hat{Q}^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}\hat{B}\|_{op}^2| &= |\|\hat{Q}^{-1/2}BB^*\hat{Q}^{-1/2}\|_{op}^2 - \|\hat{Q}^{-1/2}\hat{B}\hat{B}^*\hat{Q}^{-1/2}\|_{op}| \\ &\leq \|\hat{Q}^{-1/2}(BB^* - \hat{B}\hat{B}^*)\hat{Q}^{-1/2}\|_{op} \\ &\leq \|\hat{Q}^{-1/2}\|_{op}^2\|BB^* - \hat{B}\hat{B}^*\|_{op}.\end{aligned}$$

Now, noting that $\|\hat{Q}^{-1/2}\|_{op}^2 = \|\hat{Q}^{-1}\|_{op}$ and combining the inequalities above, we obtain

$$|\|Q^{-1/2}B\|_{op}^2 - \|\hat{Q}^{-1/2}\hat{B}\|_{op}^2| \leq \|B\|_{op}^2\|\hat{Q}^{-1}\|_{op}\|Q^{-1}\|_{op}\|Q - \hat{Q}\|_{op} + \|\hat{Q}^{-1}\|_{op}\|BB^* - \hat{B}\hat{B}^*\|_{op}. \tag{34}$$

Step 5. With the tools derived above we can proceed to bound $\widehat{D}_\lambda(f,g) - D_\lambda(f,g)$. First, note that, by Lemma 4.3, $\widehat{D}_\lambda(f,g)$ of Eq. (12) is equivalent to

$$\widehat{D}_\lambda(f,g) := \lambda\|(P_{\tilde{Y}}G_NP_{\tilde{X}}G_N^*P_{\tilde{Y}} + \lambda)^{-\frac{1}{2}}P_{\tilde{Y}}F_{\mu,N}P_{\tilde{X}}\|_{op}^2.$$

Then,

$$|D_\lambda(f,g) - \widehat{D}_\lambda(f,g)| \leq= \lambda|\|(GG^* + \lambda)^{-\frac{1}{2}}F_\mu\|_{op}^2 - \|(P_{\tilde{Y}}G_NP_{\tilde{X}}G_N^*P_{\tilde{Y}} + \lambda)^{-\frac{1}{2}}P_{\tilde{Y}}F_{\mu,N}P_{\tilde{X}}\|_{op}^2|.$$

By applying Eq. (34) with $Q = GG^* + \lambda$, $\hat{Q} = P_{\tilde{Y}}G_NP_{\tilde{X}}G_N^*P_{\tilde{Y}} + \lambda$, $B = F_\mu$ and $\hat{B} = P_{\tilde{Y}}F_{\mu,N}P_{\tilde{X}}$ and noting that $Q \succeq \lambda I$ and $\hat{Q} \succeq \lambda I$, then $\|Q^{-1}\|_{op} \leq \lambda^{-1}$ and $\|\hat{Q}^{-1}\|_{op} \leq \lambda^{-1}$, and that $\|F_\mu\|_{op} \leq L$, we have

$$\begin{aligned}|D_\lambda(f,g) - \widehat{D}_\lambda(f,g)| &\leq \lambda\|B\|_{op}^2\|\hat{Q}^{-1}\|_{op}\|Q^{-1}\|_{op}\|Q - \hat{Q}\|_{op} + \|\hat{Q}^{-1}\|_{op}\|BB^* - \hat{B}\hat{B}^*\|_{op} \\ &\leq L^2\lambda^{-1}\|Q - \hat{Q}\|_{op} + \|BB^* - \hat{B}\hat{B}^*\|_{op}.\end{aligned}$$

Now, since $\|Q - \hat{Q}\|_{op} = \|GG^* - P_{\tilde{Y}}G_NP_{\tilde{X}}\|_{op}$ and $\|BB^* - \hat{B}\hat{B}^*\|_{op} = \|F_\mu - P_{\tilde{Y}}F_{\mu,N}P_{\tilde{X}}\|_{op}$, which are bounded in Eqs. (32) and (33)

$$|D_\lambda(f,g) - \widehat{D}_\lambda(f,g)| \leq 2L'L^2\lambda^{-1}\beta' + 2L\beta.$$

□

E. Computing the optimal h and q

Using the finite-rank approximate described in the paper, we observe h and q using the following formulas, where $[h] = [h(x_1), \ldots, h(x_N)]$ and $[q] = [q(x_1), \ldots, q(x_N)]$, the values at grid points (discretization of the integral):

$$\begin{aligned}[h] &= K_{X\tilde{X}}R_X^{-1/2}\widetilde{h}, \\ [q] &= K_{X\tilde{X}}R_X^{-1/2}\widehat{G}^*(\widehat{G}\widehat{G}^* + \lambda I)^{-1}\widehat{F}\widetilde{h}.\end{aligned}$$

where $\widetilde{h}$ is computed as a byproduct of the computation of $\widehat{D}$ (for example, power-iteration).

F. Information for reproducing experiments

Dependencies We conduct our experiments using Pytorch version 1.9.0 (torchvision 0.10.0) and Numpy 1.20.1 (but our implementation does not require any special functions so should generalize to any recent version). We also make use of Kornia version 0.5.7 and PIL version 8.2.0 for IO.

Code Our source code is organized as follows:

- `warping` (dir): contains the necessary code for generating random warps. Code modified from Petrini et al. (2021) at https://github.com/pcsl-epfl/diffeomorphism.
- `did` (dir): implements objects necessary for the computation of DID.
- `imagenet.py`: implements as ImageFolder torchvision dataset for Imagenet (and Imagenette).
- `scenes` (dir), `perspective` (dir), `peppers.mat`: provide images taken us for the experiments (the `.mat` is a Matlab binary object which can be read thanks to `scipy`).
- `demo_XXX.py` are used for the demonstrations in Fig. 1. They are the easiest way to get familiar with DID.
- `exp_XXX.py` are used for the different experiments quantitative experiments.

Note that the path to the Imagenette dataset must be set in each experiment file or in the `imagenet.py` source file.

Imagenette We use images from the Imagenet dataset for our experiments. We use the subset called Imagenette, available at: https://github.com/fastai/imagenette.

ImageNet statistics We use the following Imagenet statistics to normalize *all* images:

$$\mu = [0.485, 0.456, 0.406]$$
$$\sigma = [0.229, 0.224, 0.225]$$

G. DID in action (supplementary experiments)

We propose a collection of experiments with DID on the `peppers` (see Fig. 5) image (with random square patches). We show the values DID takes and the shapes of the optimal h and q for a choice of λ.

Figure 5. Peppers image from Matlab software.

DID has parameters: $M_X = 100$, $M_Y = 16^3$, k_X Gaussian with $\sigma = 1/6$ and k_Y Abel with $a = 5$. We ran several experiments and chose $\lambda = 10^{-2}$ as the "best" value. Indeed, as shown in Fig. 6, the regions found by h and q for this value are coherent.

We consider a random square area f of size 150×150. We rotate, translate and scale it and denote it $g = \text{transform}(f)$. We then show f, h (over f), g and q (over g) as well as the value of $\widehat{D}_\lambda(f, g)$. Experiments in Fig. 6 can be reproduced with `appendix_peppers_match.py`.

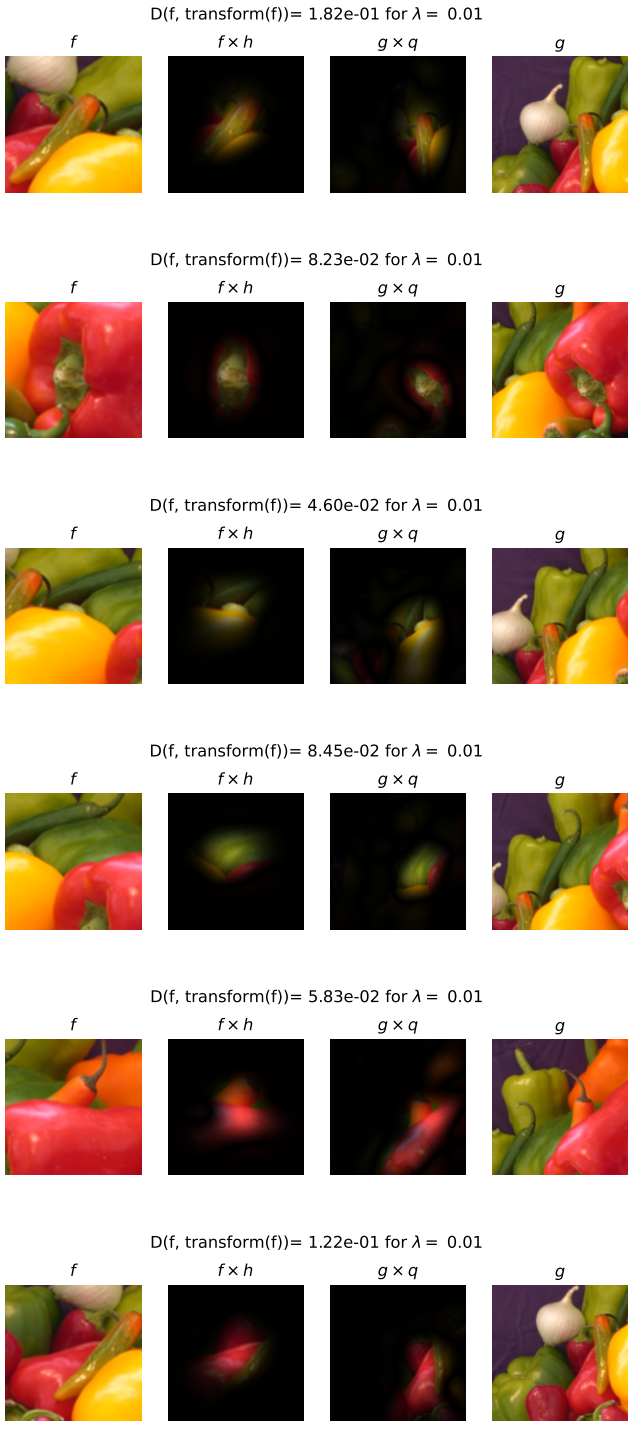
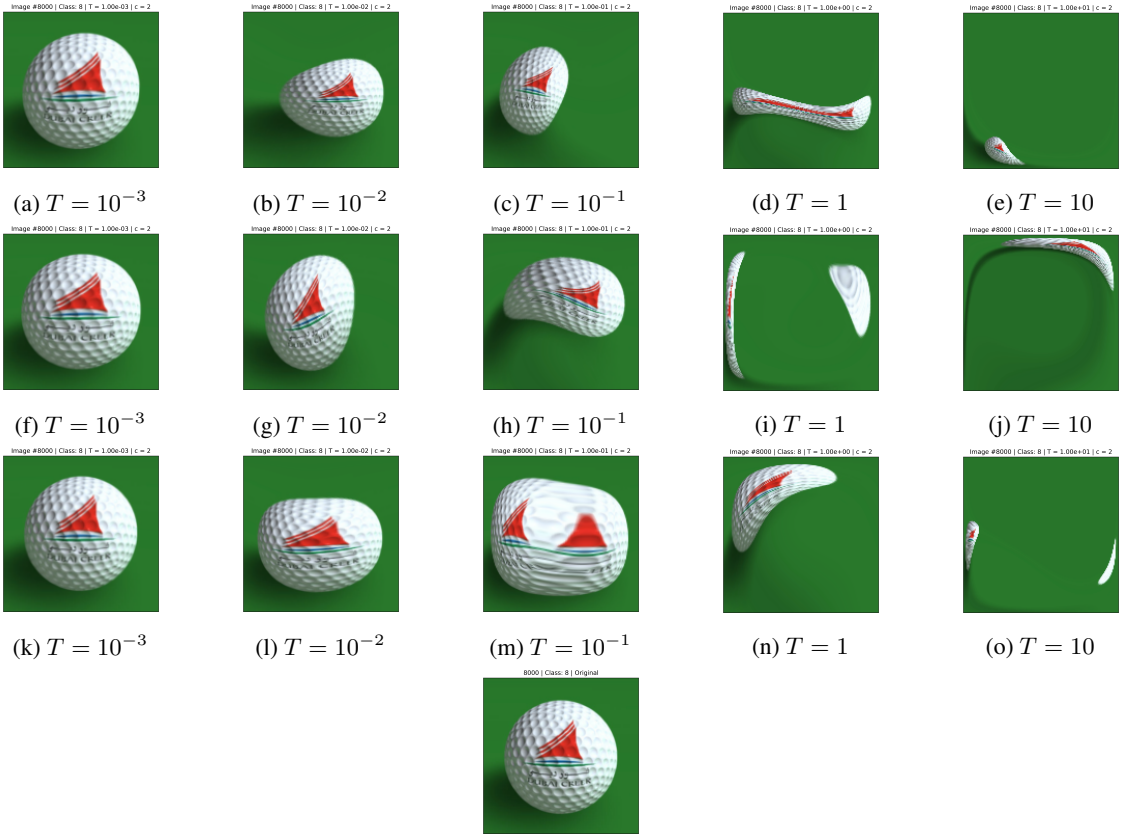


Figure 6. $\widehat{D}_\lambda(f, \text{transform}(f))$ for f random patches from `peppers`. We show f, $g = \text{transform}(f)$ as well as the optimal functions h and q.

H. Warping demonstrations

The warps below were generated using file `appendix_warp.py` in our source code, using an image from ImageNet. Rows are from different samples, columns for different warp temperatures. All waps use $c = 2$.



(a) $T = 10^{-3}$ (b) $T = 10^{-2}$ (c) $T = 10^{-1}$ (d) $T = 1$ (e) $T = 10$

(f) $T = 10^{-3}$ (g) $T = 10^{-2}$ (h) $T = 10^{-1}$ (i) $T = 1$ (j) $T = 10$

(k) $T = 10^{-3}$ (l) $T = 10^{-2}$ (m) $T = 10^{-1}$ (n) $T = 1$ (o) $T = 10$

(p) Original

Figure 7. Image #8000. Deformations warp(f, T) for different values of T ($c = 2$).