

# DO IDEAS HAVE SHAPE? PLATO'S THEORY OF FORMS AS THE CONTINUOUS LIMIT OF ARTIFICIAL NEURAL NETWORKS

HOUMAN OWHADI

**ABSTRACT.** We show that ResNets converge, in the infinite depth limit, to a generalization of image registration algorithms. In this generalization, images are replaced by abstractions (ideas) living in high dimensional RKHS spaces, and material points are replaced by data points. Whereas computational anatomy aligns images via deformations of the material space, this generalization aligns ideas by via transformations of their RKHS. This identification of ResNets as idea registration algorithms has several remarkable consequences. The search for good architectures can be reduced to that of good kernels, and we show that the composition of idea registration blocks with reduced equivariant multi-channel kernels (introduced here) recovers and generalizes CNNs to arbitrary spaces and groups of transformations. Minimizers of  $L_2$  regularized ResNets satisfy a discrete least action principle implying the near preservation of the norm of weights and biases across layers. The parameters of trained ResNets can be identified as solutions of an autonomous Hamiltonian system defined by the activation function and the architecture of the ANN. Momenta variables provide a sparse representation of the parameters of a ResNet. The registration regularization strategy provides a provably robust alternative to Dropout for ANNs. Pointwise RKHS error estimates lead to deterministic error estimates for ANNs.

## 1. Introduction

The purpose of this paper is to show that residual neural networks [34] are essentially discretized solvers for a generalization of image registration/computational anatomy variational problems. This identification allows us to initiate a theoretical understanding of deep learning from the perspective of shape analysis with images replaced by high dimensional RKHS spaces. This introduction is an articulated overview of what was learned through this work.

**1.1. The setting.** We employ the setting of supervised learning, which can be expressed as solving the following problem.

**Problem 1.** *Let  $f^\dagger$  be an unknown continuous function mapping  $\mathcal{X}$  to  $\mathcal{Y}$ . Given the information<sup>1</sup>  $f^\dagger(X) = Y$  with the data  $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$  approximate  $f^\dagger$ .*

---

*Date:* November 3, 2020.

Caltech, MC 9-94, Pasadena, CA 91125, USA, owjadi@caltech.edu.

<sup>1</sup>For a  $N$ -vector  $X = (X_1, \dots, X_N) \in \mathcal{X}^N$  and a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , write  $f(X)$  for the  $N$  vector with entries  $(f(X_1), \dots, f(X_N))$  (we will keep using this generic notation).

Given  $\lambda > 0$  and a kernel  $K$  defining an RKHS of functions mapping  $\mathcal{X}$  to  $\mathcal{Y}$ , the ridge regression solution to Problem 1 is to approximate  $f^\dagger$  with the minimizer of<sup>2</sup>

$$\min_f \lambda \|f\|_K^2 + \|f(X) - Y\|_{\mathcal{Y}^N}^2. \quad (1.1)$$

**1.2. Mechanical regression and ANNs.** Given another kernel  $\Gamma$  defining an RKHS of functions mapping  $\mathcal{X}$  to  $\mathcal{X}$ , consider the variant in which  $f^\dagger$  is approximated by

$$f^\ddagger = f \circ \phi_L \text{ where } \phi_L = (I + v_L) \circ \cdots \circ (I + v_1) \quad (1.2)$$

(write  $I$  for the identity map) is a deformation of the input space obtained from the composition of  $L$  displacements  $v_s : \mathcal{X} \rightarrow \mathcal{X}$ , and  $(v_1, \dots, v_L, f)$  is a minimizer of

$$\min_{f, v_1, \dots, v_L} \frac{\nu L}{2} \sum_{s=1}^L \|v_s\|_\Gamma^2 + \lambda \|f\|_K^2 + \|f \circ \phi_L(X) - Y\|_{\mathcal{Y}^N}^2, \quad (1.3)$$

where  $\nu$  is a positive parameter. Using the setting (Sec. 2) of operator-valued kernels [2], we show (Subsec. 6.3) that if  $\Gamma(x, x') = \varphi^T(x)\varphi(x')I_{\mathcal{X}}$  and  $K(x, x') = \varphi^T(x)\varphi(x')I_{\mathcal{Y}}$  where  $I_{\mathcal{X}}$  ( $I_{\mathcal{Y}}$ ) is the identity operator on  $\mathcal{X}$  ( $\mathcal{Y}$ ) and  $\varphi : \mathcal{X} \rightarrow \mathcal{X} \oplus \mathbb{R}$  is a nonlinear map  $\varphi(x) = (\mathbf{a}(x), 1)$  defined by an activation function  $\mathbf{a} : \mathcal{X} \rightarrow \mathcal{X}$  (e.g., an elementwise nonlinearity) then minimizers of (1.3) are of the form  $f(x) = \tilde{w}\varphi(x)$  and  $v_s(x) = w_s\varphi(x)$  where<sup>3</sup>  $\tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})$  and the  $w_s \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})$  are minimizers of

$$\min_{\tilde{w}, w_1, \dots, w_L} \frac{\nu L}{2} \sum_{s=1}^L \|w_s\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \lambda \|\tilde{w}\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})}^2 + \|f \circ \phi_L(X) - Y\|_{\mathcal{Y}^N}^2, \quad (1.4)$$

with

$$f \circ \phi_L(x) = (\tilde{w}\varphi) \circ (I + w_L\varphi) \circ \cdots \circ (I + w_1\varphi). \quad (1.5)$$

(1.5) has the structure of one ResNet block [34] and minimizing (1.4) is equivalent to training the network with  $L_2$  regularization on weights and biases<sup>4</sup>. Composing (1.5) over a hierarchy of spaces (layered in between  $\mathcal{X}$  and  $\mathcal{Y}$ , as described in Sec. 5 and 7) produces input-output functions that have the functional form of artificial neural networks (ANNs) [44] and ResNets. If  $K$  and  $\Gamma$  are reduced equivariant multichannel (REM) kernels (introduced in Sec. 9) then the input-output functions obtained by composition blocks of the form (1.5) are convolutional neural networks (CNNs) [45] and their generalization.

**1.3. Idea registration and the continuous limit of ANNs.** We show (Subsec. 3.9) that, in the limit  $L \rightarrow \infty$ , the adherence values (accumulation points) of the minimizers (1.2) of (1.3) are of the form  $f \circ \phi^v(X, 1)$  where  $(v, f)$  are minimizers of

$$\min_{v, f} \frac{\nu}{2} \int_0^1 \|v(\cdot, t)\|_\Gamma^2 dt + \lambda \|f\|_K^2 + \|f \circ \phi^v(X, 1) - Y\|_{\mathcal{Y}^N}^2 \quad (1.6)$$

<sup>2</sup>Write  $\|\cdot\|_K$  for the reproducing kernel Hilbert space (RKHS) norm defined by  $K$  and  $\|f(X) - Y\|_{\mathcal{Y}^N}^2 := \sum_{i=1}^N \|f(X_i) - Y_i\|_{\mathcal{Y}}^2$  where  $\|\cdot\|_{\mathcal{Y}}$  is a quadratic norm on  $\mathcal{Y}$ .

<sup>3</sup>Write  $\mathcal{L}(\mathcal{F}, \mathcal{X})$  for the set of linear maps from  $\mathcal{F}$  to  $\mathcal{X}$  and  $\|w_s\|_{\mathcal{L}(\mathcal{F}, \mathcal{X})}$  for the Frobenius norm of  $w_s$ .

<sup>4</sup>Writing  $\varphi(x) = (\mathbf{a}(x), 1)$  has the same effect as using a bias neuron (an always active neuron), therefore  $\tilde{w}$  and the  $w_s$  incorporate both weights and biases.

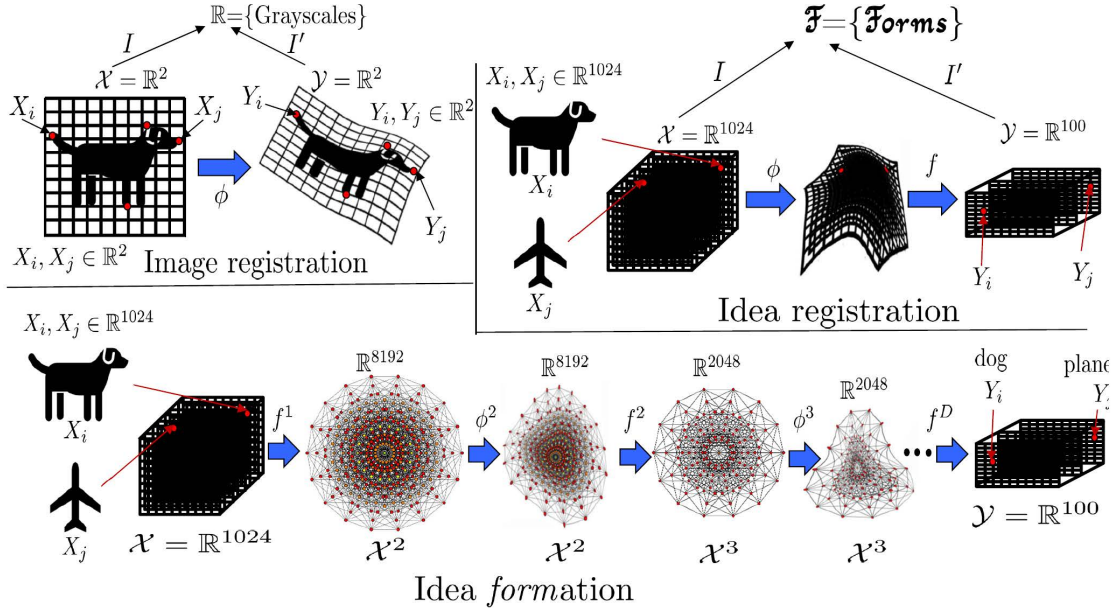


FIGURE 1. Image registration. Idea registration. Idea formation.

and  $\phi^v(x, t)$  is the flow map of  $v$  defined as the solution of

$$\begin{cases} \dot{\phi}(x, t) = v(\phi(x, t), t) & \text{for } (x, t) \in \mathcal{X} \times [0, 1] \\ \phi(x, 0) = x & \text{for } x \in \mathcal{X}. \end{cases} \quad (1.7)$$

(1.6) has the structure of variational formulations used in computational anatomy [26], image registration [12] and shape analysis [96]. Recall that the core idea of image registration is to represent the image of an anatomical structure as a function  $I$  mapping material points in  $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$  to intensities in  $\mathbb{R}_+$  (see Fig. 1). The distance between an image  $I$  and a template  $I'$  is then defined by minimizing

$$\min_v \nu \int_0^1 \|\Delta v(\cdot, t)\|_{L^2([0,1]^2)}^2 dt + \|I(\phi^v(\cdot, 1)) - I'\|_{L^2([0,1]^2)}, \quad (1.8)$$

over diffeomorphisms  $\phi^v$  of  $\mathbb{R}^2$  driven by the vector field  $v$  ( $\dot{\phi}^v = v(\phi, t)$ ) such that  $\phi^v(x, 0) = x$  [95, 88]. The regularizer  $\|\Delta v\|_{L^2}$  can be replaced by higher order Sobolev norms [24] or the  $L^2$  norm of differential operators adapted to the underlying problem [55]. *Landmark matching* [40] simplifies the loss (1.8) to

$$\min_v \nu \int_0^1 \|\Delta v\|_{L^2([0,1]^2)}^2 dt + \sum_i |\phi^v(X_i, 1) - Y_i|^2, \quad (1.9)$$

where the  $X_i$  and  $Y_i$  are a finite number of landmark/control (material) points on the two images  $I$  and  $I'$  (e.g., in Fig. 1,  $X_i$  is the tip of the tail of the first dog and  $Y_i$  is the tip of the tail of the second dog). The variational problem (1.6) looks like the image registration with landmark matching variational problem (1.9) with a few differences. The matching material/landmark points  $(X_i, Y_i) \in \mathbb{R}^2 \times \mathbb{R}^2$  are replaced by matching

data points  $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ . The deformation  $\phi$  is not acting on  $\mathbb{R}^2$  but on  $\mathcal{X}$ , which could be high dimensional. The images  $I : \mathbb{R}^2 \rightarrow \mathbb{R}_+$  and  $I' : \mathbb{R}^2 \rightarrow \mathbb{R}_+$  are replaced (see Fig. 1) by functions  $I : \mathcal{X} \rightarrow \mathcal{F}$  and  $I' : \mathcal{Y} \rightarrow \mathcal{F}$ , which we will call ideas<sup>5</sup>. The space of grayscale intensities  $\mathbb{R}_+$  is replaced by an abstract space  $\mathcal{F}$ , which we will call *space of forms* in reference to Plato's theory of forms<sup>6</sup> [71]. Since the spaces  $\mathcal{X}$  and  $\mathcal{Y}$  may be distinct, (1.6) composes the deformation  $\phi^v(\cdot, 1) : \mathcal{X} \rightarrow \mathcal{X}$  with the map  $f : \mathcal{X} \rightarrow \mathcal{Y}$  to align the ideas  $I : \mathcal{X} \rightarrow \mathcal{F}$  and  $I' : \mathcal{Y} \rightarrow \mathcal{F}$ . In that sense, (1.6) (which we call idea registration) compares ideas by creating alignments via deformations/transformations of RKHS spaces<sup>7</sup>. Since (1.4) is a particular case of (1.3), the convergence of (1.3) towards (1.6) implies that ANNs, ResNets and CNNs are discretized image/idea registration algorithms (they converge towards (1.6) in the continuous/infinite depth limit) with material/landmark points replaced by data points, and images replaced by functions in high dimensional spaces defining high dimensional shapes<sup>8</sup>. The kernel representation of ResNet blocks as (1.3) and the identification of ANNs as discretized idea registration problems have several remarkable consequences, which we will now highlight to map the content of this paper.

#### 1.4. Least action principle, Hamiltonian dynamic and energy preservation.

Minimizers of (1.3) have the representation<sup>9</sup> (Sec. 3)

$$v_s(\cdot) = \Gamma(\cdot, q^s) \Gamma(q^s, q^s)^{-1} (q^{s+1} - q^s) \quad (1.10)$$

(Thm. 3.3) where the position variables  $q^s$  are in  $\mathcal{X}^N$  started from  $q^1 = X$  and minimizing a discrete least action principle of the form ( $\Delta t := 1/L$ )

$$\min_{f, q^2, \dots, q^{L+1}} \frac{\nu}{2} \sum_{s=1}^L \left( \frac{q^{s+1} - q^s}{\Delta t} \right)^T \Gamma(q^s, q^s)^{-1} \left( \frac{q^{s+1} - q^s}{\Delta t} \right) + \lambda \|f\|_K^2 + \|f(q^{L+1}) - Y\|_{\mathcal{Y}^N}^2. \quad (1.11)$$

Therefore, introducing the momentum variables

$$p^s = \Gamma(q^s, q^s)^{-1} \frac{q^{s+1} - q^s}{\Delta t}, \quad (1.12)$$

$(q^s, p^s)$  follows the discrete Hamiltonian dynamics

$$\begin{cases} q^{s+1} &= q^s + \Delta t \Gamma(q^s, q^s) p^s \\ p^{s+1} &= p^s - \frac{\Delta t}{2} \partial_{q^{s+1}} \left( (p^{s+1})^T \Gamma(q^{s+1}, q^{s+1}) p^{s+1} \right), \end{cases} \quad (1.13)$$

and the near energy preservation of variational integrators [49, 31] implies (Thm. 3.10) that the norms  $\|w_s\|_{\mathcal{L}(\mathcal{F}, \mathcal{X})}^2$  (of weights and biases of ResNet blocks after training with  $L_2$  regularization) are nearly constant (fluctuate by at most  $\mathcal{O}(1/L)$ ) across  $i \in \{1, \dots, L\}$ .

<sup>5</sup>The etymology of ‘‘idea’’ is (<https://www.etymonline.com/word/idea>) ‘‘mental image or picture’’... from Greek *idea* ‘‘form’’... In Platonic philosophy, ‘‘an archetype, or pure immaterial pattern, of which the individual objects in any one natural class are but the imperfect copies.’’

<sup>6</sup>According to Plato's theory of forms the reason why we know that a particular dog is a dog is because there exists an ideal form (an universal intelligible archetype known as a dog) and the particular dog is a shadow (as in Plato's cave) or an imperfect copy/projection of that ideal form.

<sup>7</sup>Credit to <https://en.wikipedia.org/wiki/User:Tomruen> for the  $N$ -cube images in Fig. 1.

<sup>8</sup>Plato introduced the intriguing notion that ideas have an actual shape [71].

<sup>9</sup>Write  $\Gamma(q^s, q^s)$  for the  $N \times N$  block matrix with blocks  $\Gamma(q_i^s, q_j^s)$ , and  $\Gamma(\cdot, q^s)$  for the  $1 \times N$  block vector with blocks  $\Gamma(\cdot, q_i^s)$ .

Similarly, minimizers of (1.6) have, as in landmark matching [40], the representation (Thm. 3.8)

$$\dot{\phi}^v(x, t) = \Gamma(\phi^v(x, t), q)p, \quad (1.14)$$

where the position and momentum variables  $(q, p)$  are in  $\mathcal{X}^N \times \mathcal{X}^N$ , started from  $q(0) = X$ , and following the dynamic defined by the Hamiltonian (Thm. 3.4)

$$\mathfrak{H}(q, p) = \frac{1}{2}p^T \Gamma(q, q)p. \quad (1.15)$$

Therefore (Thm. 3.9), the norm  $\|v(\cdot, t)\|_{\Gamma}^2$  (of the weights and biases in the continuous infinite depth limit) must be a constant over  $t \in [0, 1]$ . Furthermore (1.13) is a first-order variational/symplectic integrator for approximating the Hamiltonian flow of (1.15). Momentum variables reflect the contribution of each data point to the regressor of  $f^\dagger$  (Sec. 3.11), in particular, as with support vector machines [83], if the squared loss  $\|\phi^v(X, 1) - Y\|_{\mathcal{Y}^N}^2$  is replaced by a hinge loss in (1.6), then the only points  $X_i$  with non zero momentum are those for which  $\phi^v(X_i, 1)$  is included in the (hinge loss) margin. Therefore, as in image registration [13, 89], the momentum map representation of minimizers is often sparse.

**1.5. Mean field dynamic.** Let  $\psi : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{X}, \mathcal{F})$  and  $\mathcal{F}$  be a feature map and space (Subsec. 2.2) for  $\Gamma$  ( $\Gamma(x, x') = \psi^T(x)\psi(x')$ ). Rescaling momentum variables as  $p_j = \frac{1}{N}\bar{p}_j$ , the Hamiltonian flow of (1.15) can be written (Subsec. 3.6.1 and 3.12)

$$\begin{cases} \dot{q}_i = \psi^T(q_i)\alpha \\ \dot{p}_i = -\partial_x(\bar{p}_i^T \psi^T(x)\alpha)\Big|_{x=q_i}, \end{cases} \text{ with } \alpha = \frac{1}{N} \sum_{j=1}^N \psi(q_j)\bar{p}_j. \quad (1.16)$$

$\alpha : [0, 1] \rightarrow \mathcal{F}$  is a norm preserving trajectory in feature space (such that  $t \rightarrow \|\alpha(t)\|_{\mathcal{F}}$  is constant) amenable to mean field analysis (Subsec. 3.12) and (1.14) is equivalent to

$$\dot{\phi}^v(x, t) = \psi^T(\phi^v(x, t))\alpha(t). \quad (1.17)$$

**1.6. Existence, uniqueness, and convergence of minimizers.** Minimizers of (1.3) (and therefore (1.4)) exist, and, although they may not be unique (Sec. 3.8), they are unique up the value of the initial momentum  $p^1$  in (1.12) (Thm. 3.10) which suggests that ResNets could also be trained with geodesic shooting (Sec. 3.6.3, 3.13 and 5.3) as done in image registration [1]. Minimal values of (1.3) converge (Thm. 3.11 and Cor. 3.12) to those of (1.6) and the adherence values (as  $L \rightarrow \infty$ ) of the minimizers of (1.3) are the minimizers of (1.6).

**1.7. Brittleness of ANNs.** Minimal values and minimizers of (1.3) and (1.6) may not be continuous in the data  $X$ . Furthermore, minimizing (1.6) is equivalent (Subsec. 3.10) to approximating  $f^\dagger$  with the (ridge regression) minimizer of (1.1) with the kernel  $K(x, x')$  replaced by the learned kernel  $K^v := K(\phi^v(x, 1), \phi^v(x', 1))$  (Prop. 3.13). Therefore minimizing (1.1) is equivalent (Sec. 8) to estimating  $f^\dagger(x)$  with

$$\mathbb{E}_{\xi \sim \mathcal{N}(0, K^v)}[\xi(x) \mid \xi(X) = Y] = \mathbb{E}_{\xi \sim \mathcal{N}(0, K)}[\xi(\phi^v(x, 1)) \mid \xi(\phi^v(X, 1)) = Y], \quad (1.18)$$

where  $\mathcal{N}(0, K^v)$  is the centered Gaussian process prior with covariance function  $K^v$ . This observation that ANNs can be interpreted as performing ridge regression with a data-dependent prior suggests Bayesian brittleness (the extreme lack of robustness of

Bayesian posterior values with respect to the prior [66, 65, 61]) as a cause for the high sensitivity of ANNs with respect to the testing data  $x$  or the training data  $X$  reported in [85] (this lack of stability was predicted in [50] based on [66]). This fragility endures even if the training data  $X$  is randomized [62] and may not be resolved without loss of accuracy since robustness and accuracy/consistency are conflicting requirements [65, 62]. The Hamiltonian representation (1.10) of minimizers of ANNs suggests Hamiltonian chaos [15] as another cause of the instability of ANNs (from this dynamical perspective, the instability of ANNs is related to curvature fluctuations of the metric defined by  $\Gamma(q, q)$  [15]) and that Lyapunov characteristic exponents could also be used a measure of instability for ANNs.

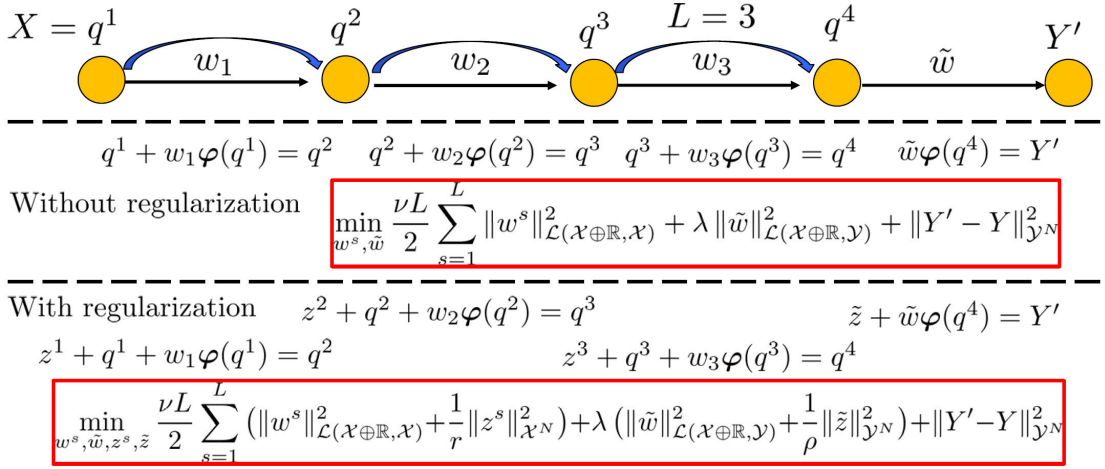


FIGURE 2. Training of one ResNet block with the proposed regularization.

**1.8. Regularization.** To ensure continuity, (1.3) and (1.6) must be regularized and generalize (Sec. 4, 6 and 9.7) an image registration regularization strategy [53] to idea registration. When performed with activation functions (Fig. 2), the proposed regularization provides a principled alternative to Dropout for ANNs [82]. In the setting of one ResNet block, this regularization does not change the functional form (1.5) of the block but replaces (Thm. 6.9) the training (1.4) of the weights and biases by the minimization of

$$\begin{aligned}
 \min_{w^s, \tilde{w}, q^s, Y'} \frac{\nu L}{2} \sum_{s=1}^L (\|w^s\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \frac{1}{r} \|q^{s+1} - q^s - w^s\varphi(q^s)\|_{\mathcal{X}^N}^2) \\
 + \lambda (\|\tilde{w}\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})}^2 + \frac{1}{\rho} \|\tilde{w}\varphi(q^{L+1}) - Y'\|_{\mathcal{Y}^N}^2) + \|Y' - Y\|_{\mathcal{Y}^N}^2,
 \end{aligned} \tag{1.19}$$

where  $\rho, r > 0$  are regularization parameters relaxing the constraint that the input data  $X$  must propagate without error through each layer of the network (the trajectory of  $q^s$  is started from  $q^1 = X$ ). Fig. 2 shows the training of one ResNet block with the



proposed regularization. Note that training with regularization is equivalent to replacing the exact propagation  $q^{s+1} = q^s + w^s \varphi(q^s)$  (and  $Y' = \tilde{w} \varphi(q^{L+1})$ ) of the input data by  $q^{s+1} = q^s + w^s \varphi(q^s) + z^s$  ( $Y' = \tilde{w} \varphi(q^{L+1}) + \tilde{z}$ ) where the  $z^s$  and  $\tilde{z}$  are propagation error variables ( $z^s \in \mathcal{X}^N$ ,  $\tilde{z} \in \mathcal{Y}^N$ ) whose norms are added to the total loss at the training stage (the  $z^s$  and  $\tilde{z}$  are set to zero at the testing stage). These slack variables have, as in Tikhonov regularization, a natural interpretation as Gaussian noise added to the output of each layer. In particular  $r$  and  $\rho$  play the same role as  $\lambda$  in the ridge regression (1.1) (they can be interpreted as variances of propagation errors) and (1.19) converges to (1.4) as  $r, \rho \downarrow 0$ . While reducing overfitting by training with noise seems to have been exhaustively explored (variants include adding noise to the input data [36], to the weights and biases [3] and to the activation functions [27]), the proposed approach seems to be distinct in the sense that the  $z^s$  and  $\tilde{z}$  are not random noise but deterministic variables to be trained alongside the weights and biases of the network. Furthermore, since the proposed strategy is equivalent to adding the nuggets  $r$  and  $\rho$  to the kernels  $\Gamma$  and  $K$  in (1.3) ( $\Gamma \rightarrow \Gamma + rI$  where  $I$  is the identity operator), it appears to be the natural generalization of the regularization strategy employed in spatial statistics.

**1.9. Idea formation.** Composing idea registration blocks (Fig. 1) produces input/output functions that have the exact functional structure of ANNs and enable their generalization to ANNs of continuous depth and acting on continuous (e.g., functional) spaces (Sec. 5 and 7). In doing so we (1) prove the existence of minimizers for  $L_2$  regularised ANNs/ResNets/CNNs (Thm. 5.1, 7.1 and 9.8) (2) characterize these minimizers as autonomous solutions of discrete Hamiltonian systems with discrete least action principles (3) derive the near-preservation of the norm of weights and biases in ResNet blocks (4) obtain their uniqueness given initial momenta (5) prove their convergence (in the sense of adherence values) in the infinite depth limit towards nested idea registration (idea formation<sup>10</sup>) characterized by continuous deformations flows in high dimensional RKHS spaces (6) deduce that training  $L^2$ -regularized ANNs could in principle be reduced to the determination of the weights and biases of the first layer (Subsec. 5.3).

**1.10. Reduced equivariant multichannel (REM) kernels.** The identification of ANNs as discretized idea formation flow maps implies that the search for good architectures for ANNs can be reduced to the search for good kernels for idea registration/formation. We introduce (in Sec. 9) reduced equivariant multichannel (REM) kernels (the equivariant component is a variant of [74]) and show that CNNs (and their ResNet variants) are particular instances of idea formation with REM kernels. REM kernels (1) enable the generalization of CNNs to arbitrary groups of transformations acting on arbitrary spaces, (2) preserve the relative pose information (see Rmk. 9.6) across layers.

**1.11. Deep learning without backpropagation.** Approximating  $f^\dagger$  with (1.2) is equivalent to performing ridge regression with the kernel  $K(\phi^L(x), \phi^L(x'))$  (Subsec. 3.10) which is also the strategy employed by the non parametric version of Kernel Flows (KF) [68] (Subsec. 10.6). Whereas ANNs are trained via backpropagation, KF is based on

<sup>10</sup>One definition of “formation” is (<https://www.merriam-webster.com/dictionary/form>) to give a particular shape, its etymology (<https://en.wiktionary.org/wiki/forma>) is borrowed from Latin *forma*, perhaps from Ancient Greek *μορφή* (*morphé*, “shape, figure”).

a cross-validation principle [20, 32] that enables its training without backpropagation (and that has been shown to be consistent in the parametric setting [20]). This suggests that deep learning could be performed by replacing backpropagation with forward cross-validation<sup>11</sup>.

**1.12. Error estimates and deep residual Gaussian processes.** Scalar-valued Gaussian processes have a natural extension to function-valued Gaussian processes (Sec. 8). This extension leads to deterministic (Cor. 8.10) and probabilistic (Subsec. 8.3) error estimates for idea registration. Minimizers of (1.6) (and its regularized variant (1.19)) have natural interpretations as MAP estimators of Brownian flows of diffeomorphisms [7, 42], which we will extend as deep residual Gaussian processes (that can be interpreted as a continuous variant of deep Gaussian processes [23]).

## 2. Operator-valued kernels

Through this manuscript we employ (with slight variations) the setting of operator-valued kernels introduced in [41] (as a generalization of vector valued kernels [2]).

**2.1. A short reminder.** Let  $\mathcal{X}$  and  $\mathcal{Y}$  be separable Hilbert spaces<sup>12</sup> endowed with the inner products  $\langle \cdot, \cdot \rangle_{\mathcal{X}}$  and  $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ . Write  $\mathcal{L}(\mathcal{Y})$  for the set of bounded linear operators mapping  $\mathcal{Y}$  to  $\mathcal{Y}$ .

**Definition 2.1.** We call  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  an **operator-valued kernel** if

- (1)  $K$  is Hermitian, i.e.

$$K(x, x') = K(x', x)^T \text{ for } x, x' \in \mathcal{X}, \quad (2.1)$$

writing  $A^T$  for the adjoint of the operator  $A$  with respect to  $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ , and

- (2) non-negative, i.e.

$$\sum_{i,j=1}^m \langle y_i, K(x_i, x_j)y_j \rangle_{\mathcal{Y}} \geq 0 \text{ for } (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, m \in \mathbb{N}. \quad (2.2)$$

We call  $K$  non-degenerate if  $\sum_{i,j=1}^m \langle y_i, K(x_i, x_j)y_j \rangle_{\mathcal{Y}} = 0$  implies  $y_i = 0$  for all  $i$  whenever  $x_i \neq x_j$  for  $i \neq j$ .

The following definition provides a simple example of operator-valued kernels obtained from scalar-valued kernels.

**Definition 2.2.** We say that  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  is **scalar** if  $K(x, x') = k(x, x')I_{\mathcal{Y}}$  (writing  $I_{\mathcal{Y}}$  for the identity operator on  $\mathcal{Y}$ ) for some scalar-valued kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , i.e.

$$\langle y, K(x, x')y' \rangle_{\mathcal{Y}} = k(x, x')\langle y, y' \rangle_{\mathcal{Y}} \text{ for } x, x' \in \mathcal{X} \text{ and } y, y' \in \mathcal{Y}. \quad (2.3)$$

<sup>11</sup>See [94] for how KF/cross-validation could also be applied to the direct training of the inner layers of ANNs.

<sup>12</sup>Although  $\mathcal{X}$  and  $\mathcal{Y}$  are finite-dimensional in all practical applications, and although we will restrict some of our proofs to the finite-dimensional setting to minimize technicalities, as demonstrated in [58], it is useful to keep the infinite-dimensional viewpoint in the identification of discrete models with desirable attributes inherited from the infinite-dimensional setting.



**Example 2.3.**  $\mathcal{X} = \mathbb{R}^{d_x}$ ,  $\langle x, x' \rangle_{\mathcal{X}} = x^T x'$ ,  $\mathcal{Y} = \mathbb{R}^{d_y}$  and  $\langle y, y' \rangle_{\mathcal{Y}} = y^T y'$  are prototypical examples. For ease of presentation, we will continue using the notation  $\langle y, y' \rangle_{\mathcal{Y}} = y^T y'$  even when  $\mathcal{Y}$  is arbitrary.

Each non-degenerate, locally bounded and separately continuous operator-valued kernel  $K$  (which we will refer to as a Mercer's kernel) is in one to one correspondence with a reproducing kernel Hilbert space  $\mathcal{H}$  of continuous functions  $f : \mathcal{X} \rightarrow \mathcal{Y}$  obtained [41, Thm. 1,2] as the closure of the linear span of functions  $z \rightarrow K(z, x)y$  ( $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ) with respect to the inner product identified by the reproducing property

$$\langle f, K(\cdot, x)y \rangle_{\mathcal{H}} = \langle f(x), y \rangle_{\mathcal{Y}} \quad (2.4)$$

**2.2. Feature maps.** Let  $\mathcal{F}$  be a separable Hilbert space (with inner product  $\langle \cdot, \cdot \rangle_{\mathcal{F}}$  and norm  $\|\cdot\|_{\mathcal{F}}$ ) and let  $\psi : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y}, \mathcal{F})$  be a continuous function mapping  $\mathcal{X}$  to the space of bounded linear operators from  $\mathcal{Y}$  to  $\mathcal{F}$ .

**Definition 2.4.** We say that  $\mathcal{F}$  and  $\psi : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y}, \mathcal{F})$  are a feature space and a feature map for the kernel  $K$  if, for all  $(x, x', y, y') \in \mathcal{X}^2 \times \mathcal{Y}^2$ ,

$$y^T K(x, x')y' = \langle \psi(x)y, \psi(x')y' \rangle_{\mathcal{F}}. \quad (2.5)$$

Write  $\psi^T(x)$ , for the adjoint of  $\psi(x)$  defined as the linear function mapping  $\mathcal{F}$  to  $\mathcal{Y}$  satisfying

$$\langle \psi(x)y, \alpha \rangle_{\mathcal{F}} = \langle y, \psi^T(x)\alpha \rangle_{\mathcal{Y}} \quad (2.6)$$

for  $x, y, \alpha \in \mathcal{X} \times \mathcal{Y} \times \mathcal{F}$ . Note that  $\psi^T : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{F}, \mathcal{Y})$  is therefore a function mapping  $\mathcal{X}$  to the space of bounded linear functions from  $\mathcal{F}$  to  $\mathcal{Y}$ . Writing  $\alpha^T \alpha' := \langle \alpha, \alpha' \rangle_{\mathcal{F}}$  for the inner product in  $\mathcal{F}$  we can ease our notations by writing

$$K(x, x') = \psi^T(x)\psi(x') \quad (2.7)$$

which is consistent with the finite-dimensional setting and  $y^T K(x, x')y' = (\psi(x)y)^T (\psi(x')y')$  (writing  $y^T y'$  for the inner product in  $\mathcal{Y}$ ). For  $\alpha \in \mathcal{F}$  write  $\psi^T \alpha$  for the function  $\mathcal{X} \rightarrow \mathcal{Y}$  mapping  $x \in \mathcal{X}$  to the element  $y \in \mathcal{Y}$  such that

$$\langle y', y \rangle_{\mathcal{Y}} = \langle y', \psi^T(x)\alpha \rangle_{\mathcal{Y}} = \langle \psi(x)y', \alpha \rangle_{\mathcal{F}} \text{ for all } y' \in \mathcal{Y}. \quad (2.8)$$

We can, without loss of generality, restrict  $\mathcal{F}$  to be the range of  $(x, y) \rightarrow \psi(x)y$  so that the RKHS  $\mathcal{H}$  defined by  $K$  is the (closure of) linear space spanned by  $\psi^T \alpha$  for  $\alpha \in \mathcal{F}$ . Note that the reproducing property (2.4) implies that for  $\alpha \in \mathcal{F}$

$$\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\psi(x)y \rangle_{\mathcal{H}} = \langle \psi^T(x)\alpha, y \rangle_{\mathcal{Y}} = \langle \alpha, \psi(x)y \rangle_{\mathcal{F}} \quad (2.9)$$

for all  $x, y \in \mathcal{X} \times \mathcal{Y}$ , which leads to the following theorem.

**Theorem 2.5.** The RKHS  $\mathcal{H}$  defined by the kernel (2.7) is the linear span of  $\psi^T \alpha$  over  $\alpha \in \mathcal{F}$  such that  $\|\alpha\|_{\mathcal{F}} < \infty$ . Furthermore,  $\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\alpha' \rangle_{\mathcal{H}} = \langle \alpha, \alpha' \rangle_{\mathcal{F}}$  and

$$\|\psi^T(\cdot)\alpha\|_{\mathcal{H}}^2 = \|\alpha\|_{\mathcal{F}}^2 \text{ for } \alpha, \alpha' \in \mathcal{F}. \quad (2.10)$$

**2.3. Kernel method solutions to the approximation problem 1.** In setting of Subsec. 2.1, assume the unknown function  $f^\dagger$  in Problem 1 to be contained in  $\mathcal{H}$ .

**2.3.1. The optimal recovery solution.** Using the relative error in  $\|\cdot\|_{\mathcal{H}}$ -norm as a loss, the minimax optimal recovery solution of Problem (1) is [63, Thm. 12.4,12.5] the minimizer (in  $\mathcal{H}$ ) of

$$\ell(X, Y) := \begin{cases} \text{Minimize} & \|f\|_{\mathcal{H}}^2 \\ \text{subject to} & f(X) = Y \end{cases} \quad (2.11)$$

By the representer theorem [52], the minimizer of (2.11) is

$$f(\cdot) = \sum_{j=1}^N K(\cdot, X_j) Z_j, \quad (2.12)$$

where the coefficients  $Z_j \in \mathcal{Y}$  are identified by solving the system of linear equations

$$\sum_{j=1}^N K(X_i, X_j) Z_j = Y_i \text{ for all } i \in \{1, \dots, N\}, \quad (2.13)$$

i.e.  $K(X, X)Z = Y$  where  $Z = (Z_1, \dots, Z_N)$ ,  $Y = (Y_1, \dots, Y_N) \in \mathcal{Y}^N$  and  $K(X, X)$  is the  $N \times N$  block-operator matrix<sup>13</sup> with entries  $K(X_i, X_j)$ . Therefore, writing  $K(\cdot, X)$  for the vector  $(K(\cdot, X_1), \dots, K(\cdot, X_N)) \in \mathcal{H}^N$ , the minimizer of (2.11) is

$$f(\cdot) = K(\cdot, X)K(X, X)^{-1}Y, \quad (2.14)$$

which implies

$$\ell(X, Y) = \|f\|_{\mathcal{H}}^2 = Y^T K(X, X)^{-1}Y, \quad (2.15)$$

where  $K(X, X)^{-1}$  is the inverse of  $K(X, X)$  (whose existence is implied by the non-degeneracy of  $K$  combined with  $X_i \neq X_j$  for  $i \neq j$ ).

**2.3.2. The ridge regression solution.** Let  $\lambda > 0$  and  $\ell_{\mathcal{Y}} : \mathcal{Y}^N \times \mathcal{Y}^N \rightarrow [0, \infty]$  be an arbitrary continuous positive loss. A ridge regression solution (also known as Tikhonov regularizer) to Problem 1 is a minimizer of

$$\ell(X, Y) := \inf_{f \in \mathcal{H}} \lambda \|f\|_{\mathcal{H}}^2 + \ell_{\mathcal{Y}}(f(X), Y). \quad (2.16)$$

Prototypical examples for  $\ell_{\mathcal{Y}}$  are the **empirical squared error**

$$\ell_{\mathcal{Y}}(Y', Y) = \sum_{i=1}^N \|Y'_i - Y_i\|_{\mathcal{Y}}^2, \quad (2.17)$$

used for general regression and the **hinge loss** [83]

$$\ell_{\mathcal{Y}}(Y', Y) = \sum_{i=1}^N \left( Y'_{i, \text{class}(Y_i)} - \max_{j \neq \text{class}(Y_i)} Y'_{i,j} - 1 \right)_+, \quad (2.18)$$

<sup>13</sup>For  $N \geq 1$  let  $\mathcal{Y}^N$  be the  $N$ -fold product space endowed with the inner-product  $\langle Y, Z \rangle_{\mathcal{Y}^N} := \sum_{i,j=1}^N \langle Y_i, Z_j \rangle_{\mathcal{Y}}$  for  $Y = (Y_1, \dots, Y_N), Z = (Z_1, \dots, Z_N) \in \mathcal{Y}^N$ .  $\mathbf{A} \in \mathcal{L}(\mathcal{Y}^N)$  given by  $\mathbf{A} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,N} \\ \vdots & & \vdots \\ A_{N,1} & \cdots & A_{N,N} \end{pmatrix}$  where  $A_{i,j} \in \mathcal{L}(\mathcal{Y})$ , is called a block-operator matrix. Its adjoint  $\mathbf{A}^T$  with respect to  $\langle \cdot, \cdot \rangle_{\mathcal{Y}^N}$  is the block-operator matrix with entries  $(A^T)_{i,j} = (A_{j,i})^T$ .

used for classification problems<sup>14</sup>. By the representer theorem, (2.16) admits a minimizer of the form  $f(\cdot) = K(\cdot, X)Z$  where  $Z \in \mathcal{Y}^N$  is identified as the minimizer of

$$\ell(X, Y) = \inf_{Z \in \mathcal{Y}^N} \lambda Z^T K(X, X)Z + \ell_{\mathcal{Y}}(K(X, X)Z, Y). \quad (2.19)$$

In particular, for  $\ell_{\mathcal{Y}}$  defined as in (2.17), the minimizer of (2.16) is

$$f(x) = K(x, X)(K(X, X) + \lambda I)^{-1}Y, \quad (2.20)$$

(writing  $I$  for the identity matrix) and the value of (2.16) at the minimum is

$$\ell(X, Y) = \lambda Y^T (K(X, X) + \lambda I)^{-1}Y. \quad (2.21)$$

### 3. Mechanical regression and idea registration

**3.1. Mechanical regression.** Motivated by the structure of Residual Neural Networks [34] we seek to approximate  $f^\dagger$  in Problem 1 by a function of the form

$$f^\dagger = f \circ \phi_L, \quad (3.1)$$

where (writing  $I$  for the identity map on  $\mathcal{X}$ )

$$\phi_L := (I + v_L) \circ \cdots \circ (I + v_1) \quad (3.2)$$

is a function (large deformation) mapping  $\mathcal{X}$  to itself obtained from the unknown residuals (small deformations)  $v_k : \mathcal{X} \rightarrow \mathcal{X}$  and  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is a ridge regression approximation of an unknown function mapping  $\phi_L(X)$  (the image of the data  $X$  under the deformation  $\phi_L$ ) to  $Y$ . We penalize the lack of regularity of the  $v_k$  and  $f$  by introducing an RKHS  $\mathcal{V}$  of functions mapping  $\mathcal{X}$  to itself and an RKHS  $\mathcal{H}$  of functions mapping  $\mathcal{X}$  to  $\mathcal{Y}$  and identify  $v_1, \dots, v_L$  and  $f$  by minimizing

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} L \sum_{s=1}^L \|v_s\|_{\mathcal{V}}^2 + \lambda \|f\|_{\mathcal{H}}^2 + \ell_{\mathcal{Y}}(f \circ (I + v_L) \circ \cdots \circ (I + v_1)(X), Y) \\ \text{over} & v_1, \dots, v_L \in \mathcal{V} \text{ and } f \in \mathcal{H}, \end{cases} \quad (3.3)$$

where  $\nu$  is a strictly positive parameter balancing the regularity of  $\phi_L$  with that of  $f$  and  $\lambda > 0$  balances the regularity of  $f$  with the loss  $\ell_{\mathcal{Y}}=(2.17)$ .

**3.2. Ridge regression loss.** The variational problem (3.3) can be written

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} L \sum_{s=1}^L \|v_s\|_{\mathcal{V}}^2 + \ell((I + v_L) \circ \cdots \circ (I + v_1)(X), Y) \\ \text{over} & v_1, \dots, v_L \in \mathcal{V}, \end{cases} \quad (3.4)$$

where  $\ell : \mathcal{X}^N \times \mathcal{Y}^N \rightarrow [0, \infty]$  is the **ridge regression loss** (2.16)=(2.21). The following condition (which we will from now on assume to be satisfied) ensures the continuity of  $\ell=(2.21)$ .

**Condition 3.1.** *Assume that (1)  $x \rightarrow K(x, x')$  is continuous and for all  $x'$  (2)  $\mathcal{X}$  and  $\mathcal{Y}$  are finite-dimensional.*

<sup>14</sup>(2.18) seeks to maximize the margin between correct and incorrect labels and is defined for  $\mathcal{Y} = \mathbb{R}^{d_{\mathcal{Y}}}$  by writing  $Y'_{i,j}$  for the entries of  $Y'_i$ , using  $\operatorname{argmax}_j Y'_{i,j}$  for the predicted label for the data  $i$ , setting  $\operatorname{class}(Y_i) = j$  if the label/class of  $X_i$  is  $j$  and writing  $a_+ := \max(a, 0)$ .

We will now focus on the reduction of (3.4) and only assume  $\ell : \mathcal{X}^N \times \mathcal{Y}^N \rightarrow [0, \infty]$  to be continuous and positive.

**3.3. Reduction to a discrete least action principle.** Although  $\ell$  may not be convex, the first part of (3.4) is quadratic and can be reduced as shown in Thm 3.3. We will from now on work under the following condition<sup>15</sup> where  $\Gamma$  is the kernel defined by  $\mathcal{V}$ .

**Condition 3.2.** Assume that (1) there exists  $r > 0$  such that  $Z^T \Gamma(X, X) Z \geq r Z^T Z$  for all  $Z \in \mathcal{X}^N$ , (2)  $\Gamma$  admits  $\mathcal{F}$  and  $\psi$  as feature space/map,  $\mathcal{F}$  is finite-dimensional,  $\psi$  and its first and second order partial derivatives are continuous and uniformly bounded, and (3)  $\mathcal{X}$  is finite-dimensional.

**Theorem 3.3.**  $v_1, \dots, v_L \in \mathcal{V}$  is a minimizer of (3.4) if and only if

$$v_s(x) = \Gamma(x, q^s) \Gamma(q^s, q^s)^{-1} (q^{s+1} - q^s) \text{ for } x \in \mathcal{X}, s \in \{1, \dots, L\}, \quad (3.5)$$

where  $q^1, \dots, q^{L+1} \in \mathcal{X}^N$  is a minimizer of (write  $\Delta t := 1/L$ )

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \sum_{s=1}^L \left( \frac{q^{s+1} - q^s}{\Delta t} \right)^T \Gamma(q^s, q^s)^{-1} \left( \frac{q^{s+1} - q^s}{\Delta t} \right) \Delta t + \ell(q^{L+1}, Y) \\ \text{over} & q^2, \dots, q^{L+1} \in \mathcal{X}^N \text{ with } q^1 = X. \end{cases} \quad (3.6)$$

*Proof.* Introduce the variables  $q_i^{s+1} = (I + v_s) \circ \dots \circ (I + v_1)(X_i)$  for  $2 \leq s \leq L$ , and  $q_i^1 = X_i$ . (3.4) is then equivalent to

$$\begin{cases} \text{Minimize} & \nu \frac{L}{2} \sum_{s=1}^L \|v_s\|_{\mathcal{V}}^2 + \ell(q^{L+1}, Y) \\ \text{over} & v_1, \dots, v_s \in \mathcal{V}, \quad q^1, \dots, q^{L+1} \in \mathcal{X}^N \\ \text{subject to} & q^1 = X \text{ and } v_s(q^s) = q^{s+1} - q^s \text{ for all } s \end{cases} \quad (3.7)$$

Minimizing with respect to the  $v_s$  first we obtain  $\|v_s\|_{\mathcal{V}}^2 = (q^{s+1} - q^s)^T \Gamma(q^s, q^s)^{-1} (q^{s+1} - q^s)$  and (3.5). (3.7) can then be reduced to (3.6).  $\square$

**3.4. Continuous limit and neural least action principle.** Interpreting  $\Delta t = 1/L$  as the time step, (3.6) is the discrete least action principle [49] obtained by using the approximation

$$\left( \frac{q^{s+1} - q^s}{\Delta t} \right)^T \Gamma(q^s, q^s)^{-1} \left( \frac{q^{s+1} - q^s}{\Delta t} \right) \approx \dot{q}_{\frac{s}{L}}^T \Gamma(q_s, q_s) \dot{q}_{\frac{s}{L}}$$

in the continuous least action principle

$$\begin{cases} \text{Minimize} & \nu \mathcal{A}[q] + \ell(q(1), Y) \\ \text{over} & q \in C^1([0, 1], \mathcal{X}^N) \text{ subject to } q(0) = X. \end{cases} \quad (3.8)$$

where  $\mathcal{A}[q]$  is the action

$$\mathcal{A}[q] := \int_0^1 \mathfrak{L}(q, \dot{q}) dt \quad (3.9)$$

defined by the Lagrangian

$$\mathfrak{L}(q, \dot{q}) := \frac{1}{2} \dot{q}^T \Gamma(q, q)^{-1} \dot{q}, \quad (3.10)$$

<sup>15</sup>Note that Cond. 3.2.(1) is equivalent to the non singularity of  $\Gamma(X, X)$  and (2) implies that  $(x, x') \rightarrow \Gamma(x, x')$  and its first and second order partial derivatives are continuous and uniformly bounded.

and  $C^1([0, 1], \mathcal{X}^N)$  is the set of continuously differentiable functions  $q : [0, 1] \rightarrow \mathcal{X}^N$  mapping  $s \in [0, 1]$  to  $q_s \in \mathcal{X}^N$ . Consequently, minimizing (3.7) corresponds to using a first-order variational symplectic integrator (simulating a nearby mechanical system [31]) to approximate (3.8). We will present convergence results in Thm. 3.11.

**3.5. Euler-Lagrange equations and geodesic motion.** Following classical Lagrangian mechanics [48], a minimizer of (3.8) follows the Euler-Lagrange equations  $\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}} - \frac{\partial \mathcal{L}}{\partial q} = 0$ , i.e.

$$\frac{d}{dt} (\Gamma(q, q)^{-1} \dot{q}) = \partial_q \left( \frac{1}{2} \dot{q}^T \Gamma(q, q)^{-1} \dot{q} \right) \quad (3.11)$$

Furthermore,  $\Gamma^{-1}(q, q)$  can be interpreted as a mass matrix or metric tensor [48, p. 3] and the Euler-Lagrange equations are equivalent to the equations of geodesic motion [48, Sec. 7.5] corresponding to minimizing the length  $\int_0^1 \sqrt{\dot{q}^T \Gamma(q, q)^{-1} \dot{q}} ds$  of the curve  $q$  connecting  $X$  to  $q(1)$  (which, using the equivalence between minimizing length and length squared, can also be recovered as a limit by replacing  $\frac{L}{2} \sum_{s=1}^L \|v_s\|_{\mathcal{V}}^2$  by  $\sum_{s=1}^L \|v_s\|_{\mathcal{V}}$  in (3.4)).

**3.6. Hamiltonian mechanics.** Introduce the momentum variable

$$p = \frac{\partial \mathcal{L}}{\partial \dot{q}} = \Gamma(q, q)^{-1} \dot{q}, \quad (3.12)$$

and the Hamiltonian  $\mathfrak{H}(q, p) = p^T \dot{q} - \mathcal{L}(q, \dot{q}) = \frac{1}{2} \dot{q}^T \Gamma(q, q)^{-1} \dot{q} = \frac{1}{2} p^T \Gamma(q, q) p = (1.15)$ . The following theorem summarizes the classical [48] correspondence between the Lagrangian and Hamiltonian viewpoints.

**Theorem 3.4.** *If  $q$  is a minimizer of the least action principle (3.8) then  $(q, p)$  follows the Hamiltonian dynamic*

$$\begin{cases} \dot{q} = \frac{\partial \mathfrak{H}(q, p)}{\partial p} = \Gamma(q, q) p \\ \dot{p} = -\frac{\partial \mathfrak{H}(q, p)}{\partial q} = -\partial_q \left( \frac{1}{2} p^T \Gamma(q, q) p \right), \end{cases} \text{ with initial value } (q(0) = X, p(0)). \quad (3.13)$$

The energy  $\mathfrak{H}(q, p)$  is conserved by this dynamic and any function  $F$  of  $(q, p)$  evolves according to the Lie derivative

$$\frac{d}{dt} F(q, p) = \{F, \mathfrak{H}\} = \partial_q F \partial_p \mathfrak{H} - \partial_p F \partial_q \mathfrak{H} = \partial_q F \Gamma(q, q) p - \partial_p F \partial_q \left( \frac{1}{2} p^T \Gamma(q, q) p \right). \quad (3.14)$$

**3.6.1. In feature space.** Let  $\mathcal{F}$  and  $\psi$  be a feature space/map of  $\Gamma$  as in Cond. 3.2. Using the identity  $\Gamma(x, x') = \psi^T(x) \psi(x')$ , the Hamiltonian system (3.13) can be written

$$\begin{cases} \dot{q}_i = \psi^T(q_i) \alpha \\ \dot{p}_i = -\partial_x (p_i^T \psi^T(x) \alpha) \Big|_{x=q_i}, \end{cases} \quad (3.15)$$

where  $\alpha$  is the time dependent element of  $\mathcal{F}$  defined by

$$\alpha := \sum_{j=1}^N \psi(q_j) p_j. \quad (3.16)$$

Energy preservation and the identity  $\|\alpha\|_{\mathcal{F}}^2 = p^T \Gamma(q, q) p$ , implies the following.

**Proposition 3.5.**  $t \rightarrow \|\alpha(t)\|_{\mathcal{F}}$  is constant.

**3.6.2. Existence and uniqueness.** Cond. 3.2 provides sufficient regularity on  $\Gamma$  for the existence and uniqueness of a solution to (3.13) in  $C^2([0, 1], \mathcal{X}^N) \times C^1([0, 1], \mathcal{X}^N)$ .

**Theorem 3.6.** (3.13) admits a unique solution in  $C^2([0, 1], \mathcal{X}^N) \times C^1([0, 1], \mathcal{X}^N)$ .

*Proof.* (3.15) implies that  $\|\dot{p}_i\|_{\mathcal{Y}} \leq \|p_i\|_{\mathcal{Y}} \|\alpha\|_{\mathcal{F}} \sup_x \|\nabla \psi(x)\|$ . Therefore Prop. 3.5 implies that  $p(t)$  remains in a bounded domain  $B$  (for  $t \in [0, 1]$ ). The regularity of  $\Gamma$  (Cond. 3.2) implies that the vector field of (3.13) is uniformly Lipschitz for  $p \in B$ . We conclude from the global version of the Picard-Lindelöf theorem [4, Thm. 1.2.3].  $\square$

**3.6.3. Geodesic shooting.** The Hamiltonian representation of minimizers of (3.8) enables its reduction to the search for an initial momentum  $p(0)$ . This method, known as geodesic shooting in image registration [1], is summarized in the following theorem.

**Theorem 3.7.** Write  $p = \Gamma(q, q)^{-1} \dot{q} = (3.12)$  for  $q \in C^1([0, 1], \mathcal{X}^N)$ .  $q$  is a minimizer of (3.8) if and only if  $(q, p)$  follows the Hamiltonian dynamic (3.13),  $q(0) = X$  and  $p(0)$  is a minimizer of

$$\mathfrak{R}(p(0), X, Y) := \frac{\nu}{2} p^T(0) \Gamma(X, X) p(0) + \ell(q(1), Y). \quad (3.17)$$

Furthermore,  $p(1)$  satisfies

$$\nu p(1) + \partial_{q(1)} \ell(q(1), Y) = 0. \quad (3.18)$$

*Proof.* Zeroing the Fréchet derivative of (3.8) with respect to the trajectory  $q(t)$  implies that a minimizer of (3.8) must satisfy the Hamiltonian dynamic (3.13) and the boundary condition (3.18) (which is analogous to the one obtained in image registration [1, Eq. 7]). Since the energy  $p^T \Gamma(q, q) p / 2$  is preserved along the Hamiltonian flow, the minimization of (3.8) can be reduced to that of (3.17) with respect to  $p(0)$ .  $\square$

**3.7. Idea registration.** Let  $C([0, 1], \mathcal{V})$  be the space of continuous functions  $v : \mathcal{X} \times [0, 1] \rightarrow \mathcal{X}$  such that  $x \rightarrow v(x, t)$  belongs to  $\mathcal{V}$  (for all  $t \in [0, 1]$ ) and is uniformly (in  $t$  and  $x$ ) Lipschitz continuous. For  $v \in C([0, 1], \mathcal{V})$  write  $\phi^v : \mathcal{X} \times [0, 1] \rightarrow \mathcal{X}$  for the solution of (1.7). By the Picard-Lindelöf theorem [4, Thm. 1.2.3] the solution of (1.7) exists and is unique if  $\mathcal{X}$  is finite-dimensional<sup>16</sup>, which is ensured by Cond. 3.2. Instead of reducing (3.4) to (3.7), consider its infinite depth limit<sup>17</sup> and observe that, in the limit  $L \rightarrow \infty$ ,  $(I + v_k) \circ \dots \circ (I + v_1)$  approximates (at time  $t_k := \frac{k}{L}$ ) the flow map  $\phi^v$  where  $v$  is a minimizer of

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 \|v\|_{\mathcal{V}}^2 dt + \ell(\phi^v(X, 1), Y) \\ \text{over} & v \in C([0, 1], \mathcal{V}). \end{cases} \quad (3.19)$$

The proof of this convergence, stated in Thm. 3.11, is based on the following reduction theorem.

<sup>16</sup>The simplicity of the proof of existence and uniqueness of solutions for (1.7) is the main reason why we work under Cond. 3.2. Although [87, Thm. 3.3] could be used when  $\dim(\mathcal{X}) = \infty$ , the existence and uniqueness of solutions for ODEs can be quite delicate in general infinite-dimensional spaces [47].

<sup>17</sup>The infinite depth limit (3.19) can be interpreted as an image registration variational problem with images replaced by abstractions deformed by arbitrary kernels. [24] shows that if  $\mathcal{V}$  is defined by a differential operator of sufficiently high order in a Sobolev space, then  $\phi^v$  is a diffeomorphism (a differentiable bijection). Although the bijectivity of  $\phi^v$  is a natural requirement in image registration, it is not needed for idea registration since two inputs may share the same label.



**Theorem 3.8.**  $v$  is a minimizer of (3.19) if and only if

$$\dot{\phi}^v(x, t) = \Gamma(\phi^v(x, t), q_t) \Gamma(q_t, q_t)^{-1} \dot{q}_t \text{ with } \phi^v(x, 0) = x \in \mathcal{X} \quad (3.20)$$

where  $q$  is a minimizer of the least action principle (3.8). Furthermore (defining  $\mathcal{A}[q]$  as in (3.9)), for  $q \in C^1([0, 1], \mathcal{X}^N)$ ,

$$\mathcal{A}[q] = \inf_{v \in C([0, 1], \mathcal{V}) : \phi^v(q(0), t) = q(t) \forall t \in [0, 1]} \int_0^1 \frac{1}{2} \|v\|_{\mathcal{V}}^2 dt, \quad (3.21)$$

and the representer PDE (3.20) can be written as (1.14), where  $(q, p)$  is the solution of the Hamiltonian system (3.13) with initial condition  $q(0) = X$  and  $p(0)$  identified as a minimizer of (3.17).

*Proof.* The proof of (3.20) and (3.21) is identical to that of Thm. 3.3. (1.14) follows from theorems 3.4 and 3.7.  $\square$

Figure 3 summarizes the correspondence between the least action principles obtained from (3.4) under reduction and/or infinite depth limit.

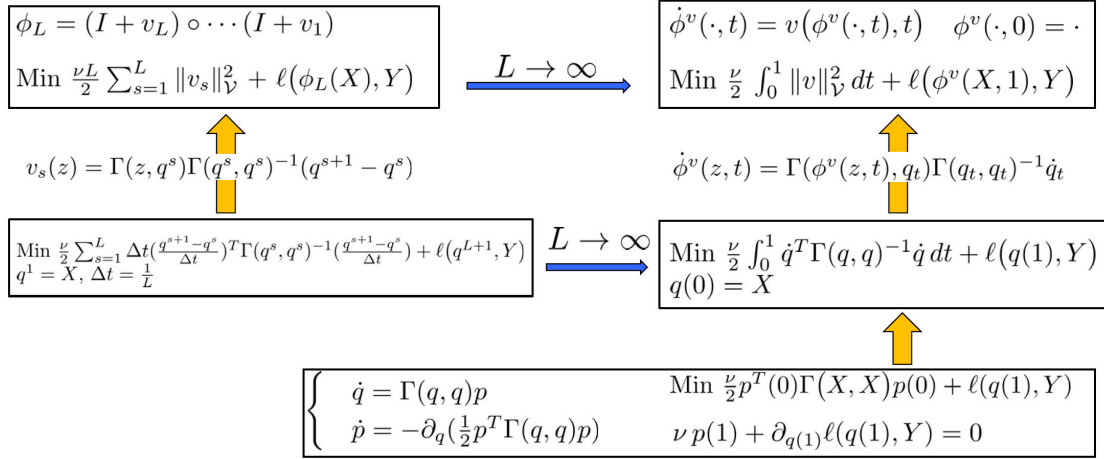


FIGURE 3. Least action principles after reduction and/or infinite depth limit.

**3.8. Existence/identification of minimizers and energy preservation.** Although it is simple to show the existence of minimizers for (3.8), (3.17) and (3.19) (see Thm. 3.9 below) we will not attempt to identify sufficient conditions for their uniqueness since pathological landmark matching examples [54] suggest that, even with smooth kernels, these minimizers may not be unique<sup>18</sup>.

<sup>18</sup>For a simple example, consider a rigid pendulum spinning about the origin. Starting from the stable equilibrium point (pendulum down, zero velocity), consider the problem of finding a minimal energy initial momentum arriving at the unstable equilibrium (pendulum up) with zero velocity. This problem is analogous to minimizing (3.8) and has two solutions. This lack of uniqueness is also related to the notions of *nonconjugate* solutions [48, Def. 7.4.4] in classical mechanics and *conjugate* points [48, p. 198] in the study of geodesics.

**Theorem 3.9.** *The minimum values of (3.8), (3.17) and (3.19) are identical. (3.8), (3.17) and (3.19) have minimizers.  $q$  is a minimizer of (3.8) if and only if  $(q, p)$  ( $p = \Gamma(q, q)^{-1}\dot{q}$ ) follows the Hamiltonian dynamic (3.13) (with  $q(0) = X$ ) and  $p(0) = \Gamma(q(0), q(0))^{-1}\dot{q}(0)$  is a minimizer of  $\mathfrak{V}(p(0), X, Y) = (3.17)$ .  $v$  is a minimizer of (3.19) if and only if  $v(x, t) = \Gamma(x, q(t))p(t)$  with  $(q, p)$  following the Hamiltonian dynamic (3.13) (with  $q(0) = X$ ) and  $p(0)$  being a minimizer of  $\mathfrak{V}(p(0), X, Y) = (3.17)$ . Therefore the minimizers of (3.8) and (3.19) can be parameterized by their initial momentum identified as a minimizer of  $\mathfrak{V}(p(0), X, Y) = (3.17)$ . Furthermore, at those minima, the energies  $\frac{1}{2}\dot{q}^T \Gamma(q, q)^{-1}\dot{q}$  and  $\frac{1}{2}\|v\|_{\mathcal{V}}^2$  are constant over  $t \in [0, 1]$  and equal to  $\frac{\nu}{2}p^T(0)\Gamma(X, X)p(0)$ .*

*Proof.* Given Thm. 3.7 and Thm. 3.8 we only need to prove the existence of minimizers for  $\mathfrak{V}(p(0), X, Y) = (3.17)$ . Let  $B_\rho := \{p(0) \in \mathcal{X}^N \mid p(0)^T p(0) \leq \rho^2\}$ . Since  $\mathcal{X}$  is finite-dimensional this ball is compact. Since  $\ell$  is continuous in  $q(1)$  and  $q(1)$  is continuous in  $p(0)$  [4, Thm. 1.4.1.], (3.17) must have a minimizer in  $B_\rho$ . Since  $\ell$  is positive, Cond. 3.2.(1) implies that (3.17) diverges towards infinity as  $p(0)^T p(0) \rightarrow \infty$ . It follows that, for  $\rho$  large enough,  $B_\rho$  contains at least one global minimizer of (3.17) and all global minimizers are contained in  $B_\rho$ . Note that by Thm. 3.6 and the representation (3.20) it holds true that a minimizer  $v$  of (3.19) must be an element of  $C([0, 1], \mathcal{V})$ .  $\square$

We will now show the existence of minimizers for (3.4) and (3.6). [91, Thm. 3.3] implies<sup>19</sup> (under the regularity conditions 3.2 on  $\Gamma$ ) that the trajectory  $q^1, \dots, q^{L+1}$  of a minimizer of the discrete least action principle (3.6) follows a first-order<sup>20</sup> symplectic integrator for the Hamiltonian system (3.13). Introducing the momentum variables (1.12) this discrete integrator is (1.13). Write

$$\mathfrak{V}_L(p^1, X, Y) := \begin{cases} \frac{\nu}{2} \sum_{s=1}^L (p^s)^T \Gamma(q^s, q^s) p^s \Delta t + \ell(q^{L+1}, Y) \\ p^s = (1.12) \text{ and } (q^s, p^s) \text{ follow (1.13) with } q^1 = X. \end{cases} \quad (3.22)$$

**Theorem 3.10.** *The minimum values of (3.4), (3.6) and  $\mathfrak{V}_L(p^1, X, Y)$  (in  $p^1$ ) are identical. (3.4), (3.6) and (3.22) have minimizers.  $q^1, \dots, q^{L+1}$  is a minimizer of (3.6) if and only if  $(q^s, p^s)$  (with  $p^s = (1.12)$ ) follows the discrete Hamiltonian map (1.13),  $q^1 = X$  and  $p^1$  is a minimizer of  $\mathfrak{V}_L(p^1, X, Y) = (3.22)$ .  $v_1, \dots, v_L$  is a minimizer of (3.4) if and only if  $v_s(x) = \Delta t \Gamma(x, q^s) p^s = (3.5)$  where  $(q^s, p^s)$  follows the discrete Hamiltonian map (1.13) with  $q^1 = X$  and  $p^1$  is a minimizer of  $\mathfrak{V}_L(p^1, X, Y) = (3.22)$ . Therefore the minimizers of (3.4) and (3.6) can be parameterized by their initial momentum identified as a minimizer of  $\mathfrak{V}_L(p^1, X, Y) = (3.22)$ . At those minima, the energies  $\frac{1}{2}(p^s)^T \Gamma(q^s, q^s) p^s$  and  $\frac{1}{2}\|v_s\|_{\mathcal{V}}^2$  are equal and fluctuate by at most  $\mathcal{O}(1/L)$  over  $s \in \{1, \dots, L\}$ .*

*Proof.* By Thm. 3.3 we only need to prove the result for (3.22). Since (under Cond. 3.2)  $\mathfrak{V}_L(p^1, X, Y)$  diverges towards infinity as  $(p^1)^T p^1 \rightarrow \infty$  and since  $\mathfrak{V}_L(p^1, X, Y)$  is continuous, as in proof of Thm. 3.9, for  $\rho$  large enough, (3.22) must have a global minimizer in  $B_\rho := \{p^1 \in \mathcal{X}^N \mid (p^1)^T p^1 \leq \rho^2\}$  and all global minimizers must be contained in  $B_\rho$ . By

<sup>19</sup>Such results are part of the discrete mechanics literature on discretized least action principles. General accuracy results could also be derived from [49, Sec. 2] and [10, p. 114] and  $\Gamma$ -convergence results could be derived from [57].

<sup>20</sup>Higher order symplectic partitioned Runge Kutta discretizations [31, Sec. 2.6.5] of (3.13) would lead to numerical schemes akin to Densely Connected Networks [37].

Thm. 3.3  $\|v_s\|_Y^2$  is equal to  $(p^s)^T \Gamma(q^s, q^s) p^s$  where  $(q^s, p^s)$  is obtained from the (first-order) symplectic and variational integrator (1.13) for the Hamiltonian system (3.13). The near energy preservation then follows from [31, Thm. 8.1] (derived from the fact that symplectic integrators simulate a nearby mechanical system) and the order of accuracy of (1.13).  $\square$

**3.9. Convergence of minimal values and minimizers.** Due to the possible lack of uniqueness of mechanical regression solutions (discussed in Subsec. 3.8) the convergence of minimizers must be indexed based on their initial momentum parametrization as described in Thm. 3.9 and Thm. 3.10. Write  $\mathfrak{M}_L(X, Y)$  for the set of minimizers  $p^1$  of  $\mathfrak{V}_L(p^1, X, Y) = (3.22)$ . Write  $\mathfrak{M}(X, Y)$  for the set of minimizers  $p(0)$  of  $\mathfrak{V}(p(0), X, Y) = (3.17)$ .

**Theorem 3.11.** *The minimal value of (3.4), (3.6) and (3.22) converge, as  $L \rightarrow \infty$ , towards the minimal value of (3.8), (3.17), (3.19). As  $L \rightarrow \infty$ , the set of adherence values<sup>21</sup> of  $\mathfrak{M}_L(X, Y)$  is  $\mathfrak{M}(X, Y)$ . Let  $v_s^L, q_L^s$  and  $p_L^1$  be sequences of minimizers of (3.4), (3.4) and (3.22) indexed by the same sequence  $p_L^1$  of initial momentum in  $\mathfrak{M}_L(X, Y)$  (as described in Thm. 3.10). Then, the adherence points of the sequence  $p_L^1$  are in  $\mathfrak{M}(X, Y)$  and if  $p(0)$  is such a point ( $p_L^1$  converges towards  $p(0)$  along a subsequence  $L_k$ ) then, along that subsequence: (1) The trajectory formed by interpolating the states  $q_L^s \in \mathcal{X}^N$  converges to the trajectory formed by a minimizer of (3.8) with initial momentum  $p(0)$ . (2) For<sup>22</sup>  $t \in [0, 1]$ ,  $(I + v_{\text{int}(tL)}^L) \circ \cdots \circ (I + v_1^L)(x)$  converges to  $\phi^v(x, t) = (1.7)$  where  $v$  is a minimizer of (3.19) with initial momentum  $p(0)$ . Conversely if  $p(0) \in \mathfrak{M}(X, Y)$  then it is the limit of a sequence  $p_L^1 \in \mathfrak{M}_L(X, Y)$  and the minimizers of (3.4), (3.6) and (3.22) with initial momentum  $p_L^1$  converge (in the sense given above) to the minimizers of (3.8), (3.17), (3.19) with initial momentum  $p(0)$  (as described in Thm. 3.9).*

*Proof.* By Thm. 3.10 and Thm. 3.9 the convergence of minimum values follows from that of  $\mathfrak{V}_L(p^1, X, Y)$  towards that of  $\mathfrak{V}(p(0), X, Y)$ . As shown in the proof of Thm. 3.9 and Thm. 3.10, the initial momenta minimizing  $\mathfrak{V}_L$  and  $\mathfrak{V}$  are contained in a compact set  $B_\rho$  (independent from  $L$ ). The uniform convergence (for  $p^1 = p(0) \in B_\rho, q^1 = q(0) = X$ ) of the solution of the integrator (1.13) towards the solution of the Hamiltonian system (3.13) implies that  $\lim_{L \rightarrow \infty} \min_{p^1 \in B_\rho} \mathfrak{V}_L(p^1, X, Y) = \min_{p^1 \in B_\rho} \lim_{L \rightarrow \infty} \mathfrak{V}_L(p^1, X, Y) = \min_{p^1 \in B_\rho} \mathfrak{V}(p^1, X, Y)$ . Which proves the convergence of minimum values. Similarly the uniform convergence (over  $p^1 \in B_\rho$ ) of  $\mathfrak{V}_L(p^1, X, Y)$  towards  $\mathfrak{V}(p^1, X, Y)$  (also obtained from the uniform convergence of the solution of (1.13) in  $B_\rho$ ) implies that the set of adherence values of  $\mathfrak{M}_L(X, Y)$  is  $\mathfrak{M}(X, Y)$ . Let  $p_L^1$  be a sequence in  $\mathfrak{M}_L(X, Y)$  and let  $p(0) \in \mathfrak{M}(X, Y)$  be one of its adherence points. The convergence of  $p_L^1$  towards  $p(0)$  (along a subsequence  $L_k$ ) and the uniform convergence (along  $L_k$ ) of the solution of (1.13) towards the solution of (3.13) implies (1). (1) and the representation  $v_s = \Gamma(x, q^s) p^s = (3.5)$  of Thm. 3.10 imply that  $(I + v_{\text{int}(tL)}^L) \circ \cdots \circ (I + v_1^L)(x)$  converges to  $z(t)$  where  $z$  is the solution of the ODE

$$\dot{z} = \Gamma(z, q)p \text{ with } z(0) = x, \quad (3.23)$$

<sup>21</sup>Writing  $\text{cl } A$  for the closure of a set  $A$ ,  $\cap_{L' \geq 1} \text{cl } \cup_{L \geq L'} \mathfrak{M}_L(X, Y) = \mathfrak{M}(X, Y)$ .

<sup>22</sup>Write  $\text{int}(tL)$  for the integer part of  $tL$ .

where  $(q, p)$  follows the Hamiltonian dynamic (3.13) with initial value  $q(0) = X$  and  $p(0)$ . We conclude that (2) holds true by the representation  $v(x, t) = \Gamma(x, q)p$  of Thm. 3.9 and by observing that (3.23) is a characteristic curve of (1.14). The proof of the remaining (conversely) portion of the theorem is identical.  $\square$

Observe that, with  $\ell = (2.16)$ , (3.19) is equivalent to minimizing

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 \|v\|_{\mathcal{V}}^2 dt + \lambda \|f\|_{\mathcal{H}}^2 + \ell_{\mathcal{Y}}(f \circ \phi^v(X, 1), Y) \\ \text{over} & v \in C([0, 1], \mathcal{V}) \text{ and } f \in \mathcal{H}. \end{cases} \quad (3.24)$$

Thm. 3.11 implies that minimizers of (3.3) converge towards minimizers of (3.24) in the sense of the following corollary.

**Corollary 3.12.** *As  $L \rightarrow \infty$ , (1) the minimum value of (3.3) converges towards the minimum value of (3.24). If  $(v_1, \dots, v_L, f)$  is a sequence of minimizers of (3.3) then the set of adherence values of  $f \circ (I + v_L) \circ \dots \circ (I + v_1)$  is*

$$\{f \circ \phi^v(\cdot, 1) \mid (v, f) \text{ is a minimizer of (3.24)}\}, \quad (3.25)$$

*i.e., the sequence  $(v_1, \dots, v_L, f)$  can be partitioned into subsequences such that, along each subsequence,  $f \circ (I + v_L) \circ \dots \circ (I + v_1)(x)$  converges (for all  $x \in \mathcal{X}$ ) towards  $f \circ \phi^v(x, 1)$  where  $(v, f)$  is a minimizer of (3.24).*

*Proof.* The proof is a direct consequence of Thm. 3.11. Observe that we are using the fact that a minimizer  $f$  of (3.3) is unique given  $(v_1, \dots, v_L)$  and a minimizer  $f$  of (3.24) is unique given  $v$ .  $\square$

**3.10. Mechanical regression is ridge regression with an RKHS learned from data.** In the infinite depth (continuous time) limit, the mechanical regression approach to Problem 1 is to approximate  $f^\dagger$  with

$$f^\ddagger(\cdot) = f \circ \phi^v(\cdot, 1) \quad (3.26)$$

where  $v$  and  $f$  are minimizers of (3.24) (and  $\phi^v$  solves (1.7)). The following proposition shows that mechanical regression is equivalent to performing optimal recovery or ridge regression with an RKHS  $\mathcal{H}^v$  that is learned from the data  $(X, Y)$ . The  $\nu$  penalty avoids overfitting that RKHS to the data.

**Proposition 3.13.** *Let  $\phi^v$  be the solution of (1.7) for an arbitrary  $v \in C([0, 1], \mathcal{V})$ . Let  $\mathcal{H}^v$  be the RKHS associated with  $K^v(x, x') := K(\phi^v(x, 1), \phi^v(x', 1))$ . If  $f^v$  is the minimizer of  $\lambda \|f'\|_{\mathcal{H}^v}^2 + \ell_{\mathcal{Y}}(f'(X), Y)$  over  $f' \in \mathcal{H}^v$  and  $f$  is the minimizer of  $\lambda \|f'\|_{\mathcal{H}}^2 + \ell_{\mathcal{Y}}(f' \circ \phi^v(X, 1), Y)$  over  $f' \in \mathcal{H}$ . It holds true that  $f^v(\cdot) = f \circ \phi^v(\cdot, 1)$  and*

$$\inf_{f' \in \mathcal{H}} \lambda \|f'\|_{\mathcal{H}}^2 + \ell_{\mathcal{Y}}(f' \circ \phi^v(X, 1), Y) = \inf_{f' \in \mathcal{H}^v} \lambda \|f'\|_{\mathcal{H}^v}^2 + \ell_{\mathcal{Y}}(f'(X), Y). \quad (3.27)$$

*In particular, if  $\ell_{\mathcal{Y}}$  is the  $\ell_{\mathcal{Y}}$  in the ridge regression loss (2.16), then a mechanical regression solution  $f^\ddagger = (3.26)$  to Problem (1) is a minimizer of*

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 \|v\|_{\mathcal{V}}^2 dt + \lambda \|f\|_{\mathcal{H}^v}^2 + \ell_{\mathcal{Y}}(f(X), Y) \\ \text{over} & v \in C([0, 1], \mathcal{V}) \text{ and } f \in \mathcal{H}^v, \end{cases} \quad (3.28)$$

*whose minimal value is  $\lambda \|f^v\|_{\mathcal{H}^v}^2$  for  $\ell_{\mathcal{Y}} = (2.17)$ , where  $\mathcal{H}_\lambda^v$  is the RKHS associated with  $K_\lambda^v := K^v + \lambda I$ .*

*Proof.*  $f^v(\cdot) = f \circ \phi^v(\cdot, 1)$  follows from (2.20) and (3.27) follows from (2.21).  $\square$

**3.11. Information in momentum variables and sparsity.** Consider the Hamiltonian system (3.13). While  $q$  has a clear interpretation as the displacement of the input data  $X$ , that of the momentum variable  $p$  is less transparent. The following theorem shows that the entry  $p_i$  of  $p$  is zero if  $q_i(1)$  does not contribute to the loss. Therefore, using the hinge loss (2.18), the only indices  $i$  having a non-zero momentum will be those within or at the boundary of the safety margin. In that sense  $p_i$  represents the contribution of the data point  $(X_i, Y_i)$  to the predictor  $\phi^v(\cdot, 1)$  obtained from (3.19) and (1.14). This phenomenon is analogous to the sparse representations obtained with support vector machines [83] where the predictor is represented with the subset of the training points (the support vectors) within the safety margin of the hinge loss.

**Theorem 3.14.** *Let  $(q, p)$  be the solution of the Hamiltonian system (3.13) with initial state  $q(0) = X$  and  $p(0)$  minimizing (3.17). For  $i \in \{1, \dots, N\}$ , it holds true that  $p_i(t) = 0$  for all  $t \in [0, 1]$  if and only if  $\partial_{q_i(1)} \ell(Y, q(1)) = 0$ .*

*Proof.* Combine (3.18) with Lem. 3.15.  $\square$

**Lemma 3.15.** *Let  $(q, p)$  be a solution of the Hamiltonian system (3.13). If  $p_i(t_0) = 0$  for some  $t_0 \in [0, 1]$  then  $p_i(t) = 0$  for all  $t \in [0, 1]$ .*

*Proof.*  $\dot{p} = -\partial_q(\frac{1}{2}p^T \Gamma(q, q)p)$  implies that  $\dot{p}_i(t) = 0$  if  $p_i(t) = 0$  and  $p_i(t) = 0$  for  $t \geq t_0$  follows by integration. Since the time reversed trajectory  $t \rightarrow (q, -p)(1-t)$  also satisfies the Hamiltonian system (3.13) the result also follows by integration for  $t \in [0, t_0]$ .  $\square$

Following [80, 63, 79] we will now interpret the number of particles represented by the components of  $q$  as a notion of scale and initiate a multiresolution description of the action  $\mathcal{A}$  supporting the proposed interpretation of the momentum variables. For  $q \in C^1([0, 1], \mathcal{X}^N)$  (with  $N$  being an arbitrary integer) define  $\mathcal{A}[q]$  as in (3.9). For  $q^1 \in C^1([0, 1], \mathcal{X}^{N_1})$  and  $q^2 \in C^1([0, 1], \mathcal{X}^{N_2})$  write  $\mathcal{A}[(q^1, q^2)]$  for the action of the trajectory  $t \rightarrow q(t) := (q^1(t), q^2(t))$  in  $\mathcal{X}^{N_1+N_2}$ . Note that we have the following consistency relation.

**Proposition 3.16.** *Let  $q^1 \in C^1([0, 1], \mathcal{X}^{N_1})$  and  $X^2 \in \mathcal{X}^{N_2}$  be arbitrary. It holds true that*

$$\mathcal{A}[q^1] = \inf_{q^2 \in C^1([0, 1], \mathcal{X}^{N_2}) : q^2(0) = X^2} \mathcal{A}[(q^1, q^2)] \quad (3.29)$$

*Proof.* Observe that (as in Thm. 3.8)  $\inf_{v \in C([0, 1], \mathcal{V}) : \phi^v(q^1(0), t) = q^1(t) \forall t \in [0, 1]} \int_0^1 \frac{1}{2} \|v\|_{\mathcal{V}}^2 dt$  reduces to  $\mathcal{A}[q^1]$  and is also equal to the minimum of

$$\inf_{v \in C([0, 1], \mathcal{V}) : \phi^v((q^1, q^2)(0), t) = (q^1, q^2)(t) \forall t \in [0, 1]} \int_0^1 \frac{1}{2} \|v\|_{\mathcal{V}}^2 dt$$

over  $q_2 \in C^1([0, 1], \mathcal{X}^{N_2})$  such that  $q_2(0) = X^2$ .  $\square$

**Proposition 3.17.** *For an arbitrary trajectory  $t \rightarrow q^1(t)$  let  $q^2$  be a minimizer of*

$$\nu \mathcal{A}[(q^1, q^2)] + \ell((Y^1, Y^2), (q^1, q^2)(1)). \quad (3.30)$$

Write  $q := (q^1, q^2)$ ,  $Y = (Y^1, Y^2)$  and  $p = (p^1, p^2) := \Gamma(q, q)^{-1}\dot{q}$ . It holds true that  $(q, p)$  is a solution of the dynamical system

$$\begin{cases} \dot{q}^2 &= \Gamma(q^2, q^1)p^1 + \Gamma(q^2, q^2)p^2 \\ \dot{p}^2 &= -\frac{1}{2}\partial_{q^2}p^T\Gamma(q, q)p \\ p^1 &= \Gamma(q^1, q^1)^{-1}(\dot{q}^1 - \Gamma(q^1, q^2)p^2) \end{cases} \quad (3.31)$$

with the boundary condition

$$\nu p^2(1) + \partial_{q^2(1)}\ell(Y, q(1)) = 0. \quad (3.32)$$

Furthermore, if  $\partial_{q^2(1)}\ell(Y, q(1)) = 0$  then  $p^2(t) = 0$  for all  $t \in [0, 1]$ , and (3.31) reduces to  $\dot{q}^2 = \Gamma(q^2, q^1)\Gamma(q^1, q^1)^{-1}\dot{q}^1$  (which corresponds to (3.23)).

*Proof.* The result follows by zeroing the Fréchet derivative of (3.30) with respect to  $q^2$ . Note that  $(p^1, p^2) := \Gamma(q, q)^{-1}\dot{q}$  implies  $\dot{q}^1 = \Gamma(q^1, q^1)p^1 + \Gamma(q^1, q^2)p^2$  which allows us to identify  $p^1$  as in the third line of (3.31). Note that if  $\partial_{q^2(1)}\ell(Y, q(1)) = 0$  then (3.32) implies  $p^2(1) = 0$  and the second equation of (3.31) implies  $p^2(t) = 0$  for all  $t \in [0, 1]$ .  $\square$

**3.12. Hydrodynamic/mean-field limit.** (3.16) and (3.15) are, as in the ensemble analysis of gradient descent [51, 75], natural candidates for a hydrodynamic/mean-field limit analysis. Indeed, using (as in Subsec. 1.5) the change of variables  $p_j = \frac{1}{N}\bar{p}_j$ , (3.16) and the Hamiltonian system (3.15) are equivalent to (1.16) and (3.20) is equivalent to (1.17). Let  $\mu_N := \frac{1}{N}\sum_{i=1}^N \delta_{(q_i, \bar{p}_i)}$  be the empirical distribution of the particles  $(q_i, \bar{p}_i)$ . Then by (3.15) the average of a test function  $f(\tilde{q}, \tilde{p})$  against  $\mu_N$  obeys the dynamic

$$\frac{d}{dt}\mu_N[f] = \mu_N\left[\partial_{\tilde{q}}f\psi^T(\tilde{q}) - \partial_{\tilde{p}}f\partial_x(\tilde{p}^T\psi^T(x))\Big|_{x=\tilde{q}}\right]\mu_N[\psi(\tilde{q})\tilde{p}], \quad (3.33)$$

which leads to the following theorem.

**Theorem 3.18.** *If, as  $N \rightarrow \infty$ ,  $\mu_N$  and its first-order derivatives weakly converge towards  $\mu$  then minimizers of (3.19) converge to the solution of  $\dot{\phi}(x, t) = \psi^T(\phi^v(x, t))\mu[\psi(\tilde{q})\tilde{p}]$  and*

$$\partial_t\mu = \left[-\operatorname{div}_{\tilde{q}}(\mu\psi^T(\tilde{q})) + \operatorname{div}_{\tilde{p}}(\mu\partial_x(\tilde{p}^T\psi^T(x))\Big|_{x=\tilde{q}})\right]\mu[\psi(\tilde{q})\tilde{p}]. \quad (3.34)$$

**3.13. Numerical implementation as geodesic shooting.** Given the picture depicted in Fig. 3, the geodesic shooting solution to Problem 1 is summarized in the pseudo-algorithm (1). Except for the structure of the end loss  $\ell$ , this method is similar to the one introduced for computational anatomy [56, 1] (where the constraint  $\dot{q} = \Gamma(q, q)p$  may also be relaxed [14]). This section will implement this strategy on the Swiss roll dataset to illustrate the impact of the value of  $\nu$  on the deformation of the space and the sparsity of momentum variables.

**3.13.1. Geometric integration.** The Hamiltonian system (3.13) is characterized by structural and geometric invariants (the canonical symplectic form, volumes in the phase space, the energy, etc.) [48]. Symplectic integrators [31] have been developed to approximate the continuous system while exactly (e.g., for the symplectic form) or nearly (e.g., for the energy) preserving these invariants. The main idea of these integrators is to



**Algorithm 1** Shooting solution to Problem 1

- 
- 1: Define the loss  $\ell$  via (2.11) or (2.16).
  - 2: Discretize the Hamiltonian system (3.13) with a stable and accurate integrator.
  - 3: Minimize (3.17) (via gradient descent and the discretized Hamiltonian system) to identify the initial momentum  $p(0)$ .
  - 4: Approximate  $f^\dagger$  with  $f^\ddagger(\cdot) = f \circ \phi(\cdot, 1)$  where  $\phi$  is obtained from the numerical integration of (1.14) (using the solution of the discretized Hamiltonian system with optimal initial momentum  $p(0)$ ) and  $f$  is obtained as the minimizer of  $\ell(\phi(X, 1), Y)$  in (2.11) or (2.16).
- 

simulate a nearby discrete mechanical system rather than a nearby discrete ODE. Within this class of symplectic integrators, explicit ones are preferred for their computational efficiency/tractability. Since the Hamiltonian (3.13) is nonseparable, classical symplectic integrators [31] such as Störmer-Verlet are implicit [30]. Although the Euler-Lagrange scheme associated with the discrete least action principle (3.6) is symplectic (since it is variational [49]), it is also implicit and therefore difficult to simulate. In this paper we will simply discretize the Hamiltonian system with the Leapfrog method [31] as follows:

$$\begin{cases} p & \leftarrow p - \frac{h}{2} \partial_q \left( \frac{1}{2} p^T \Gamma(q, q) p \right) \\ q & \leftarrow q + h \Gamma(q, q) p \\ p & \leftarrow p - \frac{h}{2} \partial_q \left( \frac{1}{2} p^T \Gamma(q, q) p \right). \end{cases} \quad (3.35)$$

For simplicity, although (3.35) is not symplectic for our non-separable system (3.13), it is explicit, time-reversible and sufficiently stable for our example is.

**Remark 3.19.** *M. Tao has recently introduced [86] an ingenious method for deriving explicit symplectic integrators for general non-separable Hamiltonian systems that could be employed for (3.13). Tao's idea is to consider the augmented Hamiltonian system*

$$\bar{\mathfrak{H}}(q, p, \bar{q}, \bar{p}) = \mathfrak{H}(q, \bar{p}) + \mathfrak{H}(\bar{q}, p) + \omega(\|q - \bar{q}\|_2^2 + \|p - \bar{p}\|_2^2) \quad (3.36)$$

*in which the first two terms are copies of the original system with mixed-up positions and momenta and the last term is an artificial restraint ( $\omega$  is a constant controlling the binding of the two copies). Discretizing (3.36) via Strang splitting leads to explicit symplectic integrators of arbitrary even order.*

**3.13.2. Swiss roll dataset.** We implement the pseudo-algorithm (1) for the Swiss roll dataset illustrated in Fig. 4. We use the optimal recovery loss (2.11) to define  $\ell$  and  $f$ . For this example  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{Y} = \mathbb{R}$ ,  $N = 200$ ,  $Y_i = +1$  for the first 100 points, and  $Y_i = -1$  for the remaining 100 points.  $\Gamma$  is a separable Gaussian kernel (with a nugget  $r$ ) of the form  $\Gamma(z, z') = (k(z, z') + r)I$  with  $k(z, z') = e^{-|z-z'|^2/s^2}$  for  $z, z' \in \mathcal{X}$ ,  $s = 5$  and  $r = 0.1$ . We simply take  $K$  to be the scalar kernel  $k(z, z') + r$  with the same parameters as for  $\Gamma$ . Fig. 4 shows the locations of the points  $q_i(t)$  for  $i = 1, \dots, 200$  which is a solution of the numerical discretization of the Hamiltonian system (3.13) with the Leapfrog method (3.35) and  $h = 0.2$ . This Hamiltonian system is initialized with the momentum  $p(0)$  identified by minimizing (3.17) via gradient descent for three different values of  $\nu$  of the

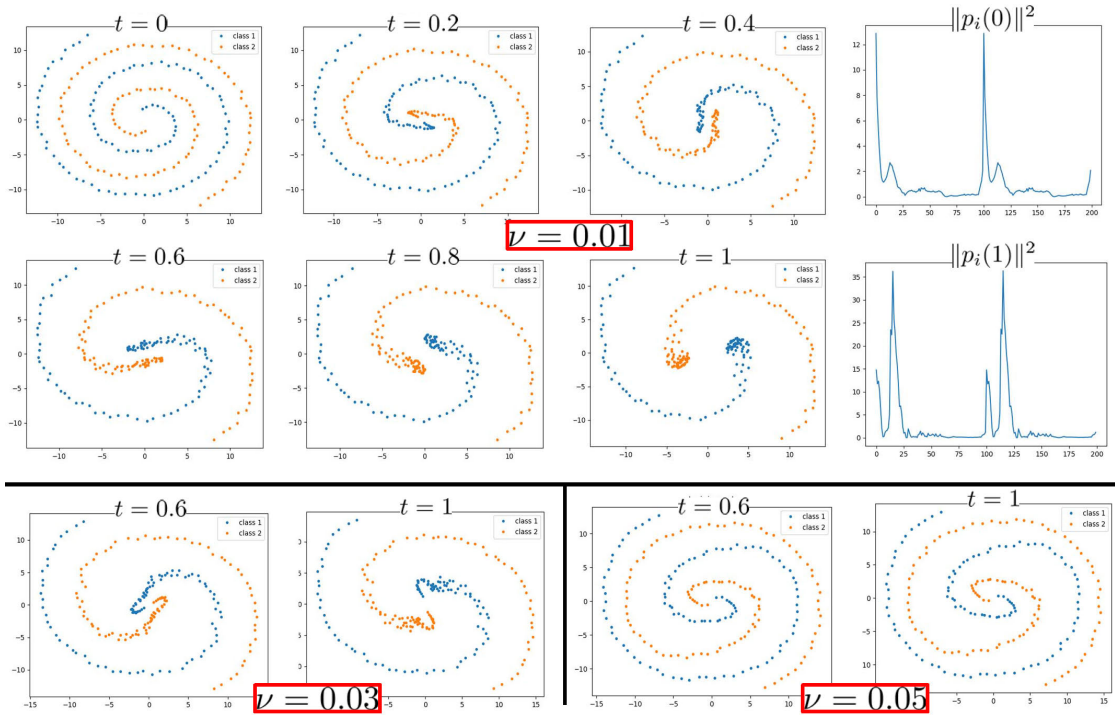


FIGURE 4. Swiss roll data set. Locations of the points  $q_i(t)$  for  $\nu = 0.01$  (top)  $\nu = 0.03$  (bottom left) and  $\nu = 0.05$  (bottom right).

regularizing parameter balancing, in (3.4), the RKHS norm of the deformation of the space with that of the regressor  $f$ . Note that as  $\nu$  increases a greater penalty is placed on that deformation and the points  $q_i(1)$  remain closer to their original position. On the other hand for a small value of  $\nu$ , the space will deform to a greater degree to minimize the RKHS norm of the regressor. Fig. 4 is also showing the norm of the entries of initial momentum  $p(0)$  and final momentum  $p(1)$ . As discussed in Subsec. 3.11 the domination of a few entries supports the suggestion that momentum variables promote sparsity in the representation of the regressor.

#### 4. Regularization

The landmark matching [40] setting of Sec. 3 requires non-overlapping data, and minimizers, and minimal values obtained from that setting may depend non-continuously on the input  $X$  (since  $\Gamma(X, X)$  will become singular as  $X_i \rightarrow X_j$  for some  $i \neq j$ ). To ensure continuity and avoid singularities, idea registration must be regularized as it is commonly done in image registration [53]. The proposed regularization provides an alternative to Dropout for ANNs [82].

**4.1. Regularized mechanical regression.** In the setting of Subsec. 3.1, let  $\ell_{\mathcal{Y}} : \mathcal{Y}^N \times \mathcal{Y}^N \rightarrow [0, \infty]$  be an arbitrary continuous positive loss. The proposed regularized mechanical regression solution to Problem 1 is to approximate  $f^\dagger$  by a function of

the form  $f^\ddagger = f \circ \phi_L = (3.1)$  where  $\phi_L = (I + v_L) \circ \dots \circ (I + v_1) = (3.2)$ , and  $(v_1, \dots, v_L, f)$  are identified by minimizing the following regularized version of (3.3).

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} L \sum_{s=1}^L (\|v_s\|_{\mathcal{Y}}^2 + \frac{1}{r} \|q^{s+1} - (I + v_s)(q^s)\|_{\mathcal{X}^N}^2) \\ & + \lambda (\|f\|_{\mathcal{H}}^2 + \frac{1}{\rho} \|f(q^{L+1}) - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & v_1, \dots, v_L \in \mathcal{V}, f \in \mathcal{H}, q^1, \dots, q^{L+1} \in \mathcal{X}^N, q^1 = X, Y' \in \mathcal{X}^N, \end{cases} \quad (4.1)$$

where  $r, \rho > 0$  are regularization parameters (akin to the nuggets employed in Kriging/spatial statistics [79]) and  $\|X\|_{\mathcal{X}^N} := \sum_{i=1}^N \|X_i\|_{\mathcal{X}}^2$ ,  $\|Y\|_{\mathcal{Y}^N} := \sum_{i=1}^N \|Y_i\|_{\mathcal{Y}}^2$ . With this regularization Cond. 3.1 and 3.2 can, as described below, be relaxed to the following conditions, which we assume to hold true in this section.

**Condition 4.1.** Assume that (1)  $\mathcal{X}$  and  $\mathcal{Y}$  are finite-dimensional (2)  $x \rightarrow K(x, x')$  is continuous for all  $x'$ , and (3)  $(x, x') \rightarrow \Gamma(x, x')$  and its first and second order partial derivatives are continuous and uniformly bounded.

**4.2. Reduction to a discrete least action principle.** Minimizing in  $f$  and  $Y'$  first, (4.1) is equivalent to

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} L \sum_{s=1}^L (\|v_s\|_{\mathcal{Y}}^2 + \frac{1}{r} \|q^{s+1} - (I + v_s)(q^s)\|_{\mathcal{X}^N}^2) + \ell(q^{L+1}, Y) \\ \text{over} & v_1, \dots, v_L \in \mathcal{V}, q^1, \dots, q^{L+1} \in \mathcal{X}^N, q^1 = X. \end{cases} \quad (4.2)$$

where  $\ell : \mathcal{X}^N \times \mathcal{Y}^N \rightarrow [0, \infty]$  is the loss defined by

$$\ell(X', Y) := \begin{cases} \text{Minimize} & \lambda (\|f\|_{\mathcal{H}}^2 + \frac{1}{\rho} \|f(X') - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & f \in \mathcal{H}, Y' \in \mathcal{Y}^N. \end{cases} \quad (4.3)$$

For  $q \in \mathcal{X}^N$ , write  $\Gamma_r(q, q) := \Gamma(q, q) + rI$  and  $K_\rho(q, q) := K(q, q) + \rho I$  for the  $N \times N$  block operator matrices with blocks  $\Gamma(q_i, q_j) + r\delta_{i,j}I_{\mathcal{X}}$  and  $K(X_i, X_j) + \rho\delta_{i,j}I_{\mathcal{Y}}$  (writing  $I_{\mathcal{X}}$  ( $I_{\mathcal{Y}}$ ) for the identity operator on  $\mathcal{X}$  ( $\mathcal{Y}$ )).

**Theorem 4.2.**  $(v_1, \dots, v_L, f)$ , is a minimizer of (4.1) if and only if

$$v_s(x) = \Gamma(x, q^s) \Gamma_r(q^s, q^s)^{-1} (q^{s+1} - q^s) \text{ for } x \in \mathcal{X}, s \in \{1, \dots, L\}, \quad (4.4)$$

where  $q^1, \dots, q^{L+1} \in \mathcal{X}^N$  is a minimizer of (write  $\Delta t := 1/L$ )

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \sum_{s=1}^L (\frac{q^{s+1} - q^s}{\Delta t})^T \Gamma_r(q^s, q^s)^{-1} (\frac{q^{s+1} - q^s}{\Delta t}) \Delta t + \ell(q^{L+1}, Y) \\ \text{over} & q^2, \dots, q^{L+1} \in \mathcal{X}^N \text{ with } q^1 = X, \end{cases} \quad (4.5)$$

and  $f$  is a minimizer of (4.3) defining  $\ell(q^{L+1}, Y)$ . Furthermore  $f$  is a minimizer of (4.3) defining  $\ell(X', Y)$  if and only if

$$f(\cdot) = K(\cdot, X')Z \quad (4.6)$$

where  $Z$  is a minimizer of

$$\ell(X', Y) = \inf_{Z \in \mathcal{Y}^N} \lambda Z^T K_\rho(X', X')Z + \ell_{\mathcal{Y}}(K(X, X)Z, Y) = (4.3). \quad (4.7)$$

Finally, under Cond. 4.1,  $\ell = (4.3) = (4.7)$  is continuous (in both arguments), positive and admits a minimizer  $f \in \mathcal{H}$  that is unique if  $Y' \rightarrow \ell_{\mathcal{Y}}(Y', Y)$  is convex.

*Proof.* (2.20) implies (4.4). The representer theorem implies (4.6). Using (2.21) we get (4.5) and (4.7) from  $\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + \frac{1}{\rho} \|f(X') - Y'\|_{\mathcal{Y}^N}^2 = (Y')^T K_\rho(X', X')^{-1} Y'$  and  $\min_{v_s \in \mathcal{V}} (\|v_s\|_{\mathcal{V}}^2 + \frac{1}{r} \|q^{s+1} - (I + v_s)(q^s)\|_{\mathcal{X}^N}^2) = (q^{s+1} - q^s)^T \Gamma_r(q^s, q^s)^{-1} (q^{s+1} - q^s)$ .  $Z^T K_\rho(X', X') Z \geq \rho Z^T Z$  ensures that the variable  $Z$  in (4.7) can be restricted to live in a compact set. The continuity of  $\ell$  then follows from [84, Lem. 5.3,5.4] and uniqueness follows from the (strict) convexity of (4.7) in  $Z$ .  $\square$

**4.3. Continuous least action principle and Hamiltonian system.** The continuous limit of the discrete least action principle (4.5) is

$$\begin{cases} \text{Minimize} & \nu \mathcal{A}_r[q] + \ell(q(1), Y) \\ \text{over} & q \in C^1([0, 1], \mathcal{X}^N) \text{ subject to } q(0) = X, \end{cases} \quad (4.8)$$

where  $\mathcal{A}_r$  is the continuous action defined by

$$\mathcal{A}_r[q] := \int_0^1 \frac{1}{2} \dot{q}^T \Gamma_r(q, q)^{-1} \dot{q} dt, \quad (4.9)$$

whose regularized Lagrangian  $\mathfrak{L}_r(q, \dot{q}) = \frac{1}{2} \dot{q}^T \Gamma_r(q, q)^{-1} \dot{q}$  is identical to (3.10) with  $\Gamma(q, q)$  replaced by  $\Gamma_r(q, q)$ . We will now show that all the results of Sec. 3 remain true under regularization and the relaxed conditions 4.1 with  $\Gamma(q, q)$  replaced by  $\Gamma_r(q, q)$ .

**Theorem 4.3.** *For  $q \in C^1([0, 1], \mathcal{X}^N)$  introduce the momentum*

$$p = \Gamma_r(q, q)^{-1} \dot{q}. \quad (4.10)$$

*$q$  is minimizer of (4.8) if and only if  $q(0) = X$ ,  $(q, p)$  follows the Hamiltonian dynamic*

$$\begin{cases} \dot{q} = \Gamma_r(q, q)p \\ \dot{p} = -\partial_q(\frac{1}{2}p^T \Gamma_r(q, q)p), \end{cases} \quad (4.11)$$

*defined by the regularized Hamiltonian  $\mathfrak{H}_r(q, p) = \frac{1}{2}p^T \Gamma_r(q, q)p$ , and  $p(0)$  is a minimizer of*

$$\mathfrak{W}^r(p(0), X, Y) := \frac{\nu}{2} p^T(0) \Gamma_r(X, X) p(0) + \ell(q(1), Y). \quad (4.12)$$

*Furthermore, (4.11) admits a unique solution in  $C^2([0, 1], \mathcal{X}^N) \times C^1([0, 1], \mathcal{X}^N)$ , the energy  $\mathfrak{H}_r(q, p) = \frac{1}{2}p^T \Gamma_r(q, q)p$  is an invariant of the dynamic, and  $p(1)$  satisfies (3.18).*

*Proof.* The proof is identical to those presented in Sec. 3. Since  $rp^T p \leq p^T \Gamma_r(q, q)p$ , (4.12) and energy preservation imply that  $p(t)$  is confined to a compact set.  $\square$

**4.4. Regularized idea registration.** In the limit  $L \rightarrow \infty$ ,  $(I + v_k) \circ \dots \circ (I + v_1)$  (obtained from (4.2)) approximates (at time  $t_k := \frac{k}{L}$ ) the flow map  $\phi^v$  (defined as the solution of (1.7)) where  $v$  is a minimizer of

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 (\|v\|_{\mathcal{V}}^2 + \frac{1}{r} \|\dot{q} - v(q, t)\|_{\mathcal{X}^N}^2) dt + \ell(q(1), Y) \\ \text{over} & v \in C([0, 1], \mathcal{V}), q \in C^1([0, 1], \mathcal{X}^N), q(0) = X. \end{cases} \quad (4.13)$$

The proof of the following theorem is identical to that of Thm. 3.8.

**Theorem 4.4.**  $v$  is a minimizer of (4.13) if and only if

$$v(x, t) = \Gamma(x, q(t))\Gamma_r(q(t), q(t))^{-1}\dot{q}(t) \quad (4.14)$$

where  $q$  is a minimizer of the least action principle (4.8). Furthermore the value of (4.13) after minimization over  $v$  is (4.8) and the representer ODE (3.20) can be written as in (1.14) where  $(q, p)$  is the solution of the Hamiltonian system (4.11) with initial condition  $q(0) = X$  and  $p(0)$  identified as a minimizer of (4.12).

**4.5. Existence/identification of minimizers and energy preservation.** As in Subsec. 3.8, minimizers of (4.13) may not be unique. The proof of the following theorem is identical to that of Thm. 3.9.

**Theorem 4.5.** The minimum values of (4.8), (4.12) and (4.13) are identical. (4.8), (4.12) and (4.13) have minimizers.  $q$  is a minimizer of (4.8) if and only if  $(q, p)$  ( $p = \Gamma_r(q, q)^{-1}\dot{q}$ ) follows the Hamiltonian dynamic (4.11) (with  $q(0) = X$ ) and  $p(0) = \Gamma_r(q(0), q(0))^{-1}\dot{q}(0)$  is a minimizer of  $\mathfrak{V}^r(p(0), X, Y) = (4.12)$ .  $v$  is a minimizer of (4.13) if and only if

$$v(x, t) = \Gamma(x, q(t))p(t) \quad (4.15)$$

with  $(q, p)$  following the Hamiltonian dynamic (4.11) (with  $q(0) = X$ ) and  $p(0)$  being a minimizer of  $\mathfrak{V}^r(p(0), X, Y) = (4.12)$ . Therefore the minimizers of (4.8) and (4.13) can be parameterized by their initial momentum  $p(0)$ , identified as a minimizer of  $\mathfrak{V}^r(p(0), X, Y) = (4.12)$ . Furthermore, at those minima, the energies  $\frac{1}{2}p\Gamma_r(q, q)p = \frac{1}{2}\dot{q}^T\Gamma_r(q, q)^{-1}\dot{q} = \frac{1}{2}(\|v\|_{\mathcal{V}}^2 + \frac{1}{r}\|\dot{q} - v(q, t)\|_{\mathcal{X}^N}^2) = \frac{1}{2}\|v\|_{\mathcal{V}_r}^2$  (writing  $\|\cdot\|_{\mathcal{V}_r}$  for the RKHS norm defined by the kernel  $\Gamma(x, x') + r\delta(x - x')I$ ) are constant over  $t \in [0, 1]$  and equal to  $\frac{\nu}{2}p^T(0)\Gamma_r(X, X)p(0)$ .

As in Sec. 3 the trajectory  $q^1, \dots, q^{L+1}$  of a minimizer of (4.5) follows

$$\begin{cases} q^{s+1} &= q^s + \Delta t \Gamma_r(q^s, q^s)p^s \\ p^{s+1} &= p^s - \frac{\Delta t}{2} \partial_{q^{s+1}}((p^{s+1})^T \Gamma_r(q^{s+1}, q^{s+1})p^{s+1}), \end{cases} \quad (4.16)$$

where  $p^s$  is the momentum

$$p^s = \Gamma_r(q^s, q^s)^{-1} \frac{q^{s+1} - q^s}{\Delta t}. \quad (4.17)$$

Write

$$\mathfrak{V}_L^r(p^1, X, Y) := \begin{cases} \frac{\nu}{2} \sum_{s=1}^L (p^s)^T \Gamma_r(q^s, q^s) p^s \Delta t + \ell(q^{L+1}, Y) \\ p^s = (4.17) \text{ and } (q^s, p^s) \text{ follow (4.16) with } q^1 = X. \end{cases} \quad (4.18)$$

The proof of the following theorem is identical to that of Thm. 3.10.

**Theorem 4.6.** The minimum values of (4.2), (4.5) and  $\mathfrak{V}_L^r(p^1, X, Y)$  (in  $p^1$ ) are identical. (4.2), (4.5) and (4.18) have minimizers.  $q^1, \dots, q^{L+1}$  is a minimizer of (4.5) if and only if  $(q^s, p^s)$  (with  $p^s = (4.17)$ ) follows the discrete Hamiltonian map (4.16),  $q^1 = X$  and  $p^1$  is a minimizer of  $\mathfrak{V}_L^r(p^1, X, Y) = (4.18)$ .  $v_1, \dots, v_L$  is a minimizer of (4.2) if and only if

$$v_s(x) = \Gamma(x, q^s)p^s = (4.4), \quad (4.19)$$

where  $(q^s, p^s)$  follows the discrete Hamiltonian map (4.16) with  $q^1 = X$  and  $p^1$  is a minimizer of  $\mathfrak{V}_L^r(p^1, X, Y) = (4.18)$ . Therefore the minimizers of (4.2) and (4.5) can be parameterized by their initial momentum identified as a minimizer of  $\mathfrak{V}_L^r(p^1, X, Y) = (4.18)$ . At those minima, the energies  $\frac{1}{2}(p^s)^T \Gamma_r(q^s, q^s) p^s$  and  $\frac{1}{2} \|v_s\|_{\mathcal{V}_r}^2 = \frac{1}{2} (\|v_s\|_{\mathcal{V}}^2 + \frac{1}{r} \|q^{s+1} - (I + v_s)(q^s)\|_{\mathcal{X}^N}^2)$  are equal and fluctuate by at most  $\mathcal{O}(1/L)$  over  $s \in \{1, \dots, L\}$ .

**4.6. Continuity of minimal values.** The following theorem does not have an equivalent in Sec. 3 since it does not hold true without regularization.

**Theorem 4.7.** *The minimal values of (4.8), (4.12), (4.13), (4.2), (4.5) and (4.18) are continuous in  $(X, Y)$ .*

*Proof.* By Thm. 4.5 it is then sufficient (for the discrete setting) to prove the continuity of the minimum value of (4.12) with respect to  $(X, Y)$ . By [4, Thm. 1.4.1] if  $(q, p)$  follows the Hamiltonian dynamic (4.11) then, under Cond. 4.1,  $q(1)$  is continuous with respect to  $p(0)$ . The continuity of  $\ell$  then implies that of (4.12) with respect to  $p(0)$  (with  $q(1)$  being a function of  $p(0)$ ). Under Cond. 4.1  $p(0)$  can be restricted to a compact set in the minimization of (4.12). [84, Lem. 5.3,5.4] then implies the continuity of the minimum value of (4.12) with respect to  $(X, Y)$ . By Thm. 4.6 the continuity of the minimum value of (4.1) follows from that of  $\mathfrak{V}_L^r(p^1, X, Y) = (4.18)$  which is continuous in all variables. Under Cond. 4.1,  $p^1$  can be restricted to a compact set in the minimization of (4.18). We conclude by using [84, Lem. 5.3,5.4].  $\square$

**4.7. Convergence of minimal values and minimizers.** Write  $\mathfrak{M}_L^r(X, Y)$  for the set of minimizers  $p^1$  of  $\mathfrak{V}_L^r(p^1, X, Y) = (4.18)$ . Write  $\mathfrak{M}^r(X, Y)$  for the set of minimizers  $p(0)$  of  $\mathfrak{V}^r(p(0), X, Y) = (4.12)$ . The proof of the following theorem is identical to that of Thm. 3.11.

**Theorem 4.8.** *The common minimal value of (4.2), (4.5) and (4.18) converge, as  $L \rightarrow \infty$ , towards the common minimal value of (4.8), (4.12), (4.13). As  $L \rightarrow \infty$ , the set of adherence values of  $\mathfrak{M}_L^r(X, Y)$  is  $\mathfrak{M}^r(X, Y)$ . Let  $v_s^L, q_L^s$  and  $p_L^s$  be sequences of minimizers of (4.2), (4.5) and (4.18) indexed by the same sequence  $p_L^1$  of initial momentum in  $\mathfrak{M}_L^r(X, Y)$  (as described in Thm. 4.6). Then, the adherence points of the sequence  $p_L^1$  are in  $\mathfrak{M}^r(X, Y)$  and if  $p(0)$  is such a point ( $p_L^1$  converges towards  $p(0)$  along a subsequence  $L_k$ ) then, along that subsequence: (1) The trajectory formed by interpolating the states  $q_L^s \in \mathcal{X}^N$  converges to the trajectory formed by a minimizer of (4.8) with initial momentum  $p(0)$ . (2) For  $t \in [0, 1]$ ,  $(I + v_{int(tL)}^L) \circ \dots \circ (I + v_1^L)(x)$  converges to  $\phi^v(x, t) = (1.7)$  where  $v$  is a minimizer of (4.13) with initial momentum  $p(0)$ . Conversely if  $p(0) \in \mathfrak{M}^r(X, Y)$  then it is the limit of a sequence  $p_L^1 \in \mathfrak{M}_L^r(X, Y)$  and the minimizers of (4.2), (4.5) and (4.18) with initial momentum  $p_L^1$  converge (in the sense given above) to the minimizers of (4.8), (4.12), (4.13) with initial momentum  $p(0)$  (as described in Thm. 4.5).*

Note that the continuous limit  $L \rightarrow \infty$  solution to Problem 1 is to approximate  $f^\dagger$  with  $f^\ddagger = f \circ \phi^v(\cdot, 1)$  where  $f$  and  $v$  are minimizers of



$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 (\|v\|_{\mathcal{V}}^2 + \frac{1}{r} \|\dot{q} - v(q, t)\|_{\mathcal{X}^N}^2) dt \\ & + \lambda (\|f\|_{\mathcal{H}}^2 + \frac{1}{\rho} \|f(q(1)) - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & v \in C([0, 1], \mathcal{V}), q \in C^1([0, 1], \mathcal{X}^N), q(0) = X, f \in \mathcal{H}, Y' \in \mathcal{Y}^N. \end{cases} \quad (4.20)$$

## 5. Idea formation

We will now compose ridge regression, mechanical regression, and idea registration blocks across layers of abstraction between  $\mathcal{X}$  and  $\mathcal{Y}$ . The resulting input/output functions generalize ANNs and deep kernel learning [92].

**5.1. Block diagram representation.** For ease of presentation and conceptual simplicity, we will first summarize Sec. 4 in block diagram representation.

**5.1.1. Mechanical regression.** Given  $p_0 \in \mathcal{X}^N$ ,  $X \in \mathcal{X}^N$ ,  $x \in \mathcal{X}$ , let  $(q^s, p^s)$  be the solution of (4.16) with initial value  $(q^1, p^1) = (X, p_0)$ , let  $v_s = \Gamma(\cdot, q^s)p^s = (4.19)$  and set  $x' = (I + v_L) \circ \dots \circ (I + v_1)(x)$ ,  $X' = q^{L+1}$  and  $\mathfrak{V} = \frac{1}{2} \sum_{s=1}^L (p^s)^T \Gamma_r(q^s, q^s) p^s \Delta t$  with  $\Delta t = 1/L$ . We represent the corresponding multivariate function  $(x, X, p_0) \rightarrow (x', X', \mathfrak{V})$  with the diagram

$$\begin{array}{ccc} & p_0 & \\ & \downarrow & \\ X \rightarrow & \boxed{\Gamma \mid r \mid L} & \rightarrow X' \\ x \rightarrow & & \rightarrow x' \\ & \downarrow \mathfrak{V} & \end{array} \quad (5.1)$$

A deformation  $x' = \phi_L(x) = (I + v_L) \circ \dots \circ (I + v_1)(x)$  obtained by minimizing (4.2) must be (Thm. 4.6) of the form (5.1) where  $\mathfrak{V}$  is  $\frac{1}{2} L \sum_{s=1}^L (\|v_s\|_{\mathcal{V}}^2 + \frac{1}{r} \|q^{s+1} - (I + v_s)(q^s)\|_{\mathcal{X}^N}^2)$ . Furthermore (by the proof of Thm. 4.7) (5.1) is uniformly continuous in  $x, X, p_0$  if  $p_0$  is restricted to a compact set, and  $\mathfrak{V}$  diverges uniformly towards  $\infty$  as  $p_0^T p_0 \rightarrow \infty$ .

**5.1.2. Idea registration.** Given  $p_0 \in \mathcal{X}^N$ ,  $X \in \mathcal{X}^N$ ,  $x \in \mathcal{X}$ , let  $(q(t), p(t))$  be the solution of (4.11) with initial value  $(q(0), p(0)) = (X, p_0)$ , let  $v(\cdot, t) = \Gamma(\cdot, q)p = (4.15)$  and set  $x' = \phi^v(x, 1)$  (where  $\phi^v$  is defined as the solution of (1.7)),  $X' = q(1)$  and  $\mathfrak{V} = \frac{1}{2} p^T(0) \Gamma_r(X, X) p(0)$ . We represent the corresponding multivariate function with the following diagram

$$\begin{array}{ccc} & p_0 & \\ & \downarrow & \\ X \rightarrow & \boxed{\Gamma \mid r \mid \infty} & \rightarrow X' \\ x \rightarrow & & \rightarrow x' \\ & \downarrow \mathfrak{V} & \end{array} \quad (5.2)$$

A deformation  $\phi^v$  obtained by minimizing (4.13) must be (Thm. 4.5) of the form (5.2) and  $\mathfrak{V}$  is  $\frac{1}{2} \int_0^1 (\|v\|_{\mathcal{V}}^2 + \frac{1}{r} \|\dot{q} - v(q, t)\|_{\mathcal{X}^N}^2) dt$ . Furthermore, if  $p_0$  is restricted to a compact set then (5.2) is uniformly continuous in  $x, X, p_0$  and (5.1) converges<sup>23</sup> uniformly towards (5.2).

<sup>23</sup>By Thm. 4.6, given same inputs, all the outputs of (5.1) converge to the outputs of (5.2).

**5.1.3. Ridge regression.** Given  $Z \in \mathcal{X}^N$ ,  $X \in \mathcal{X}^N$  and  $x \in \mathcal{X}$ , set  $Y' = K(X, X)Z$ ,  $y = K(x, X)Z$  and  $\mathfrak{W} = Z^T K_\rho(X, X)Z$ . We represent the corresponding multivariate function with the following diagram

$$\begin{array}{ccc} & Z & \\ & \downarrow & \\ X \rightarrow & \boxed{K \mid \rho} & \rightarrow Y' \\ x \rightarrow & \downarrow & \\ & \mathfrak{W} & \end{array} . \quad (5.3)$$

A function  $f$  minimizing (4.3) must be of the form (5.3) and  $\mathfrak{W}$  is the value of  $\|f\|_{\mathcal{H}}^2 + \frac{1}{\rho}\|f(X) - Y'\|_{\mathcal{Y}^N}^2$ .

**5.1.4. Composing blocks.** Using  $\begin{array}{ccc} Y' & \rightarrow & \boxed{\ell_{\mathcal{Y}}} \\ Y & \rightarrow & \rightarrow \ell_{\mathcal{Y}} \end{array}$  for the block diagram representation of the loss  $\ell_{\mathcal{Y}}(Y', Y)$ , mechanical regression can be represented with the diagram

$$\begin{array}{ccccccc} & p_0 & & Z & & & \\ & \downarrow & & \downarrow & & & \\ X \rightarrow & \boxed{\Gamma \mid r \mid L} & \rightarrow & \boxed{K \mid \rho} & \rightarrow & Y & \rightarrow \boxed{\ell_{\mathcal{Y}}} \rightarrow \ell_{\mathcal{Y}} \\ x \rightarrow & \downarrow & & \downarrow & & & \\ & \mathfrak{W} & & \mathfrak{W} & & & \end{array} , \quad (5.4)$$

and idea registration can be represented with the diagram

$$\begin{array}{ccccccc} & p_0 & & Z & & & \\ & \downarrow & & \downarrow & & & \\ X \rightarrow & \boxed{\Gamma \mid r \mid \infty} & \rightarrow & \boxed{K \mid \rho} & \rightarrow & Y & \rightarrow \boxed{\ell_{\mathcal{Y}}} \rightarrow \ell_{\mathcal{Y}} \\ x \rightarrow & \downarrow & & \downarrow & & & \\ & \mathfrak{W} & & \mathfrak{W} & & & \end{array} . \quad (5.5)$$

For both diagrams  $p_0$  and  $Z$  are identified as minimizers of the Total Loss  $= \nu\mathfrak{W} + \lambda\mathfrak{W} + \ell_{\mathcal{Y}}$  and are contained in a set that is closed and uniformly bounded (in  $X$ ). As  $L \rightarrow \infty$ , (5.4) converges uniformly to (5.5), the adherence values of the minimizers of (5.4) are the minimizers of (5.5), the total loss of (5.4) converges to that of (5.5).

**5.2. Idea formation.** Let  $\mathcal{X}_0, \dots, \mathcal{X}_{D+1}$  be finite-dimensional Hilbert spaces, with  $\mathcal{X}_0 = \mathcal{X}$  and  $\mathcal{X}_{D+1} = \mathcal{Y}$ . Let  $\ell_{\mathcal{Y}}$  be a loss function on  $\mathcal{Y}$  as in Subsec. 2.3.2. Let  $\nu_1, \dots, \nu_D$  and  $\lambda_0, \dots, \lambda_D$  be strictly positive parameters. Let  $\mathcal{V}_m$  and  $\mathcal{H}_m$  be RKHS defined by operator-valued kernels  $\Gamma^m : \mathcal{X}^m \times \mathcal{X}^m \rightarrow \mathcal{L}(\mathcal{X}^m)$  and  $K^m : \mathcal{X}^m \times \mathcal{X}^m \rightarrow \mathcal{L}(\mathcal{X}^{m+1})$  satisfying the regularity conditions 4.1. Let  $L_1, \dots, L_D$  be strictly positive integers. The discrete hierarchical mechanical regression solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition

$$F_{m+1} = f_m \circ \phi^m(F_m) \text{ with } \phi^m = (I + v^{m, L_m}) \circ \dots \circ (I + v^{m, 1}) \text{ and } F_1 = f_0, \quad (5.6)$$

where the  $v_{m,j}$  are  $f_m$  are minimizers of

$$\left\{ \begin{array}{l} \text{Min} \quad \lambda_0 (\|f_0\|_{\mathcal{H}_0}^2 + \frac{1}{\rho} \|f_0(X) - q^{1,1}\|_{\mathcal{X}_1^N}^2) \\ \quad + \sum_{m=1}^D \left( \frac{\nu_m L_m}{2} \sum_{j=1}^{L_m} (\|v_{m,j}\|_{\mathcal{V}_m}^2 + \frac{1}{r} \|q^{m,j+1} - (I + v_{m,j})(q^{m,j})\|_{\mathcal{X}_m^N}^2) \right. \\ \quad \left. + \lambda_m (\|f_m\|_{\mathcal{H}_m}^2 + \frac{1}{\rho} \|f_m(q^{m, L_m+1}) - q^{m+1,1}\|_{\mathcal{X}_{m+1}^N}^2) \right) + \ell_{\mathcal{Y}}(q^{D+1,1}, Y) \\ \text{over} \quad v_{m,j} \in \mathcal{V}_m, f_m \in \mathcal{H}_m, q^{m,j} \in \mathcal{X}_m^N. \end{array} \right. \quad (5.7)$$

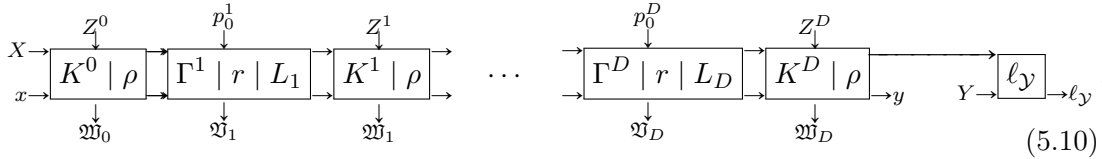
In the continuous limit  $\min_m L_m \rightarrow \infty$ , the hierarchical mechanical regression solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition

$$F_{m+1} = f_m \circ \phi^{v_m}(F_m, 1) \text{ with } F_1 = f_0, \quad (5.8)$$

where the  $v_m$  are  $f_m$  are minimizers of

$$\left\{ \begin{array}{l} \text{Min} \quad \lambda_0 (\|f_0\|_{\mathcal{H}_0}^2 + \frac{1}{\rho} \|f_0(X) - q^1(0)\|_{\mathcal{X}_1^N}^2) \\ \quad + \sum_{m=1}^D \left( \frac{\nu_m}{2} \int_0^1 (\|v_m(\cdot, t)\|_{\mathcal{V}_m}^2 + \frac{1}{r} \|\dot{q}^m - v_m(q^m, t)\|_{\mathcal{X}_m^N}^2) dt \\ \quad + \lambda_m (\|f_m\|_{\mathcal{H}_m}^2 + \frac{1}{\rho} \|f_m(q^m(1)) - q^{m+1}(0)\|_{\mathcal{X}_{m+1}^N}^2) \right) + \ell_Y(q^{D+1}(0), Y) \\ \text{over} \quad v_m \in C([0, 1], \mathcal{V}_m), f_m \in \mathcal{H}_m, q^m \in C([0, 1], \mathcal{X}_m^N). \end{array} \right. \quad (5.9)$$

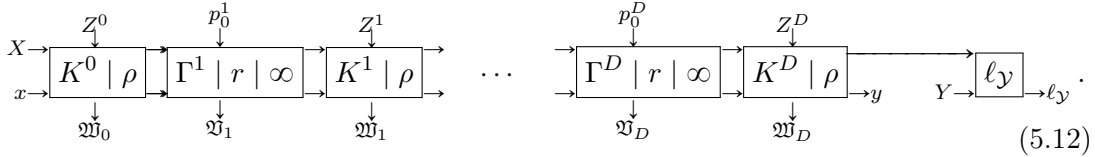
**Theorem 5.1.** *The map  $y = F_{D+1}(x)$  obtained from (5.6) is equal to the output  $y$  produced by the block diagram*



where the initial momenta  $p_0^m$  and  $Z^m$  are identified as minimizers of the total loss

$$\text{Total loss} = \lambda_0 \mathfrak{W}_0 + \sum_{m=1}^D (\nu_m \mathfrak{W}_m + \lambda_m \mathfrak{W}_m) + \ell_Y. \quad (5.11)$$

All the minimizers  $p_0^m$  and  $Z^m$  of (5.10) are contained in a compact set. The map  $y = F_{D+1}(x)$  obtained from (5.8) is equal to the output  $y$  produced by the block diagram



where the initial momenta  $p_0^m$  and  $Z^m$  are identified as minimizers of (5.11). All the minimizers  $p_0^m$  and  $Z^m$  of (5.12) are contained in a compact set. The multivariate input/output maps (5.10) and (5.12) are uniformly continuous (for  $p_0^m$  and  $Z^m$  in compact sets). As  $\min_m L_m \rightarrow \infty$ , (1) the multivariate input/output map (5.10) converges uniformly (for  $p_0^m$  and  $Z^m$  in compact sets) to the multivariate input/output map (5.12) (2) The minimal value of total loss of (5.10) converges to the minimal value of total loss of (5.12) (3) The adherence values of the momenta ( $p_0^m$  and  $Z^m$ ) minimizing (5.10) is the set of momenta minimizing (5.12) (4) The adherence values of  $F_{D+1}$  obtained from (5.10) is the set of  $F_{D+1}$  obtained from (5.12).

*Proof.* The proof is a direct consequence of the results of Sec. 4 summarized in Subsec. 5.1. Note that for  $r, \rho > 0$ , (5.11) diverges towards infinity as  $\max_m (p_0^m)^T p_0^m + \max_m (Z^m)^T Z^m \rightarrow \infty$ . Therefore the search for minimizers can be restricted to a compact set.  $\square$

**5.3. Further reduction.** Minimizing over  $f_m$  and  $v_m$ , (5.9) reduces (as in Sec. 4) to

$$\left\{ \begin{array}{l} \text{Min} \quad \lambda_0 (q^1(0))^T K_\rho^0(X, X)^{-1} q^1(0) + \sum_{m=1}^D \left( \frac{\nu_m}{2} \int_0^1 \dot{q}^m \Gamma_r^m(q^m, q^m)^{-1} \dot{q}^m dt \right. \\ \quad \left. + \lambda_m (q^{m+1}(0))^T K_\rho^m(q^m(1), q^m(1))^{-1} q^{m+1}(0) \right) + \ell_Y(q^{D+1}(0), Y) \\ \text{over} \quad q^m \in C([0, 1], \mathcal{X}_m^N). \end{array} \right. \quad (5.13)$$

Introduce the momentum variables  $p^m = \Gamma_r(q^m, q^m)^{-1}\dot{q}^m$ . Taking the Fréchet derivative of (5.13) with respect to  $q^m$ , implies that (for  $1 \leq m \leq D$ )  $(q^m, p^m)$  satisfies the Hamiltonian dynamic (4.11) (with  $\Gamma_r$  replaced by  $\Gamma_r^m$ ) and the boundary equations

$$\begin{cases} 2\lambda_{m-1}K_\rho^{m-1}(q^{m-1}(1), q^{m-1}(1))^{-1}q^m(0) - \nu_m p^m(0) & = 0 \\ \nu_m p^m(1) + \lambda_m \partial_{q^m(1)}((q^{m+1}(0))^T K_\rho^m(q^m(1), q^m(1))^{-1}q^{m+1}(0)) & = 0 \end{cases} \quad (5.14)$$

We deduce (Prop. 5.2) that, in the search for minimizers of (5.12), the initial momenta  $p_0^m = p^m(0)$  can be expressed as explicit functions of the  $Z^{m-1}$  and the  $Z_m$  can be expressed as implicit functions of the  $Z^{m-1}$ . Therefore, the search for minimizers of (5.12) could, in theory, be reduced to a shooting method (the selection of the initial momentum  $Z^0$ ).

**Proposition 5.2.** *In the setting of Thm. 5.1 the minimizers of (5.12) satisfy  $q^1(0) = K^0(X, X)Z^0$  and (for  $m \in \{1, \dots, D\}$ )*

$$q^m(0) = K^{m-1}(q^{m-1}(1), q^{m-1}(1))Z^{m-1}, \quad (5.15)$$

$$p^m(0) = 2\frac{\lambda_{m-1}}{\nu_m}K_\rho^{m-1}(q^{m-1}(1), q^{m-1}(1))^{-1}K^{m-1}(q^{m-1}(1), q^{m-1}(1))Z^{m-1}, \quad (5.16)$$

$$\partial_x((Z^m)^T K^m(q^m(1), q^m(1))K_\rho^m(x, x)^{-1}K^m(q^m(1), q^m(1))Z^m) \Big|_{x=q^m(1)} = -\frac{\nu_m}{\lambda_m}p^m(1). \quad (5.17)$$

*Proof.* (5.15) follows from (5.3). Combining (5.14) with (5.15) implies (5.16) and (5.17).  $\square$

Let us now consider the discrete setting. Minimizing over  $v_{m,j}$  and  $f_m$  (5.18) reduces to

$$\begin{cases} \text{Min} & \lambda_0(q^{1,1})^T K_\rho^0(X, X)^{-1}q^{1,1} + \sum_{m=1}^D \left( \frac{\nu_m L_m}{2} \sum_{j=1}^{L_m} (q^{m,j+1} - q^{m,j})^T \Gamma_r^m(q^{m,j}, q^{m,j})^{-1} \right. \\ & \left. (q^{m,j+1} - q^{m,j}) + \lambda_m (q^{m+1,1})^T K_\rho^m(q^{m,L_m+1}, q^{m,L_m+1})^{-1}q^{m+1,1} \right) + \ell_Y(q^{D+1,1}, Y) \\ \text{over} & q^{m,j} \in \mathcal{X}_m^N. \end{cases} \quad (5.18)$$

Introduce the discrete momenta  $p^{m,j} = L_m \Gamma_r^m(q^{m,j}, q^{m,j})^{-1}(q^{m,j+1} - q^{m,j})$ . Taking the Fréchet derivative of (5.18) with respect to  $q^{m,j}$  implies that  $(q^{m,j}, p^{m,j})$  satisfies the discrete Hamiltonian dynamic (4.16) (with  $\Gamma_r$  replaced by  $\Gamma_r^m$  and  $\Delta t = 1/L_m$ ) and the boundary equations presented in the following proposition (that is the analogue of Prop. 5.2).

**Proposition 5.3.** *In the setting of Thm. 5.1 the minimizers of (5.10) satisfy  $q^{1,1} = K^0(X, X)Z^0$  and (for  $m \in \{1, \dots, D\}$ )*

$$q^{m,1} = K^{m-1}(q^{m-1, L_{m-1}+1}, q^{m-1, L_{m-1}+1})Z^{m-1} \quad (5.19)$$

$$p^{m,1} - \frac{1}{2L_m} \partial_{q^m} (p^{m,1})^T \Gamma(q^{m,1}, q^{m,1}) p^{m,1} = 2\frac{\lambda_{m-1}}{\nu_m} K_\rho^{m-1}(q^{m-1, L_{m-1}+1}, q^{m-1, L_{m-1}+1})^{-1} q^{m,1} \quad (5.20)$$

$$\nu_m p^{m, L_m} + \lambda_m \partial_{q^{m, L_m+1}} ((q^{m+1,1})^T K_\rho^m(q^{m, L_m+1}, q^{m, L_m+1})^{-1} q^{m+1,1}) = 0 \quad (5.21)$$

## 6. With feature maps and activation functions

Image registration is based on two main strategies [33, 1]: (1) discretize  $v : \mathcal{X} \times [0, 1] \rightarrow \mathcal{X}$  on a space/time mesh and minimize (3.19); or (2) simulate the Hamiltonian system (3.13). Although these strategies work well in computational anatomy where the dimension of  $\mathcal{X}$  is 2 or 3, and the number of landmark points small, they are not suitable for industrial scale machine learning where the dimension of  $\mathcal{X}$  and the number of data points are large.

**6.1. With feature maps.** Feature space representations of kernels can overcome these limitations and lead to numerical schemes that include and generalize those currently used in deep learning.

**6.1.1. Mechanical regression.** Let  $\mathcal{F}$  and  $\psi : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{X}, \mathcal{F})$  be a feature space and map associated with the kernel  $\Gamma$  of the RKHS  $\mathcal{V}$  in (3.4) and (3.19). We will now work under Cond. 4.1. The following theorems reformulate the least action principles and Hamiltonian dynamics of Sec. 4 in the feature map setting of Subsec. 2.2.

**Theorem 6.1.**  $\alpha_1, \dots, \alpha_L, q^1, \dots, q^{L+1}$  minimize

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} L \sum_{s=1}^L (\|\alpha_s\|_{\mathcal{F}}^2 + \frac{1}{r} \|q^{s+1} - q^s - \psi^T(q^s)\alpha_s\|_{\mathcal{X}^N}^2) + \ell(q^{L+1}, Y) \\ \text{over} & \alpha_1, \dots, \alpha_L \in \mathcal{F}, q^2, \dots, q^{L+1} \in \mathcal{X}^N, q^1 = X, \end{cases} \quad (6.1)$$

if and only if the  $v_s = \psi^T \alpha_s, q^s$  minimize (4.2). Furthermore  $\|\alpha_s\|_{\mathcal{F}}^2 + \frac{1}{r} \|q^{s+1} - q^s - \psi^T(q^s)\alpha_s\|_{\mathcal{X}^N}^2$  fluctuates by at most  $\mathcal{O}(1/L)$  over  $s$ .

*Proof.* The proof is a simple consequence of Thm. 2.5 and Thm. 4.6.  $\square$

**Theorem 6.2.**  $\alpha \in C([0, 1], \mathcal{F})$  and  $q \in C^1([0, 1], \mathcal{X}^N)$  minimize

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 (\|\alpha(t)\|_{\mathcal{F}}^2 + \frac{1}{r} \|\dot{q}(t) - \psi^T(q(t))\alpha(t)\|_{\mathcal{X}^N}^2) dt + \ell(q(1), \phi^v(X, 1)) \\ \text{over} & \alpha \in C([0, 1], \mathcal{F}), q \in C^1([0, 1], \mathcal{X}^N), q(0) = X, \end{cases} \quad (6.2)$$

if and only if  $v(\cdot, t) = \psi^T(\cdot)\alpha(t)$  and  $q(t)$  minimize (4.13). Furthermore, at the minimum,  $\|\alpha(t)\|_{\mathcal{F}}^2 + \frac{1}{r} \|\dot{q}(t) - \psi^T(q(t))\alpha(t)\|_{\mathcal{X}^N}^2$  is constant over  $t \in [0, 1]$ .

*Proof.* The proof is a simple consequence of Thm. 2.5 and Thm. 4.5.  $\square$

Let  $\mathcal{F}_2$  and  $\psi_2 : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y}, \mathcal{F}_2)$  be a feature map and space associated with the RKHS  $\mathcal{H}$  in the loss (4.3). The following is a direct corollary of Thm. 6.2.

**Corollary 6.3.** The maps  $v = \psi^T \alpha$  and  $f = \psi_2^T \alpha_2$  obtained from minimizing

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 (\|\alpha(t)\|_{\mathcal{F}}^2 + \frac{1}{r} \|\dot{q}(t) - \psi^T(q(t))\alpha(t)\|_{\mathcal{X}^N}^2) dt \\ & + \lambda (\|\alpha_2\|_{\mathcal{F}_2}^2 + \frac{1}{\rho} \|\psi_2^T(q(1))\alpha_2 - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & \alpha \in C([0, 1], \mathcal{F}), \alpha_2 \in \mathcal{F}_2, q \in C^1([0, 1], \mathcal{X}^N), q(0) = X, Y' \in \mathcal{Y}^N. \end{cases} \quad (6.3)$$

are identical to those obtained by minimizing (4.20).

**6.2. With feature maps of scalar operator-valued kernels.** We will now describe mechanical regression with the feature maps of scalar operator-valued kernels.

**6.2.1. Feature maps of scalar operator-valued kernels.** We will first describe feature spaces and maps of scalar operator-valued kernels in the setting of Sec 2. Let  $K(x, x') = k(x, x')I_{\mathcal{Y}}$  be as in Definition 2.2 and write  $\mathfrak{F}$  and  $\varphi : \mathcal{X} \rightarrow \mathfrak{F}$  for a feature space and map associated with the scalar-valued kernel  $k$ . Write  $\mathcal{L}(\mathfrak{F}, \mathcal{Y})$  for the space of bounded linear operators from  $\mathfrak{F}$  to  $\mathcal{Y}$ . For  $\beta \in \mathfrak{F}$  and  $y \in \mathcal{Y}$  write  $y\beta^T \in \mathcal{L}(\mathfrak{F}, \mathcal{Y})$  for the outer product between  $y$  and  $\beta$  defined as the linear function mapping  $\beta' \in \mathfrak{F}$  to  $y\langle\beta, \beta'\rangle_{\mathfrak{F}} \in \mathcal{Y}$ .

**Theorem 6.4.** *A feature space of the operator-valued kernel  $K$  is  $\mathcal{F} := \mathcal{L}(\mathfrak{F}, \mathcal{Y})$  and its feature map is defined by*

$$\psi(x)y = y\varphi^T(x) \text{ for } (x, y) \in \mathcal{X} \times \mathcal{Y}. \quad (6.4)$$

Furthermore,

$$\psi^T(x)\alpha = \alpha\varphi(x) \text{ for } x \in \mathcal{X} \text{ and } \alpha \in \mathcal{L}(\mathfrak{F}, \mathcal{Y}), \quad (6.5)$$

and, writing  $\mathcal{H}$  for the RKHS defined by  $K$ ,

$$\|\alpha\varphi(\cdot)\|_{\mathcal{H}}^2 = \|\alpha\|_{\mathcal{F}}^2 = \|\alpha\|_{\mathcal{L}(\mathfrak{F}, \mathcal{Y})}^2 = \text{Tr}[\alpha^T\alpha]. \quad (6.6)$$

where  $\text{Tr}$  is the trace operator.

*Proof.* Write  $\psi$  and  $\mathcal{F}$  for a feature map/space associated with  $K$ . The identity

$$y^T K(x, x')y' = \langle\psi(x)y, \psi(x')y'\rangle_{\mathcal{F}} = \langle\varphi(x), \varphi(x')\rangle_{\mathfrak{F}}\langle y, y'\rangle_{\mathcal{Y}}, \quad (6.7)$$

implies (6.4) and the identification of  $\mathcal{F}$  with  $\mathcal{L}(\mathfrak{F}, \mathcal{Y})$  endowed with inner product

$$\left\langle \sum_{i,j} c_{i,j} y_i \beta_j^T, \sum_{i',j'} c'_{i',j'} y_{i'} (\beta_{j'}^T)^T \right\rangle_{\mathcal{F}} = \sum_{i,i',j,j'} c_{i,j} c'_{i',j'} \langle\beta_j, \beta_{j'}\rangle_{\mathfrak{F}} \langle y_i, y_{i'}\rangle_{\mathcal{Y}}. \quad (6.8)$$

Thm. 2.5 implies the first identity in (6.6). (6.8) combined with the matrix representation of  $\alpha \in \mathcal{L}(\mathfrak{F}, \mathcal{Y})$  over bases of  $\mathcal{Y}$  and  $\mathfrak{F}$  imply the last equality in (6.6).  $\square$

**6.2.2. Mechanical regression.** Now consider the setting of (6.3) in the situation where  $\Gamma$  and  $K$  (the kernels associated with  $\mathcal{V}$  and  $\mathcal{H}$  in the derivation of (6.3)) are scalar, i.e.  $\Gamma(z, z') = k(z, z')I_{\mathcal{X}}$  and  $K(z, z') = k_2(z, z')I_{\mathcal{Y}}$  and write  $\mathfrak{F}$ ,  $\mathfrak{F}_2$ ,  $\varphi$  and  $\varphi_2$  for feature spaces and maps associated with  $k$  and  $k_2$ . The following proposition follows from Subsec. 6.2.1.

**Proposition 6.5.** *The maps  $v(\cdot, t) = \alpha(t)\varphi(\cdot)$  and  $f = \alpha_2\varphi_2$  obtained by minimizing*

$$\begin{cases} \text{Minimize} & \frac{\nu}{2} \int_0^1 (\|\alpha(t)\|_{\mathcal{L}(\mathfrak{F}, \mathcal{X})}^2 + \frac{1}{r} \|\dot{q}(t) - \alpha(t)\varphi(q(t))\|_{\mathcal{X}^N}^2) dt \\ & + \lambda (\|\alpha_2\|_{\mathcal{F}_2}^2 + \frac{1}{\rho} \|\alpha_2\varphi_2(q(1)) - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & \alpha \in C([0, 1], \mathcal{F}), \alpha_2 \in \mathcal{F}_2, q \in C^1([0, 1], \mathcal{X}^N), q(0) = X, Y' \in \mathcal{Y}^N. \end{cases} \quad (6.9)$$

are identical to those obtained by minimizing (6.3).

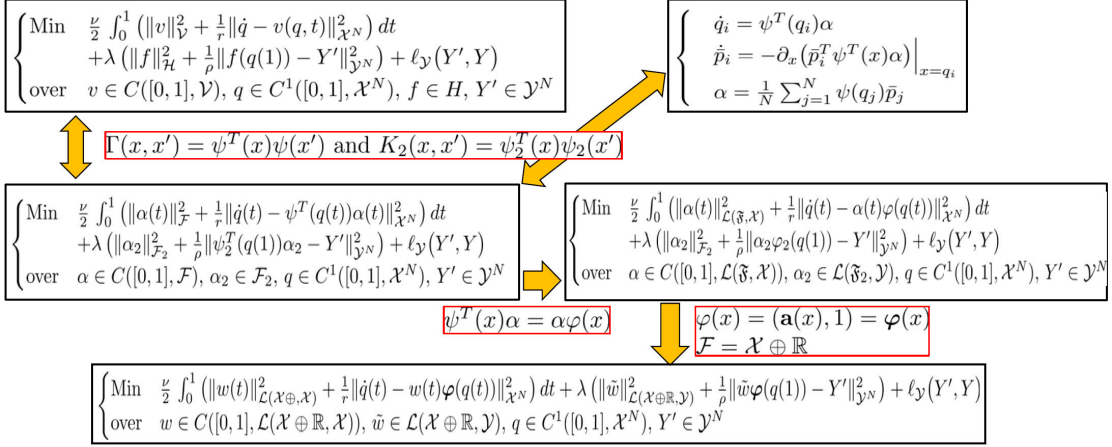


FIGURE 5. Mechanical regression with feature maps and activation functions.

**6.3. With activation functions.** Let  $\mathfrak{F} = \mathfrak{F}' \oplus \mathfrak{F}''$  where  $\mathfrak{F}'$  and  $\mathfrak{F}''$  are  $\langle \cdot, \cdot \rangle_{\mathfrak{F}}$ -orthogonal separable Hilbert sub-spaces of  $\mathfrak{F}$ . Let  $A \in \mathcal{L}(\mathcal{X}, \mathfrak{F}')$  be a bounded linear operator from  $\mathcal{X}$  to  $\mathfrak{F}'$  such that<sup>24</sup>  $A^T A = I$ ,  $c \in \mathfrak{F}''$  such that  $c^T c = 1$ , and  $\varphi : \mathcal{X} \rightarrow \mathfrak{F}' \oplus \mathfrak{F}''$  defined by

$$\varphi(x) = A\mathbf{a}(x) + c, \quad (6.10)$$

where  $\mathbf{a} : \mathcal{X} \rightarrow \mathcal{X}$ , is an arbitrary nonlinear **activation function**.

**Proposition 6.6.** *The operator-valued kernel defined by (6.10) is  $\Gamma(x, x') = (\mathbf{a}(x)^T \mathbf{a}(x') + 1)I$ . In particular  $\Gamma$  satisfies Cond. 4.1 if  $x \rightarrow \mathbf{a}(x)$  and its first and second order partial derivatives are continuous and uniformly bounded.*

**Remark 6.7.** *We call  $\mathbf{a}$  an **elementwise nonlinearity** if  $\mathbf{a}(x) = \sum_{i=1}^{d_{\mathcal{X}}} e_i \bar{\mathbf{a}}(x_i)$  for  $x = \sum_{i=1}^{d_{\mathcal{X}}} x_i e_i \in \mathcal{X}$  where  $e_1, \dots, e_{d_{\mathcal{X}}}$  is some basis of  $\mathcal{X}$  and  $\bar{\mathbf{a}}$  is a scalar-valued (nonlinear) function. To satisfy the regularity requirements of Prop. 6.6 we must then assume that  $z \rightarrow \bar{\mathbf{a}}(z)$ ,  $\partial_z \bar{\mathbf{a}}(z)$ ,  $\partial_z^2 \bar{\mathbf{a}}(z)$  are continuous and uniformly bounded. Observe that the sigmoid nonlinearity  $\bar{\mathbf{a}}(z) = \tanh(z)$  satisfies these regularity conditions. Although  $\text{ReLU}(\bar{\mathbf{a}}(z)) = \max(z, 0)$  is not bounded and lacks the required regularity one could use a bounded and smoothed version such as the following variant of softplus  $\bar{\mathbf{a}}(z) = \ln(1 + e^z)/(1 + \epsilon \ln(1 + e^z))$  which behaves like ReLU for  $z \in (-\infty, 1/\epsilon)$  and  $0 < \epsilon \ll 1$ . For ease of presentation, we will also write  $\mathbf{a}$  for  $\bar{\mathbf{a}}$  when  $\mathbf{a}$  is an elementwise nonlinearity.*

Prop. 6.6 implies that, as long as  $A$  and  $c$  are unitary, their particular choice has no influence on the kernel  $\Gamma$ . We will therefore from now on, in the setting of activation functions, select  $\mathcal{F} = \mathcal{X} \oplus \mathbb{R}$  ( $\mathfrak{F}' = \mathcal{X}$  and  $\mathfrak{F}'' = \mathbb{R}$ ) and use the identity matrix/vector for  $A$  and  $c$ . (6.10) can then be written

$$\varphi(x) = \varphi(x) \text{ with } \varphi(x) = (\mathbf{a}(x), 1), \quad (6.11)$$

<sup>24</sup> $\dim(\mathfrak{F}') \geq \dim(\mathcal{X})$  suffices for the existence of such an  $A$ .



and  $\mathbf{a}$  is, from now on, assumed to satisfy the regularity conditions of Prop. 6.6. Similarly we select  $\mathfrak{F}_2 = \mathcal{X} \oplus \mathbb{R}$  and

$$\varphi_2(x) = \varphi(x) \text{ with } \varphi(x) = (\mathbf{a}(x), 1). \quad (6.12)$$

Note that the operator-valued kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  defined by (6.12) is

$$K(x, x') = \varphi^T(x)\varphi(x') = (\mathbf{a}^T(x)\mathbf{a}(x') + 1)I_{\mathcal{Y}}. \quad (6.13)$$

Also note that for  $\tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})$  we have  $\tilde{w}\varphi(x) = W\mathbf{a}(x) + b$  where the **weight**  $W \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$  is defined by  $Wz = \tilde{w}(z, 0)$  for  $z \in \mathcal{X}$  and the **bias**  $b \in \mathcal{Y}$  is defined by  $b = \tilde{w}(0, 1)$ . Therefore (6.12) allows us to incorporate weights and biases into a single variable  $\tilde{w}$ .

Write  $\|\cdot\|_{\mathcal{L}(\mathcal{X}, \mathcal{Y})}$  for the Frobenius norm on  $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ . The following theorem shows that mechanical regression with feature maps (6.11) and (6.12) can be expressed as a ResNet with  $L_2$  regularization on weights and biases. The following two theorems are straightforward, and Fig. 5 summarises the results of this section.

**Theorem 6.8.** *If  $\varphi$  and  $\varphi_2$  are as in (6.11) and (6.12) then  $v(\cdot, t) = w(t)\varphi(\cdot)$ ,  $f = \tilde{w}\varphi(\cdot)$  and  $q, Y'$  obtained by minimizing*

$$\begin{cases} \text{Min} & \frac{\nu}{2} \int_0^1 (\|w(t)\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \frac{1}{r} \|\dot{q}(t) - w(t)\varphi(q(t))\|_{\mathcal{X}^N}^2) dt + \\ & \lambda (\|\tilde{w}\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})}^2 + \frac{1}{\rho} \|\tilde{w}\varphi(q(1)) - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & w \in C([0, 1], \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})), \tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y}), \\ & q \in C^1([0, 1], \mathcal{X}^N), q(0) = X, Y' \in \mathcal{Y}^N, \end{cases} \quad (6.14)$$

are identical to those obtained by minimizing (6.9). Therefore (6.14) has minimizers and if  $w, q$  are minimizers of (6.14) then the energy  $\frac{1}{2}(\|w(t)\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \frac{1}{r} \|\dot{q}(t) - w(t)\varphi(q(t))\|_{\mathcal{X}^N}^2)$  is constant over  $t \in [0, 1]$ .

*Proof.* The proof is straightforward. Use Thm. 4.5 for the existence of minimizers and energy preservation.  $\square$

**Theorem 6.9.** *If  $\varphi$  and  $\varphi_2$  are as in (6.11) and (6.12) then  $\phi = (I + v_L) \circ \dots \circ (I + v_1)$ ,  $v_s = w^s\varphi(\cdot)$ ,  $f = \tilde{w}\varphi(\cdot)$  and  $q^s, Y'$ , obtained by minimizing*

$$\begin{cases} \text{Min} & \frac{\nu L}{2} \sum_{s=1}^L (\|w^s\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \frac{1}{r} \|q^{s+1} - q^s - w^s\varphi(q^s)\|_{\mathcal{X}^N}^2) + \\ & \lambda (\|\tilde{w}\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})}^2 + \frac{1}{\rho} \|\tilde{w}\varphi(q^{L+1}) - Y'\|_{\mathcal{Y}^N}^2) + \ell_{\mathcal{Y}}(Y', Y) \\ \text{over} & w^s \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X}), \tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y}), q^s \in \mathcal{X}^N, Y' \in \mathcal{Y}^N, q^1 = X, \end{cases} \quad (6.15)$$

are identical to those obtained by minimizing (6.1) and, as  $L \rightarrow \infty$ , converge (in the sense of the adherence values as in Subsec. 4.7), towards those obtained by minimizing (3.19). Furthermore, (6.15) has minimizers and if the  $w^s, q^s$  are minimizers of (6.15) then the energy  $\frac{1}{2}(\|w^s\|_{\mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})}^2 + \frac{1}{r} \|q^{s+1} - q^s - w^s\varphi(q^s)\|_{\mathcal{X}^N}^2)$  fluctuates by at most  $\mathcal{O}(1/L)$  over  $s \in \{1, \dots, L\}$ .

*Proof.* The proof of the is straightforward. Convergence follows from Thm. 4.8. Near energy preservation and existence follow from Thm. 4.6.  $\square$

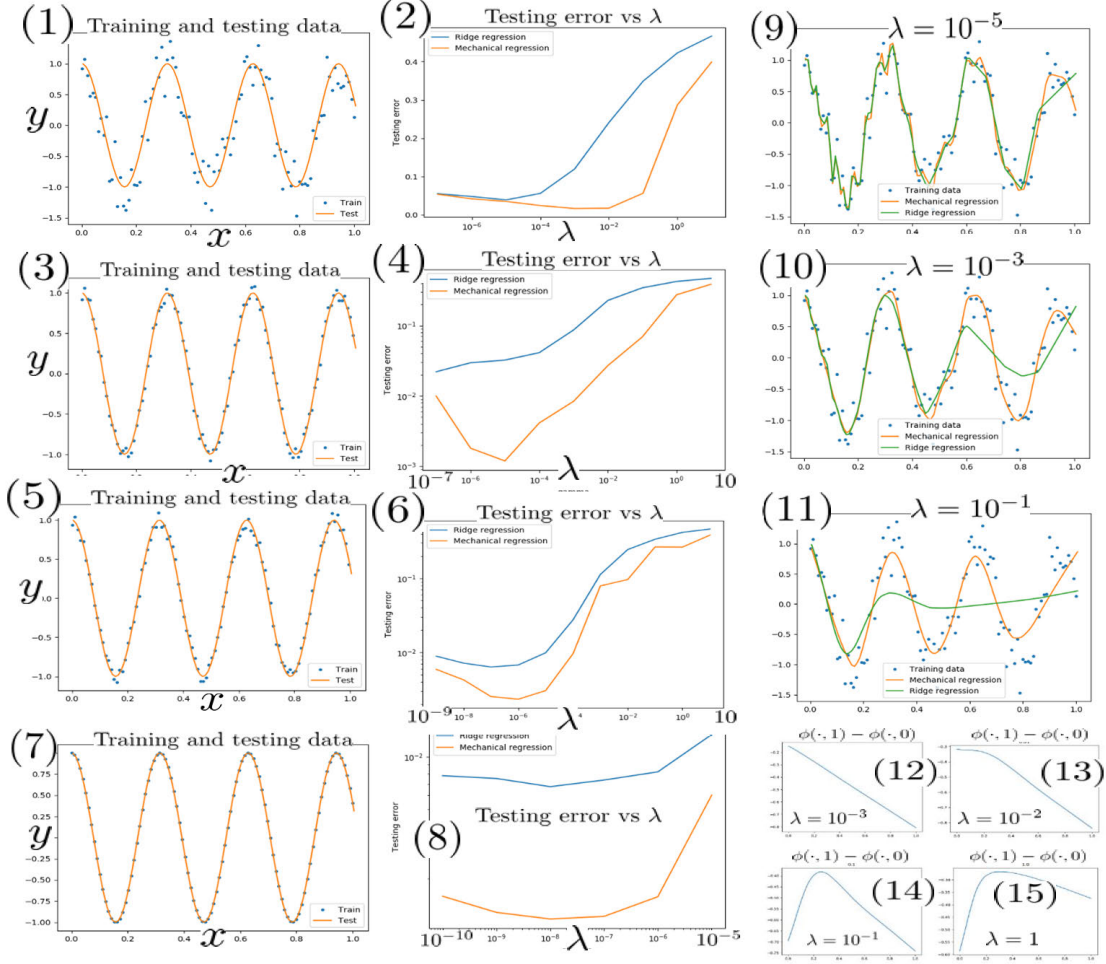


FIGURE 6. Mechanical regression vs. ridge regression. (1,3,5,7) Target function and noisy training data with  $\sigma_z = 1$  for (1),  $\sigma_z = 0.2$  for (3,5) and  $\sigma_z = 0$  for (7). (2,4,6,8) testing errors vs.  $\lambda$  corresponding to the left column for ridge regression and mechanical regression. The  $y$ -axis of (2) is in linear scale. The  $y$ -axis of (4,6,8) is in log scale. (9,10,11) ridge and mechanical regressors corresponding to (1) for  $\lambda = 10^{-5}, 10^{-3}, 10^{-1}$ . (12,13,14,15)  $\phi(\cdot, 1) - \phi(\cdot, 0)$  corresponding to mechanical regression for (1) for  $\lambda = 10^{-3}, 10^{-2}, 10^{-1}, 1$ .

**6.4. Numerical experiments.** In the following experiments we use the variational formulation (6.9) with,  $r = \rho = 0$ ,  $\ell_Y(Y', Y) = \|Y' - Y\|_{Y,N}^2$  and use random features<sup>25</sup> to construct  $\varphi$  and  $\varphi_2$ . We select  $\varphi(x) = \mathbf{a}(Wx + b)$  and  $\varphi_2(x) = \mathbf{a}(W^2x + b^2)$  with  $\mathbf{a}(\cdot) = \max(\cdot, 0)$ ,  $W \in \mathbb{R}^{\dim(\mathfrak{F}) \times \dim(\mathcal{X})}$ ,  $b \in \mathbb{R}^{\dim(\mathfrak{F})}$ ,  $W^2 \in \mathbb{R}^{\dim(\mathfrak{F}_2) \times \dim(\mathcal{X})}$ ,  $b^2 \in \mathbb{R}^{\dim(\mathfrak{F}_2)}$ . All the

<sup>25</sup>Using random features improves computational complexity without incurring significant loss in accuracy, see [72, 58].

entries of  $W, W^2, b, b^2$  are independent and we select  $W_{i,j}, W_{i,j}^2 \sim (1.5/\sqrt{\dim(\mathcal{X})})\mathcal{N}(0, 1)$  and  $b_i, b_i^2 \sim 0.1\mathcal{N}(0, 1)$ .

**6.4.1. One dimensional regression.** To goal of this experiment is to approximate the function  $f^\dagger(x) = \cos(20x)$  in the interval  $[0, 1]$  from the observation of  $N = 100$  data (training) points  $(X_i, Y_i)$  where  $X_i = i/100$ ,  $Y_i = \cos(20X_i) + \sigma_z Z_i$  and the  $Z_i$  are i.i.d. random variables uniformly distributed in  $[-0.5, 0.5]$ . Here  $\mathcal{X} = \mathcal{Y} = \mathbb{R}$  and we also use 100 points  $(x_i, y_i)_{1 \leq i \leq 100}$  to compute testing errors (we take  $x_i = i/100 - 1/200$  and  $y_i = f^\dagger(x_i)$ ). We select  $\mathfrak{F} = \mathbb{R}^{200}$  and  $\mathfrak{F}_2 = \mathbb{R}^{800}$ . Fig. 6 compares classical ridge regression ( $\nu = \infty$ ) with mechanical regression with  $\nu = 0$ . Note that mechanical regression has significantly smaller testing errors than ridge regression over a broad range of values for  $\lambda$ .

**6.4.2. MNIST and Fashion MNIST.** For this experiment we use the MNIST and Fashion MNIST datasets. We use  $N = 1000$  points  $(X_i, Y_i)$  for training and 10000 points for testing.  $f^\dagger$  maps a  $28 \times 28$  image  $X_i \in \mathbb{R}^{28 \times 28}$  to a one-hot-vector  $Y_i \in \mathbb{R}^{10}$  ( $Y_{i,j} = 1$  if the class of  $X_i$  is  $j$  and  $Y_{i,j} = 0$  otherwise). We select  $\mathfrak{F} = \mathbb{R}^{784}$  and  $\mathfrak{F}_2 = \mathbb{R}^{800}$ . Fig. 7 compares classical ridge regression ( $\nu = \infty$ ) with mechanical regression with  $\nu = 0$ . Mechanical regression has significantly smaller testing errors than ridge regression over a broad range of values for  $\lambda$ , and the deformation of the space  $\phi(\cdot, 1)$  seems to regularize the classification problem.

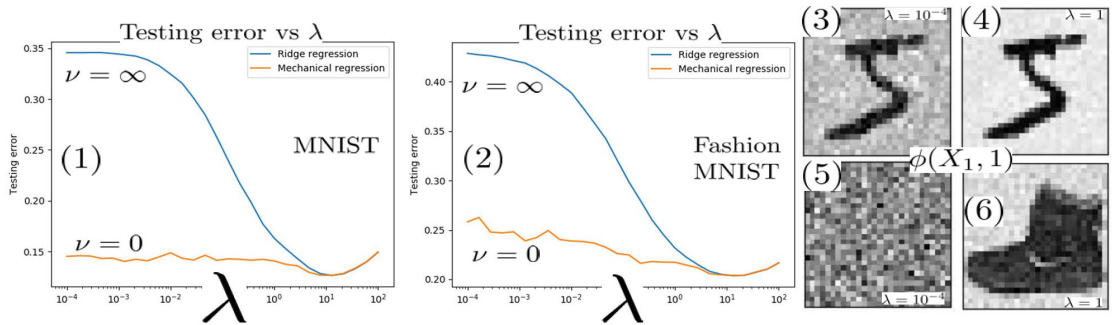


FIGURE 7. Mechanical regression vs. ridge regression. (1), (2) Testing errors vs.  $\lambda$  for MNIST and Fashion-MNIST for ridge regression ( $\nu = \infty$ ) and mechanical regression (with  $\nu = 0$ ). (3-6)  $\phi(X_1, 1)$  (3,4) MNIST (5,6) Fashion-MNIST (3,5)  $\lambda = 10^{-4}$  (4,6)  $\lambda = 1$ .

## 7. Continuous limit of ANNs

Consider the setting of Subsec. 5.2. Let  $\varphi(x) = (\mathbf{a}(x), 1)$  where  $\mathbf{a}$  is an activation function obtained as an elementwise nonlinearity satisfying the regularity conditions of Rmk. 6.7. The ANN solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition

$$F_{m+1} = \tilde{w}^m \varphi(\phi^m(F_m)) \text{ with } \phi^m = (I + w^{m, L_m} \varphi(\cdot)) \circ \dots \circ (I + w^{m, 1} \varphi(\cdot)) \text{ and } F_1 = \tilde{w}^0 \varphi(\cdot), \quad (7.1)$$

where the  $\tilde{w}^m$  and  $w^{m,j}$  are minimizers of

$$\left\{ \begin{array}{l} \text{Min} \quad \lambda_0 (\|\tilde{w}^0\|_{\mathcal{L}(\mathcal{X}_0 \oplus \mathbb{R}, \mathcal{X}_1)}^2 + \frac{1}{\rho} \|\tilde{w}^0 \varphi(X) - q^{1,1}\|_{\mathcal{X}_1^N}^2) \\ \quad + \sum_{m=1}^D \left( \frac{\nu_m L_m}{2} \sum_{j=1}^{L_m} (\|w^{m,j}\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m)}^2 + \frac{1}{r} \|q^{m,j+1} - q^{m,j} - w^{m,j} \varphi(q^{m,j})\|_{\mathcal{X}_m^N}^2) \right. \\ \quad \left. + \lambda_m (\|\tilde{w}^m\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_{m+1})}^2 + \frac{1}{\rho} \|\tilde{w}^m \varphi(q^{m, L_m+1}) - q^{m+1,1}\|_{\mathcal{X}_m^N}^2) \right) + \ell_Y(q^{D+1,1}, Y) \\ \text{over} \quad w^{m,j} \in \mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m), \tilde{w}^m \in \mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_{m+1}), q^{m,j} \in \mathcal{X}_m^N. \end{array} \right. \quad (7.2)$$

Note that in the limit  $\nu_m \rightarrow \infty$  we have  $F_{m+1} = \tilde{w}^m \varphi(F_m)$ , whereas the traditional way is to use  $h_m = \varphi(F_m)$  as variables and represent ANNs as  $h_{m+1} = \varphi(\tilde{w}^m h_m)$ . Furthermore the  $\phi^m$  represent concatenation of ResNet blocks [34]. In the continuous ( $\min_m L_m \rightarrow \infty$ ) limit, the idea formation solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition

$$F_{m+1} = \tilde{w}^m \varphi(\phi^{v_m}(F_m, 1)) \text{ with } v_m(x, t) = w^m(t) \varphi(x) \text{ and } F_1 = \tilde{w}^0 \varphi(\cdot), \quad (7.3)$$

where the  $\tilde{w}^m$  and  $w^m$  are minimizers of

$$\left\{ \begin{array}{l} \text{Min} \quad \lambda_0 (\|\tilde{w}^0\|_{\mathcal{L}(\mathcal{X}_0 \oplus \mathbb{R}, \mathcal{X}_1)}^2 + \frac{1}{\rho} \|\tilde{w}^0 \varphi(X) - q^1(0)\|_{\mathcal{X}_1^N}^2) \\ \quad + \sum_{m=1}^D \left( \frac{\nu_m}{2} \int_0^1 (\|w^m(\cdot, t)\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m)}^2 + \frac{1}{r} \|\dot{q}^m - w^m(t) \varphi(q^m(t))\|_{\mathcal{X}_m^N}^2) dt \right. \\ \quad \left. + \lambda_m (\|\tilde{w}^m\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_{m+1})}^2 + \frac{1}{\rho} \|\tilde{w}^m \varphi(q^m(1)) - q^{m+1}(0)\|_{\mathcal{X}_m^N}^2) \right) + \ell_Y(q^{D+1}(0), Y) \\ \text{over} \quad w^m \in C^1([0, 1], \mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m)), \tilde{w}^m \in \mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_{m+1}), q^m \in C^1([0, 1], \mathcal{X}_m^N). \end{array} \right. \quad (7.4)$$

The following theorem is a direct consequence of the equivalence established in Subsec. 6.3 and Thm. 5.1.

**Theorem 7.1.** (7.2) and (7.4) have minimizers. Minimal values of (7.2) and (7.4) are continuous in  $(X, Y)$ . Minimal values and minimizers  $F_{D+1}=(7.1)$  of (7.2) converge (in the sense of adherence values of Subsec. 4.7), as  $\min_m L_m \rightarrow \infty$  towards minimal values and minimizers  $F_{D+1}=(7.3)$  of (7.4). At the minima, the  $(\|w^m(\cdot, t)\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m)}^2 + \frac{1}{r} \|\dot{q}^m - w^m(t) \varphi(q^m(t))\|_{\mathcal{X}_m^N}^2)$  are constant over  $t \in [0, 1]$  and  $(\|w^{m,j}\|_{\mathcal{L}(\mathcal{X}_m \oplus \mathbb{R}, \mathcal{X}_m)}^2 + \frac{1}{r} \|q^{m,j+1} - q^{m,j} - w^{m,j} \varphi(q^{m,j})\|_{\mathcal{X}_m^N}^2)$  fluctuates by at most  $\mathcal{O}(1/L_m)$  over  $j \in \{1, \dots, L_m\}$ . Let  $\Gamma^m(x, x') = \varphi(x)^T \varphi(x') I_{\mathcal{X}_m}$ ,  $K^m(x, x') = \varphi(x)^T \varphi(x') I_{\mathcal{X}_{m+1}}$  be the kernels defined by the activation function  $\varphi$  as in (6.13). The maps  $y = F_{D+1}(x)$  obtained from (7.2) and (7.4) are equal to the output  $y$  produced by the block diagrams (5.10) and (5.12), where the initial momenta  $p_0^m$  and  $Z^m$  are identified as minimizers of the total loss (5.11). In particular the results of Thm. 5.1, Prop. 5.2 and Prop. 5.3 hold true for (7.2) and (7.4).

## 8. Deep residual Gaussian processes and error estimates

This section presents a natural [63, Sec. 7&17] extension of scalar-valued Gaussian processes to function-valued Gaussian processes (Subsec. 8.1). As with Kriging, probabilistic error estimates corresponding to the conditional standard deviation of the Gaussian process (Subsec. 8.2) are identical to the deterministic ones induced by the reproducing property of its kernel (Subsec. 8.3). As suggested in [24, p. 4] minimizers of instances of (3.19) occurring in image registration have a natural interpretation as MAP estimators of Brownian flows of diffeomorphisms [7, 42], which we will extend (Subsec. 8.4) to

the setting of function-valued GPs as deep residual Gaussian processes (that could be interpreted as a continuous variant of deep Gaussian processes [23]).

**8.1. Function-valued Gaussian processes.** The following definition of function-valued Gaussian processes is a natural extension of scalar-valued Gaussian fields as presented in [63, Sec. 7&17].

**Definition 8.1.** Let  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  be an operator-valued kernel as in Sec. 2. Let  $m$  be a function mapping  $\mathcal{X}$  to  $\mathcal{Y}$ . We call  $\xi : \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y}, \mathbf{H})$  a function-valued Gaussian process if  $\xi$  is a function mapping  $x \in \mathcal{X}$  to  $\xi(x) \in \mathcal{L}(\mathcal{Y}, \mathbf{H})$  where  $\mathbf{H}$  is a Gaussian space<sup>26</sup> and  $\mathcal{L}(\mathcal{Y}, \mathbf{H})$  is the space of bounded linear operators from  $\mathcal{Y}$  to  $\mathbf{H}$ . Abusing notations we write  $\langle \xi(x), y \rangle_{\mathcal{Y}}$  for  $\xi(x)y$ . We say that  $\xi$  has mean  $m$  and covariance kernel  $K$  and write  $\xi \sim \mathcal{N}(m, K)$  if  $\langle \xi(x), y \rangle_{\mathcal{Y}} \sim \mathcal{N}(m(x), y^T K(x, x)y)$  and

$$\text{Cov}(\langle \xi(x), y \rangle_{\mathcal{Y}}, \langle \xi(x'), y' \rangle_{\mathcal{Y}}) = y^T K(x, x')y'. \quad (8.1)$$

We say that  $\xi$  is centered if it is of zero mean.

If  $K(x, x)$  is trace class ( $\text{Tr}[K(x, x)] < \infty$ ) then  $\xi(x)$  defines a measure on  $\mathcal{Y}$  (i.e. a  $\mathcal{Y}$ -valued random variable), otherwise it only defines a (weak) cylinder-measure in the sense of Gaussian fields (see [63, Sec. 17]).

**Theorem 8.2.** The distribution of a function-valued Gaussian process is uniquely determined by its mean and covariance kernel  $K$ . Conversely given  $m$  and  $K$  there exists a function-valued Gaussian process having mean  $m$  and covariance kernel  $K$ . In particular if  $K$  has feature space  $\mathcal{F}$  and map  $\psi$ , the  $e_i$  form an orthonormal basis of  $\mathcal{F}$ , and the  $Z_i$  are i.i.d.  $\mathcal{N}(0, 1)$  random variables, then

$$\xi = m + \sum_i Z_i \psi^T e_i \quad (8.2)$$

is a function-valued GP with mean  $m$  and covariance kernel  $K$ .

*Proof.* The proof is classical, see [63, Sec. 7&17]. Note that the separability of  $\mathcal{F}$  ensures the existence of the  $e_i$ . Furthermore  $\mathbb{E}[(\xi - m)(\xi - m)^T] = \psi^T \psi = K$ .  $\square$

**8.2. Probabilistic error estimates for function-valued GP regression.** The conditional covariance of the Gaussian process  $\xi \sim \mathcal{N}(m, K)$  (conditioned on the data  $(X, Y)$ ) provides natural a priori probabilistic error estimates for the testing data. The following theorem identifies this conditional covariance kernel.

**Theorem 8.3.** Let  $\xi$  be a centered function-valued GP with covariance kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$ . Let  $X, Y \in \mathcal{X}^N \times \mathcal{Y}^N$ . Let  $Z = (Z_1, \dots, Z_N)$  be a random Gaussian vector, independent from  $\xi$ , with i.i.d.  $\mathcal{N}(0, \lambda I_{\mathcal{Y}})$  entries ( $\lambda \geq 0$  and  $I_{\mathcal{Y}}$  is the identity map on  $\mathcal{Y}$ ). Then  $\xi$  conditioned on  $\xi(X) + Z$  is a function-valued GP with mean

$$\mathbb{E}[\xi(x) | \xi(X) + Z = Y] = K(x, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1}Y = (2.20) \quad (8.3)$$

and conditional covariance operator

$$K^\perp(x, x') := K(x, x') - K(x, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1}K(X, x'). \quad (8.4)$$

<sup>26</sup>That is a Hilbert space of centered Gaussian random variables, see [63, Sec. 7&17].

In particular, if  $K$  is trace class, then

$$\sigma^2(x) := \mathbb{E} \left[ \left\| \xi(x) - \mathbb{E}[\xi(x) | \xi(X) + Z = Y] \right\|_{\mathcal{Y}}^2 \middle| \xi(X) + Z = Y \right] = \text{Tr} [K^\perp(x, x)]. \quad (8.5)$$

*Proof.* The proof is a generalization of the classical setting [63, Sec. 7&17]. Writing  $\xi^T(x)y$  for  $\langle \xi(x), y \rangle_{\mathcal{Y}}$ , observe that  $y^T \xi(x) \xi^T(x') y = y^T K(x, x') y'$  implies  $\mathbb{E}[\xi(x) \xi^T(x')] = K(x, x')$ . Since  $\xi$  and  $Z$  share the same Gaussian space the expectation of  $\xi(x)$  conditioned on  $\xi(X) + Z$  is  $A(\xi(X) + Z)$  where  $A$  is a linear map identified by  $0 = \text{Cov}(\xi(x) - A(\xi(X) + Z), \xi(X) + Z) = \mathbb{E}[\xi(x) - A(\xi(X) + Z)(\xi^T(X) + Z^T)] = K(x, X) - A(K(X, X) + \lambda I_{\mathcal{Y}})$ , which leads to  $A = K(x, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1}$  and (8.3). The conditional covariance is then given by  $K^\perp(x, x') = \mathbb{E} \left[ \left( \xi(x) - K(x, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1}(\xi(X) + Z) \right) \left( \xi(x') - K(x', X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1}(\xi(X) + Z) \right)^T \right]$  which leads to (8.4).  $\square$

**8.3. Deterministic error estimates for function-valued Kriging.** For  $\lambda = 0$ ,  $f(x) = (8.3)$  is the optimal recovery solution (2.14) to Problem 1. For  $\lambda > 0$ ,  $f(x) = (8.3)$  is the ridge regression solution (2.20) to Problem 1. The following theorem shows that the standard deviation (8.5) provides deterministic a priori error bounds on the accuracy of the ridge regressor (8.3) to  $f^\dagger$  in Problem 1. Local error estimates such as (8.6) are classical in Kriging [93] where  $\sigma^2(x)$  is known as the power function/kriging variance (see also [59][Thm. 5.1] for applications to PDEs).

**Theorem 8.4.** *Let  $f^\dagger$  be the unknown function of Problem 1 and let  $f(x) = (8.3) = (2.20)$  be its GPR/ridge regression solution. Let  $\mathcal{H}$  be the RKHS associated with  $K$  and let  $\mathcal{H}_\lambda$  be the RKHS associated with the kernel  $K_\lambda := K + \lambda I_{\mathcal{Y}}$ . It holds true that*

$$\|f^\dagger(x) - f(x)\|_{\mathcal{Y}} \leq \sigma(x) \|f^\dagger\|_{\mathcal{H}} \quad (8.6)$$

and

$$\|f^\dagger(x) - f(x)\|_{\mathcal{Y}} \leq \sqrt{\sigma^2(x) + \lambda \dim(\mathcal{Y})} \|f^\dagger\|_{\mathcal{H}_\lambda}, \quad (8.7)$$

where  $\sigma(x)$  is the standard deviation (8.5).

*Proof.* Let  $y \in \mathcal{Y}$ . Using the reproducing property (2.4) and  $Y = f^\dagger(X)$  we have

$$\begin{aligned} y^T (f^\dagger(x) - f(x)) &= y^T f^\dagger(x) - y^T K(x, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1} f^\dagger(X) \\ &= \langle f^\dagger, K(\cdot, x)y - K(\cdot, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1} K(X, x)y \rangle_{\mathcal{H}}. \end{aligned}$$

Using Cauchy-Schwartz inequality, we deduce that

$$\left| y^T (f^\dagger(x) - f(x)) \right|^2 \leq \|f^\dagger\|_{\mathcal{H}}^2 y^T K^\perp(x, x)y \quad (8.8)$$

where  $K^\perp$  is the conditional covariance (8.4). Summing over  $y$  ranging in basis of  $\mathcal{Y}$  implies (8.6). The proof of (8.7) is similar, simply observe that

$$\begin{aligned} y^T (f^\dagger(x) - f(x)) &= \langle f^\dagger, K_\lambda(\cdot, x)y - K_\lambda(\cdot, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1} K(X, x)y \rangle_{\mathcal{H}_\lambda} \\ &\leq \|f^\dagger\|_{\mathcal{H}_\lambda} \|K_\lambda(\cdot, x)y - K_\lambda(\cdot, X)(K(X, X) + \lambda I_{\mathcal{Y}})^{-1} K(X, x)y\|_{\mathcal{H}_\lambda}, \end{aligned}$$



which implies

$$\left| y^T (f^\dagger(x) - f(x)) \right|^2 \leq \|f^\dagger\|_{\mathcal{H}}^2 (\lambda y^T y + y^T K^\perp(x, x) y). \quad (8.9)$$

□

**Remark 8.5.** *Since Thm. 8.4 does not require  $\mathcal{X}$  to be finite-dimensional, its estimates do not suffer from the curse of dimensionality but from finding a good kernel for which both  $\|f^\dagger\|_{\mathcal{H}}$  and  $y^T K^\perp(x, x) y$  are small (over  $x$  sampled from the testing distribution). Indeed both (8.6) and (8.7) provide a priori deterministic error bounds on  $f^\dagger - f$  depending on the RKHS norms  $\|f^\dagger\|_{\mathcal{H}}$  and  $\|f^\dagger\|_{\mathcal{H}_\lambda}$ . Although these norms can be controlled in the PDE setting [59] via compact embeddings of Sobolev spaces, there is no clear strategy for obtaining a-priori bounds on these norms for general machine learning problems<sup>27</sup>.*

**8.4. Deep residual Gaussian processes.** Write  $\zeta$  for the centered GP (independent from  $\xi$ ) defined by the quadratic norm  $\int_0^1 \|v(\cdot, t)\|_{\mathcal{V}}^2 dt$  on  $L^2([0, 1], \mathcal{V})$ . Recall [63, Sec. 7&17] that  $\zeta$  is an isometry mapping  $L^2([0, 1], \mathcal{V})$  to a Gaussian space (defined by  $\int_0^1 \langle \zeta, v \rangle_{\mathcal{V}} dt \sim \mathcal{N}(0, \int_0^1 \|v(\cdot, t)\|_{\mathcal{V}}^2 dt)$  for  $v \in L^2([0, 1], \mathcal{V})$ ). The following proposition presents a construction/representation of the GP  $\zeta$ .

**Proposition 8.6.** *Write  $\Gamma$  for the kernel associated with  $\mathcal{V}$ . Let  $\psi$  and  $\mathcal{F}$  be a feature map and (separable) feature space for  $\Gamma$ . Let the  $e_i$  form an orthonormal basis of  $\mathcal{F}$  and let the  $B^i$  be independent one dimensional Brownian motions. Then*

$$\zeta(x, t) = \sum_i \frac{dB_t^i}{dt} \psi^T(x) e_i \quad (8.10)$$

is a representation of  $\zeta$ .

*Proof.* Thm. 2.5 implies that  $v \in L^2([0, 1], \mathcal{V})$  admits the representation  $v(x, t) = \sum_i \alpha_i(t) \psi^T(x, t) e_i$  where the  $\alpha_i$  are scalar-valued functions in  $L^2([0, 1], dt)$  such that  $\sum_i \int_0^1 \alpha_i^2(t) dt = \int_0^1 \|v(\cdot, t)\|_{\mathcal{V}}^2 dt < \infty$ . We conclude by observing that (using Thm. 2.5 again)  $\int_0^1 \langle \zeta, v \rangle_{\mathcal{V}} dt = \sum_i \int_0^1 \alpha_i(t) dB_t^i \sim \mathcal{N}(0, \sum_i \int_0^1 \alpha_i^2(t) dt)$ . □

Let  $\phi^\zeta$  be the solution of (1.7) with  $v = \zeta$ . We call this solution a **deep residual Gaussian process**. Note that whereas deep Gaussian processes are defined by composing function-valued Gaussian processes [23], we define deep residual Gaussian processes as the flow of map of the stochastic dynamic system

$$\dot{z} = \zeta(z, t) \text{ with } z(0) = x \quad (8.11)$$

driven by the function-valued GP vector field  $\zeta$ . Evidently, the existence and uniqueness of solutions to (8.11) require the Cameron-Martin space of  $\zeta$  to be sufficiently regular. As shown in the following proposition this result is a simple consequence of the regularity of the feature map in the finite-dimensional setting.

<sup>27</sup>Although deep learning estimates derived from Barron spaces [6, 25] have a priori Monte-Carlo (dimension independent) convergence rates, they also suffer from this problem since they require bounding the Barron norm of the target function.



**Proposition 8.7.** *Using the notations of Prop. 8.6, if  $\mathcal{F}$  and  $\mathcal{X}$  are finite-dimensional and if  $\psi$  is uniformly Lipschitz continuous then (8.11) (and therefore (1.7) with  $v = \zeta$ ) has a unique strong solution.*

*Proof.* (8.11) corresponds to the finite-dimensional SDE  $dz = \sum_i \psi^T(z) e_i dB_t^i$  which is known [42] to have unique strong solutions if  $\psi$  is uniformly Lipschitz.  $\square$

Let  $\xi \sim \mathcal{N}(0, K)$  (independent from  $\zeta$ ) where  $K$  is the kernel associated with  $\mathcal{H}$  in (2.16).  $\xi \circ \phi^\zeta(\cdot, 1)$  provides a probabilistic solution to Problem (1) in the sense of the following proposition (whose proof is classical).

**Proposition 8.8.** *Let  $(v, f)$  minimize (3.24) with  $\ell_{\mathcal{Y}} = (2.17)$ . Let  $\phi^v$  be the solution of (1.7) obtained from  $v$ . Then*

$$f^\ddagger(\cdot) = f \circ \phi^v(\cdot, 1) \quad (8.12)$$

is a MAP estimator of  $\xi \circ \phi^{\sqrt{\frac{\lambda}{\nu}} \zeta}(\cdot, 1)$  given the information

$$\xi \circ \phi^{\sqrt{\frac{\lambda}{\nu}} \zeta}(X, 1) + \sqrt{\lambda} Z = Y, \quad (8.13)$$

where  $Z = (Z_1, \dots, Z_N)$  is a centered random Gaussian vector, independent from  $\zeta$  and  $\xi$ , with i.i.d.  $\mathcal{N}(0, I_{\mathcal{Y}})$  entries.

The following proposition generalizes Prop. 8.8 to the regularized setting of Sec. 4.

**Proposition 8.9.** *Let  $\xi, \zeta, Z$  be as in Prop. 8.8. Write  $\kappa$  for the centered GP defined<sup>28</sup> by the norm  $\int_0^1 \|\cdot\|_{\mathcal{X}^N}^2$ . Let  $z$  be the stochastic process defined as the solution of  $\dot{z} = \sqrt{\frac{\lambda}{\nu}}(\zeta(z, t) + \sqrt{r}\kappa)$  with initial value  $z(0) = X$ . Let  $(v, f)$  be a minimizer of (4.20), and let  $\phi^v$  be the solution of (1.7). The regularized solution  $f^\ddagger = f \circ \phi^v(\cdot, 1)$  to Problem (1) is a MAP estimator of  $\xi \circ \phi^{\sqrt{\frac{\lambda}{\nu}} \zeta}(\cdot, 1)$  given the information  $\xi(z(1)) + \sqrt{\lambda + \rho} Z = Y$ .*

**8.5. Errors estimates for mechanical regression.** As in Subsec. 8.2 the conditional posterior distribution of  $\xi \circ \phi^{\sqrt{\frac{\lambda}{\nu}} \zeta}(\cdot, 1)$  (conditioned on (8.13)) provides natural probabilistic error estimates on accuracy of a mechanical regression solution  $f$  to Problem 1. We will now derive deterministic error estimates.

**Corollary 8.10.** *In the setting of Prop. 3.13 it holds true that*

$$\|f^\ddagger(x) - f \circ \phi^v(x, 1)\|_{\mathcal{Y}} \leq \sigma(x) \|f^\ddagger\|_{\mathcal{H}^v}, \quad (8.14)$$

and

$$\|f^\ddagger(x) - f \circ \phi^v(x, 1)\|_{\mathcal{Y}} \leq \sqrt{\sigma^2(x) + \lambda \dim(\mathcal{Y})} \|f^\ddagger\|_{\mathcal{H}_\lambda^v}, \quad (8.15)$$

with

$$\sigma^2(x) := \text{Tr} [K^v(x, x) - K^v(x, X)(K^v(X, X) + \lambda I_{\mathcal{Y}})^{-1} K^v(X, x)]. \quad (8.16)$$

*Proof.* Cor. 8.10 is a direct consequence of Thm. 8.4 and Prop. 3.13.  $\square$

<sup>28</sup>For  $q \in L^2([0, 1], \mathcal{X}^N)$ ,  $\int_0^1 \kappa(t)u(t) \sim \mathcal{N}(0, \int_0^1 \|u\|_{\mathcal{X}^N}^2 dt)$  [63, Sec. 7&17].

**Remark 8.11.** (8.14) and (8.15) are a priori error estimate similar to those found in PDE numerical analysis. However, although compactness and ellipticity can be used in PDE analysis [59] to bound  $\|f^\dagger\|_{\mathcal{H}^v}$  or  $\|f^\dagger\|_{\mathcal{H}_\lambda^v}$ , such upper bounds are not available for general machine learning problems. Furthermore, a (deterministic) a posteriori analysis can only provide lower bounds on  $\|f^\dagger\|_{\mathcal{H}^v}$  and  $\|f^\dagger\|_{\mathcal{H}_\lambda^v}$ . Examples of such bounds are  $\|f\|_{\mathcal{H}^v} \leq \|f^\dagger\|_{\mathcal{H}^v}$  and

$$\|f^\dagger\|_{\mathcal{H}_\lambda^v} \leq \|f^\dagger\|_{\mathcal{H}^v}, \quad (8.17)$$

(implied by Prop. 8.10 and  $\ell_{\mathcal{Y}}(f^\dagger(X), Y) = 0$ ) for  $f^\dagger(\cdot) = f \circ \phi^v(\cdot, 1)$ . Note that Prop. 3.13 and Cor. 8.10 do not make assumptions on  $\phi^v$ . If  $\phi^v$  is selected as a minimizer of (3.19) then  $f^\dagger(\cdot) = f \circ \phi^v(\cdot, 1)$  is a mechanical regression solution (3.26) to Problem 1. Given the identity (3.27), in the limit  $\nu \downarrow 0$ , the variational formulations (3.19) and (3.28) seek to minimize the norm  $\|f^\dagger\|_{\mathcal{H}_\lambda^v}$  (which acts in (8.17) as lower bound for the term  $\|f^\dagger\|_{\mathcal{H}_\lambda^v}$  appearing in the error bound (8.15)). Ignoring the gap between  $\|f^\dagger\|_{\mathcal{H}_\lambda^v}$  and  $\|f^\dagger\|_{\mathcal{H}^v}$  in (8.17) mechanical regression seems to select a kernel  $K^v(x, x') = K(\phi^v(x, 1), \phi^v(x', 1))$  making the bound (8.15) as sharp as possible. The penalty  $\nu \int_0^1 \frac{1}{2} \|v\|_{\mathcal{Y}}^2 dt$  in (3.19) can then be interpreted as a regularization term whose objective is to avoid a large gap in (8.17) that could be created if  $\phi^v(x, 1)$  overfits the data.

## 9. Reduced equivariant multi-channel (REM) kernels and feature maps

**9.1. Reduced kernels.** In the setting of Sec. 2, let  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  be an operator-valued kernel and let  $P : \mathcal{X} \rightarrow \mathcal{X}$  and  $R : \mathcal{Y} \rightarrow \mathcal{Y}$  be linear projections. Note that  $RK(Px, Px')R : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(R\mathcal{Y})$  is also an operator-valued kernel. The following proposition generalizes (2.11) and (2.14) to partial measurements on the inputs and outputs of the unknown function  $f^\dagger$  in Problem 1.

**Proposition 9.1.** *Using the relative error in  $\|\cdot\|_{\mathcal{H}}$ -norm as a loss, the minimax optimal recovery of an unknown function  $f^\dagger \in \mathcal{H}$  given  $Rf^\dagger(PX) = Z$  (with  $X := (X_1, \dots, X_N) \in \mathcal{X}^N$  and  $Z := (Z_1, \dots, Z_N) \in (R\mathcal{Y})^N$ ) is the minimizer of*

$$\begin{cases} \text{Minimize} & \|f\|_{\mathcal{H}} \\ \text{subject to} & Rf(PX) = Z, \end{cases} \quad (9.1)$$

which admits the representation

$$f(\cdot) = K(\cdot, PX)R(RK(PX, PX)R)^{-1}Z \text{ with } \|f\|_{\mathcal{H}}^2 = Z^T(RK(PX, PX)R)^{-1}Z, \quad (9.2)$$

where  $RK(PX, PX)R$  is the  $N \times N$  block-operator matrix with entries  $RK(PX_i, PX_j)R$  and  $K(\cdot, PX)R$  is the  $N$ -vector with entries  $K(\cdot, PX_i)R$ .

*Proof.* The proof of minimax optimality of the minimizer of (9.1) is similar to that of [63, Thm. 12.4, 12.5]. The representation (9.2) follows by observing that that  $Rf(PX_i) = Z_i$  and that  $f$  is  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ -orthogonal to the set of  $g \in \mathcal{H}$  such that  $Rg(PX_i) = 0$  (since  $f$  has the representation  $f = \sum_i K(\cdot, PX_i)RV_i$  with  $V_i \in \mathcal{Y}$  and  $\langle K(\cdot, PX_i)RV_i, g \rangle_{\mathcal{H}} = \langle g(PX_i), RV_i \rangle_{\mathcal{Y}} = \langle Rg(PX_i), V_i \rangle_{\mathcal{Y}} = 0$  via the reproducing identity).  $\square$

**9.2. Equivariant multi-channel kernels.** We will now present a generalization of the equivariant kernels of [74].

**9.2.1. The unitary group of transformations on the base space.** Let  $\mathfrak{X}$  be a separable Hilbert space. Let  $\mathcal{G}$  be a (compact, possibly finite) group of linear unitary transformations acting on  $\mathfrak{X}$ :  $g \in \mathcal{G}$  maps  $\mathfrak{X}$  to  $\mathfrak{X}$ ,  $\mathcal{G}$  is closed under composition,  $\mathcal{G}$  contains the identity map  $i_d$ ,  $g \in \mathcal{G}$  has an inverse  $g^{-1}$  such that  $gg^{-1} = g^{-1}g = i_d$ , and  $\langle gx, gx' \rangle_{\mathfrak{X}} = \langle x, x' \rangle_{\mathfrak{X}}$  for  $g \in \mathcal{G}$  and  $x, x' \in \mathfrak{X}$  (i.e.  $g^T = g^{-1}$  where  $g^T$  is the adjoint of  $g$ ). Write  $dg$  for the Haar measure associated with  $\mathcal{G}$  and  $|\mathcal{G}| := \int_{\mathcal{G}} dg$  for the volume of the group ( $|\mathcal{G}| = \text{Card}(\mathcal{G})$  when the group is finite) and assume  $\mathcal{G}$  to be unimodular ( $dg$  is invariant under both the left and right action of the group, i.e.  $\int_{\mathcal{G}} f(g) dg = \int_{\mathcal{G}} f(gg') dg = \int_{\mathcal{G}} f(g'g) dg$  for  $g' \in \mathcal{G}$ ). Write  $\mathbb{E}_{\mathcal{G}}$  for the expectation with respect to the probability distribution induced by  $dg/|\mathcal{G}|$  on  $\mathcal{G}$ .

**9.2.2. Extension to multiple channels.** Let  $c$  be a strictly positive integer called the number of channels. Let  $\mathfrak{X}^c$  be the  $c$ -fold product space of  $\mathfrak{X}$  endowed with the scalar product defined by  $\langle x, x' \rangle_{\mathfrak{X}^c} := \sum_{i=1}^c \langle x_i, x'_i \rangle_{\mathfrak{X}}$  for  $x = (x_1, \dots, x_c) \in \mathfrak{X}^c$  and  $x' \in \mathfrak{X}^c$ . The action of the group  $\mathcal{G}$  can be naturally diagonally be extended to  $\mathfrak{X}^c$  by

$$g(x_1, \dots, x_c) := (gx_1, \dots, gx_c) \text{ for } g \in \mathcal{G} \text{ and } (x_1, \dots, x_c) \in \mathfrak{X}^c. \quad (9.3)$$

Note that  $\mathcal{G}$  remains unitary on  $\mathfrak{X}^c$  ( $\langle gx, gx' \rangle_{\mathfrak{X}^c} = \langle x, x' \rangle_{\mathfrak{X}^c}$ ).

**9.2.3. Equivariant multi-channel kernels.** We will now introduce equivariant multi-channel kernels in the setting of Subsec. 2.1.

**Definition 9.2.** Let  $\mathcal{X} = \mathfrak{X}^{c_1}$  and  $\mathcal{Y} = \mathfrak{X}^{c_2}$  with  $c_1, c_2 \in \mathbb{N}^*$ . We say that an operator-valued kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  is  $\mathcal{G}$ -equivariant if

$$K(gx, g'x') = gK(x, x')(g')^T \text{ for all } g, g' \in \mathcal{G}. \quad (9.4)$$

Similarly we say that a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is  $\mathcal{G}$ -equivariant if

$$f(gx) = gf(x) \text{ for all } (x, g) \in \mathcal{X} \times \mathcal{G}. \quad (9.5)$$

Set  $\mathcal{X} = \mathfrak{X}^{c_1}$  and  $\mathcal{Y} = \mathfrak{X}^{c_2}$  as in Def. 9.2.

**Proposition 9.3.** Given a (possibly non-equivariant) kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$ ,

$$K^{\mathcal{G}}(x, x') := \frac{1}{|\mathcal{G}|^2} \int_{\mathcal{G}^2} g^T K(gx, g'x') g' dg dg' := \mathbb{E}_{\mathcal{G}^2} [g^T K(gx, g'x') g'], \quad (9.6)$$

is a  $\mathcal{G}$ -equivariant kernel  $K^{\mathcal{G}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$ .

*Proof.* The proof is similar to that of [74, Prop. 2.2]. Simply observe that for  $\bar{g}, \bar{g}' \in \mathcal{G}$ ,  $\mathbb{E}_{\mathcal{G}^2} [g^T K(g\bar{g}x, g'\bar{g}'x') g'] = \bar{g} \mathbb{E}_{\mathcal{G}^2} [(g\bar{g})^T K(g\bar{g}x, g'\bar{g}'x') g'\bar{g}'] (\bar{g}')^T$ .  $\square$

We say that  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  is  $\mathcal{G}$ -invariant<sup>29</sup> if  $K(gx, g'x') = K(x, x')$  for  $(x, x', g, g') \in (\mathcal{X})^2 \times \mathcal{G}^2$ . We say that  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  is weakly  $\mathcal{G}$ -invariant if  $K(gx, gx') = K(x, x')$  for  $(x, x', g) \in \mathcal{X}^2 \times \mathcal{G}$ .

<sup>29</sup>Given a non-invariant kernel  $K$  Haar integration can also be used [28] to derive the invariant kernel  $\mathbb{E}_{\mathcal{G}^2} [K(gx, g'x')]$ .

**Remark 9.4.** If  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$  is scalar and weakly  $\mathcal{G}$ -invariant then

$$K^{\mathcal{G}}(x, x') = \mathbb{E}_{\mathcal{G}}[K(x, g'x')g'], \quad (9.7)$$

since  $\mathbb{E}_{\mathcal{G}^2}[g^T K(gx, g'x')g'] = \mathbb{E}_{\mathcal{G}^2}[K(x, g^T g'x')g^T g'] = \mathbb{E}_{\mathcal{G}}[K(x, g'x')g']$ . (9.7) matches the construction of [74] for  $c_1 = c_2 = 1$ .

The interpolant (2.14) of the data  $X_i, Y_i$  with a  $\mathcal{G}$ -equivariant kernel  $K$  (1) is a equivariant function (satisfies  $f(gx) = gf(x)$ ) and (2) is equal to the interpolant of the enriched data  $(gX_i, gY_i)_{g \in \mathcal{G}, 1 \leq i \leq N}$  with  $K$ . However, although interpolating with an equivariant kernel implicitly enriches the data, interpolating the enriched data  $(gX_i, gY_i)_{g \in \mathcal{G}, 1 \leq i \leq N}$  with a non-equivariant kernel  $K$  does not guarantee the equivariance (9.5) of the interpolant (2.14). Furthermore, we have the following variant of [74, Thm. 2.8].

**Theorem 9.5.** If  $K$  is scalar and weakly  $\mathcal{G}$ -invariant then the minimizer of (2.11) with the constraint that  $f$  must also be  $\mathcal{G}$ -equivariant is  $f^{\mathcal{G}}(\cdot) := K^{\mathcal{G}}(\cdot, X)K^{\mathcal{G}}(X, X)^{-1}Y$ .

*Proof.* By construction  $f^{\mathcal{G}}$  satisfies the constraints of (2.11) and is  $\mathcal{G}$ -equivariant. To show that  $f^{\mathcal{G}}$  is the minimizer simply observe that  $\langle f^{\mathcal{G}}, u \rangle_{\mathcal{H}} = 0$  if  $u \in \mathcal{H}$  is  $\mathcal{G}$ -equivariant and satisfies  $u(X) = 0$ . Indeed (writing  $V := K^{\mathcal{G}}(X, X)^{-1}Y$ )  $\langle f^{\mathcal{G}}, u \rangle_{\mathcal{H}} = \sum_i \mathbb{E}_{\mathcal{G}}[\langle K(\cdot, g'X_i)g'V_i, u \rangle_{\mathcal{H}}] = 0$  since (by the reproducing identity)  $\langle K(\cdot, g'X_i)g'V_i, u \rangle_{\mathcal{H}} = \langle u(g'X_i), g'V_i \rangle_{\mathcal{Y}} = \langle g'u(X_i), g'V_i \rangle_{\mathcal{Y}} = 0$ .  $\square$

**Remark 9.6.** Let  $(x, g^{\dagger}) \in \mathcal{X} \times \mathcal{G}$  and  $y = g^{\dagger}x$  and consider the problem of recovering  $g^{\dagger}$  (which we refer to as the relative pose between  $x$  and  $y$ ) from the observation of  $K(x, x)$ ,  $K(y, y)$  and  $K(x, y)$ . Although this problem is impossible if  $K$  is  $\mathcal{G}$ -invariant (since  $K(x, x) = K(y, y) = K(x, y)$ ), it remains solvable if  $K$  is  $\mathcal{G}$ -equivariant (since  $K(x, y) = K(x, x)g^T$  and  $g = (K(x, y))^T K(x, x)^{-1}$ ). Therefore, contrary to invariant kernels [28], equivariant kernels preserve the relative pose information between objects [74, Sec. 2.3]. The notion of equivariance has been used in deep learning to design convolutional neural networks [45] on non-flat manifolds [22] and for preserving intrinsic part/whole spatial relationship in image recognition [77].

The following theorem shows that interpolants/regressors obtained from mechanical regression with an equivariant kernel are also equivariant.

**Theorem 9.7.** Let  $\mathcal{H}$  be the RKHS defined by a  $\mathcal{G}$ -equivariant kernel  $\Gamma : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{X})$ . Then for  $v \in C([0, 1], \mathcal{V})$  obtained as a minimizer of (3.19) or (4.13), the solution  $\phi^v$  of (1.7) is also  $\mathcal{G}$ -equivariant in the sense that

$$\phi^v(gz, t) = g\phi^v(z, t) \text{ for all } (z, g, t) \in \mathcal{X} \times \mathcal{G} \times [0, 1]. \quad (9.8)$$

*Proof.* The proof follows from Thm. 3.8 and Thm. 4.4 by continuous induction on  $t$ . Indeed (3.20) implies that  $\dot{\phi}^v(gz, t) = g\dot{\phi}^v(z, t)$  as long as  $\phi^v(gz, t) = g\phi^v(z, t)$ .  $\square$

**9.3. REM kernels.** In the setting of Sec. 9.2, let  $R$  and  $P$  be linear projections from  $\mathfrak{X}$  onto closed linear subspaces of  $\mathfrak{X}$ . Extend the action of  $P$  to  $\mathcal{X} = \mathfrak{X}^{c_1}$  by  $P(x_1, \dots, x_{c_1}) = (Px_1, \dots, Px_{c_1})$ . Similarly extend the action of  $R$  to  $\mathcal{Y} = \mathfrak{X}^{c_2}$ . Observe that, given an operator-valued kernel  $K : P\mathcal{X} \times P\mathcal{X} \rightarrow \mathcal{L}(R\mathcal{Y})$ ,

$$\bar{K}(x, x') := RK(Px, Px')R \quad (9.9)$$

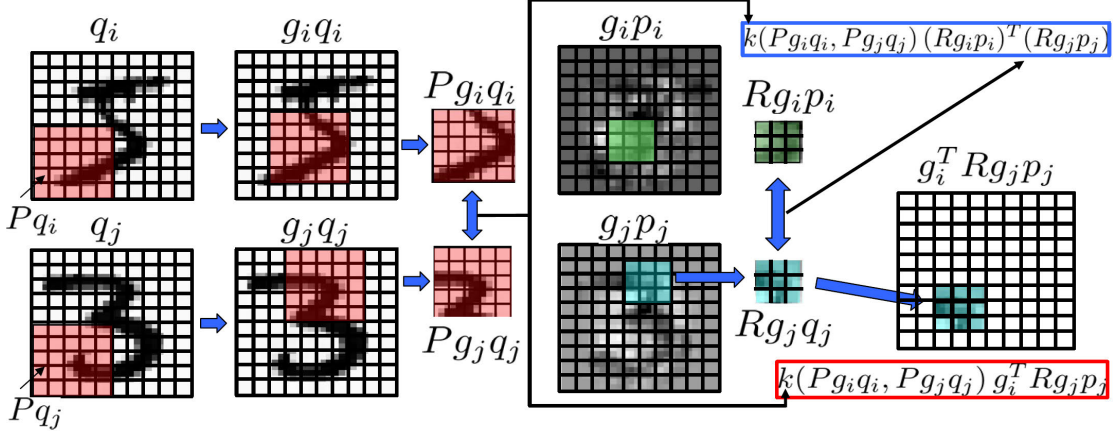


FIGURE 8. Mechanical regression with REM kernels.

is an operator-valued kernel<sup>30</sup>  $\bar{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(Y)$ . Prop. 9.3 implies that

$$C(x, x') := \bar{K}^{\mathcal{G}}(x, x') = \mathbb{E}_{\mathcal{G}^2} [g^T R K(Pg x, Pg' x') R g'] \quad (9.10)$$

is an equivariant operator-valued kernel  $C : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(Y)$ . We call (9.10) a REM (reduced equivariant multi-channel) kernel.

Fig. 8 illustrates the Hamiltonian system (3.13) for  $c_1 = c_2 = 1$ , with  $\Gamma = C = (9.10)$  obtained from a scalar kernel  $K(x, x') = k(x, x') I_{RY}$  (where the arguments of  $k$  are  $5 \times 5$  images) and the (finite) group of translations  $\mathcal{G}$  on periodized  $10 \times 10$  images. Note that  $Pq_i$  projects the image  $q_i$  to its bottom left  $5 \times 5$  sub-image and  $Rp_i$  projects the image  $p_i$  to its bottom left  $3 \times 3$  sub-image.  $Pg_i q_i$  translates  $q_i$  by  $g_i$  before the projection  $P$  which is equivalent to projecting  $q_i$  onto the  $g_i^T$  translation of the original  $5 \times 5$  patch.  $\sum_j k(Pg_i q_i, Pg_j q_j) g_i^T R g_j p_j$  creates a  $10 \times 10$  image adding (over  $j$ ) the  $g_i^T$  translates of sub-images  $g_i^T R g_j p_j$  weighted by  $k(Pg_i q_i, Pg_j q_j)$ . We will now show that this is equivalent to performing a weighted convolution, and convolutional neural networks [43] can be recovered as the feature map version of this algorithm.

**9.4. REM feature maps.** Let  $\mathcal{F}$  and  $\psi : \mathcal{X} \rightarrow \mathcal{L}(Y, \mathcal{F})$  be a feature space and map associated with the kernel  $K$  in (9.10). Then  $C(x, x') = \mathbb{E}_{\mathcal{G}^2} [g^T R \psi^T(Pg x) \psi(Pg' x') R g']$  implies that  $C$  has feature space  $\mathcal{F}$  and feature map  $\Psi$  defined by

$$\Psi(x)y = \mathbb{E}_{\mathcal{G}} [\psi(Pg x) R g y]. \quad (9.11)$$

If  $K$  is a scalar kernel as in Subsec. 6.2.1 with feature space/map  $\mathfrak{F}$  and  $\varphi : \mathcal{X} \rightarrow \mathfrak{F}$ , then  $\mathcal{F} = \mathcal{L}(\mathfrak{F}, RY)$  and  $\psi(x)Ry = Ry\varphi^T(x)$  imply  $\Psi(x)y = \mathbb{E}_{\mathcal{G}} [Rgy\varphi^T(Pg x)]$  and (for  $\alpha \in \mathcal{F}$ )

$$\Psi^T(x)\alpha = \mathbb{E}_{\mathcal{G}} [g^T \alpha \varphi(Pg x)]. \quad (9.12)$$

<sup>30</sup>Note that  $K$  can also be identified as the reduced kernel of  $\bar{K}$ . When the dimension of  $P\mathcal{X}$  is low then the interpolation of functions mapping  $P\mathcal{X}$  to  $R\mathcal{X}$  does not suffer from the curse of dimensionality.

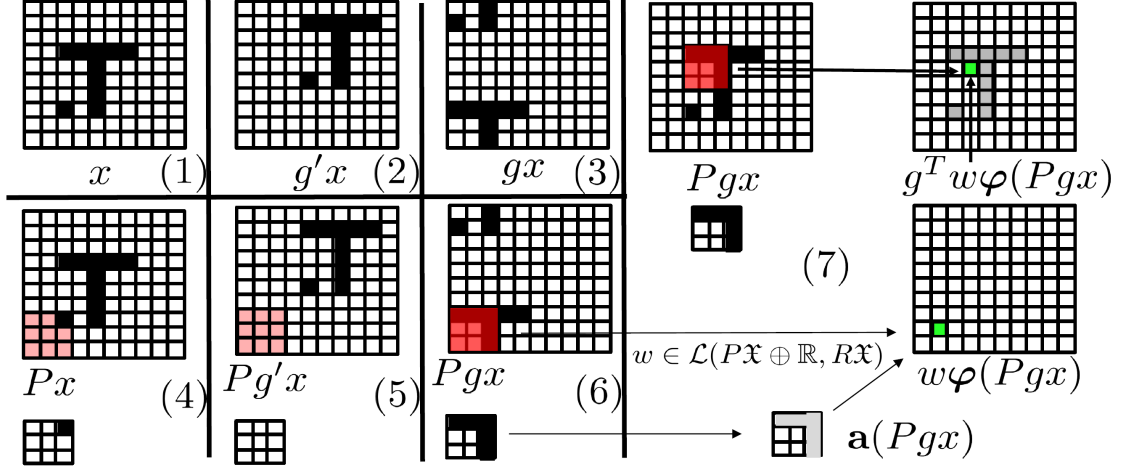


FIGURE 9. REM feature maps.

If  $\varphi$  is obtained from an elementwise nonlinearity activation function as in (6.11) and Rmk. 6.7 then  $\mathfrak{X} = P\mathcal{X} \oplus \mathbb{R}$ , and for  $\alpha = w \in \mathcal{L}(P\mathcal{X} \oplus \mathbb{R}, R\mathcal{X})$  we have

$$\Psi^T(x)\alpha = \mathbb{E}_{\mathcal{G}}[g^T(w\varphi(Pgx))]. \quad (9.13)$$

We call (9.11), (9.12) and (9.13) REM (reduced equivariant multi-channel) feature maps.

Fig. 9 shows the action of (9.13). In that illustration  $c_1 = c_2 = 1$ , the elements of  $\mathfrak{X}$  are  $10 \times 10$  images, and  $\mathcal{G}$  is the group of translations acting on  $10 \times 10$  images with periodic boundaries as shown in subimages (1-3).  $Px$  projects the  $10 \times 10$  image  $x$  onto the lower left  $3 \times 3$  sub-image (by zeroing out the pixels outside that left corner). The action of  $P$  on  $x$  and the translation of  $x$  by  $g'$  and  $g$  are illustrated in subimages (4-6). Note that translating  $x$  by  $g$  before applying  $P$  is equivalent to translating the action of  $P$  as illustrated in subimage (7) and commonly done in CNNs.  $Rx$  projects the  $10 \times 10$  image onto the green pixel at the bottom right of subimage (7). In the setting of CNNs  $w \in \mathcal{L}(P\mathfrak{X} \oplus \mathbb{R}, R\mathfrak{X})$  is one convolutional patch incorporating a  $3 \times 3$  weight matrix  $W$  and a  $1 \times 3$  vector  $b$  and computing  $g^T w\varphi(Pgx)$  is equivalent to obtaining the value of the green pixel on the top right of subimage (7) by computing  $W\mathbf{a}(Pgx) + b$ .

**9.5. Downsampling with subgrouping.** ANNs include downsampling operations such as pooling or striding. Downsampling is incorporated in REM kernels and feature maps by employing sub-groups of  $\mathcal{G}$  in the construction of the REM feature maps. Note that (9.12) and (9.13) are contained in

$$\mathcal{GR}\mathcal{X} := \oplus_{g \in \mathcal{G}} gR\mathcal{X} \quad (9.14)$$

with  $gR\mathcal{X} := \{gRx \mid x \in \mathcal{X}\}$ . Note that  $\mathcal{GR}\mathcal{X} \subset \mathcal{X}$  and this inclusion can be a strict one when  $\mathcal{G}$  is a proper subgroup of an overgroup. When  $\mathcal{G}$  is a group of translations, then subgrouping is equivalent to striding. Fig. 10 illustrates the proposed downsampling approach for  $\mathcal{X} = \mathfrak{X}$ . In that illustration, the elements of  $\mathfrak{X}$  are  $8 \times 8$  images (with periodic boundaries). (1)  $\mathcal{G}_1$  is the group of all 64 possible translations. (2)  $\mathcal{G}_2$  is a group

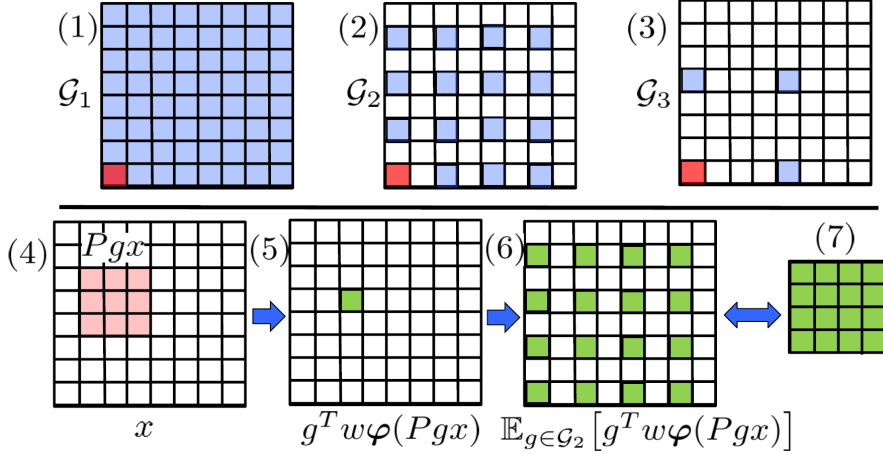


FIGURE 10. Downsampling with subgrouping.

of 16 possible translations obtained as a sub-group of  $\mathcal{G}_1$  with a stride of 2. (3)  $\mathcal{G}_3$  is a group of 4 possible translations obtained as a sub-group of  $\mathcal{G}_1$  with a stride of 4 or as a sub-group of  $\mathcal{G}_2$  with a stride of 2. The bottom row shows the action of a REM feature map constructed from the sub-group  $\mathcal{G}_2$ . (4) shows  $Pgx$  for a given  $g \in \mathcal{G}_2$  ( $Px$  is  $3 \times 3$  image). (5) shows  $g^T w \varphi(Pgx)$  ( $R\mathfrak{X}$  is a set of  $1 \times 1$  images and  $w \in \mathcal{L}(P\mathfrak{X} \oplus \mathbb{R}, R\mathfrak{X})$ ). (6) shows the average of  $g^T w \varphi(Pgx)$  over  $g \in \mathcal{G}_2$ . The range of  $\mathbb{E}_{\mathcal{G}_2}[g^T w \varphi(Pgx)]$  is the set of  $8 \times 8$  images whose pixel values are zero outside the green pixels. (7) Ignoring the white pixels (whose values are zero), the range  $\mathcal{G}_2 R\mathfrak{X}$  of  $\mathbb{E}_{\mathcal{G}_2}[g^T w \varphi(Pgx)]$  (writing  $\mathbb{E}_{\mathcal{G}_2}$  for the expectation with respect to the normalized Haar measure on  $\mathcal{G}_2$ ) can be identified with the set of  $4 \times 4$  images as it is done with CNNs.

**9.6. Idea formation with REM kernels.** We will now show that **CNNs and ResNets are particular instances of idea formation with REM kernels** as illustrated in Fig. 11. Consider the setting of Sec. 5.2 and 7. Let  $\varphi(x) = (\mathbf{a}(x), 1)$  where  $\mathbf{a}$  is an activation function obtained as an elementwise nonlinearity satisfying the regularity conditions of Rmk. 6.7. Given  $D \geq 1$ , let  $\mathfrak{X}_0, \dots, \mathfrak{X}_D$  and  $\mathcal{X}_0, \dots, \mathcal{X}_{D+1}$  be finite-dimensional Hilbert spaces constructed as follows. Set  $\mathfrak{X}_0 = \mathcal{X}_0 = \mathcal{X}$  and  $\mathcal{X}_{D+1} = \mathcal{Y}$ . For  $m \in \{1, \dots, D\}$  let  $P_m^\Gamma : \mathfrak{X}_m \rightarrow \mathfrak{X}_m$ ,  $R_m^\Gamma : \mathfrak{X}_m \rightarrow \mathfrak{X}_m$ , be linear projections, and for  $m \in \{0, \dots, D-1\}$  let  $P_m^K : \mathfrak{X}_m \rightarrow \mathfrak{X}_m$ ,  $R_m^K : \mathfrak{X}_m \rightarrow \mathfrak{X}_m$  be linear projections. Let  $c_1, \dots, c_D \in \mathbb{N}^*$ . For  $m \in \{1, \dots, D\}$ , let  $\mathcal{X}_m = \mathfrak{X}^{c_m}$ . Let  $\mathcal{G}$  be a unitary unimodular group on  $\mathcal{X}$ , write  $\mathcal{G}_0 = \mathcal{G}$  and for  $m \in \{0, \dots, D\}$  let  $\mathcal{G}_{m+1}$  be a subgroup of  $\mathcal{G}_m$  and let  $\mathfrak{X}_{m+1} = \mathcal{G}_m R_m \mathfrak{X}_m$ . For  $m \in \{0, \dots, D-1\}$  let  $K^m : \mathcal{X}_m \times \mathcal{X}_m \rightarrow \mathcal{L}(\mathcal{X}_{m+1})$  be the REM kernel defined (as in (9.13)) by  $P_m^K, R_m^K, \mathcal{G}_m$ . Let  $K^D : \mathcal{X}_D \times \mathcal{X}_D \rightarrow \mathcal{L}(\mathcal{X}_{D+1})$  be the REM kernel with feature map defined by  $\Psi^T(x)\alpha = w\varphi(x)$  with  $w \in \mathcal{L}(\mathcal{X}^D \oplus \mathbb{R}, \mathcal{Y})$ . For  $m \in \{1, \dots, D\}$  let  $\Gamma^m : \mathcal{X}_m \times \mathcal{X}_m \rightarrow \mathcal{L}(\mathcal{X}_m)$  be the REM kernel defined by the feature map (9.13) using  $P_m^\Gamma, R_m^\Gamma, \mathcal{G}_m$ .



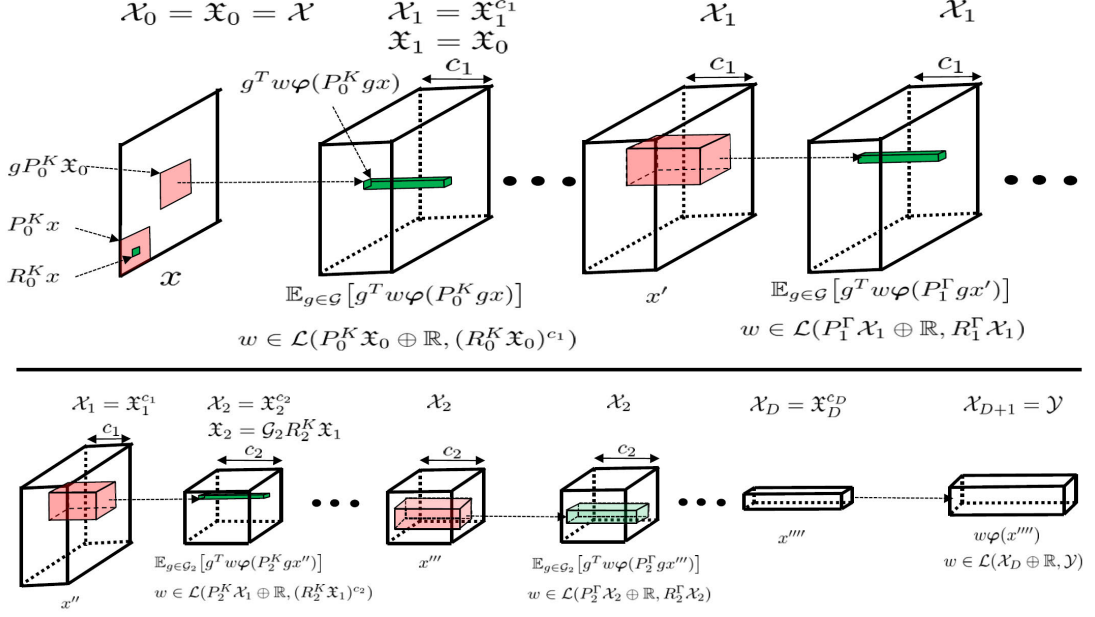


FIGURE 11. CNNs and ResNets as idea formation with REM kernels.

The corresponding discrete idea formation solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition (5.6) with

$$f_m = \mathbb{E}_{\mathcal{G}_m} [g^T \tilde{w}^m \varphi(P_m^K g \cdot)] \text{ and } v_{m,j} = \mathbb{E}_{\mathcal{G}_m} [g^T w^{m,j} \varphi(P_m^\Gamma g \cdot)], \quad (9.15)$$

where the  $\tilde{w}^m$  are  $w^{m,j}$  are minimizers of

$$\begin{cases} \text{Min} & \lambda_0 (\|\tilde{w}^0\|_{\mathcal{L}(P_0^K \mathfrak{X}_0 \oplus \mathbb{R}, (R_0^K \mathfrak{X}_0)^{c_1})}^2 + \frac{1}{\rho} \|f_0(X) - q^{1,1}\|_{\mathcal{X}_1^N}^2) + \sum_{m=1}^D \left( \frac{\nu_m L_m}{2} \sum_{j=1}^{L_m} \right. \\ & (\|w^{m,j}\|_{\mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m)}^2 + \frac{1}{r} \|q^{m,j+1} - q^{m,j} - v_{m,j}(q^{m,j})\|_{\mathcal{X}_m^N}^2) + \lambda_m \cdot \\ & (\|\tilde{w}^m\|_{\mathcal{L}(P_m^K \mathfrak{X}^{c_m} \oplus \mathbb{R}, R_m^\Gamma \mathfrak{X}^{c_{m+1}})}^2 + \frac{1}{\rho} \|f_m(q^{m,L_m+1}) - q^{m+1,1}\|_{\mathcal{X}_m^N}^2) + \ell_Y(q^{D+1,1}, Y) \\ \text{over} & w^{m,j} \in \mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m), \tilde{w}^m \in \mathcal{L}(P_m^K \mathfrak{X}^{c_m} \oplus \mathbb{R}, R_m^\Gamma \mathfrak{X}^{c_{m+1}}), q^{m,j} \in \mathcal{X}_m^N. \end{cases} \quad (9.16)$$

In the continuous ( $\min_m L_m \rightarrow \infty$ ) limit, the idea formation solution to Problem 1 is to approximate  $f^\dagger$  with  $F_{D+1}$  defined by inductive composition (5.8) with  $f_m = \mathbb{E}_{\mathcal{G}_m} [g^T \tilde{w}^m \varphi(P_m^K g \cdot)]$  and  $v_m(x, t) = \mathbb{E}_{\mathcal{G}_m} [g^T w^m(t) \varphi(P_m^\Gamma g \cdot)]$  where the  $\tilde{w}^m$  and  $w^m$  are minimizers of

$$\begin{cases} \text{Min} & \lambda_0 (\|\tilde{w}^0\|_{\mathcal{L}(P_0^K \mathfrak{X}_0 \oplus \mathbb{R}, (R_0^K \mathfrak{X}_0)^{c_1})}^2 + \frac{1}{\rho} \|f_0(X) - q^1(0)\|_{\mathcal{X}_1^N}^2) + \sum_{m=1}^D \left( \frac{\nu_m}{2} \int_0^1 \right. \\ & (\|w^m(t)\|_{\mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m)}^2 + \frac{1}{r} \|\dot{q}^m - v_m(q^m, t)\|_{\mathcal{X}_m^N}^2 dt) + \\ & \lambda_m (\|\tilde{w}^m\|_{\mathcal{L}(P_m^K \mathfrak{X}^{c_m} \oplus \mathbb{R}, R_m^\Gamma \mathfrak{X}^{c_{m+1}})}^2 + \frac{1}{\rho} \|f_m(q^m(1)) - q^{m+1}(0)\|_{\mathcal{X}_m^N}^2) + \ell_Y(q^{D+1}(0), Y) \\ \text{over} & w^m \in C^1([0, 1], \mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m)), \tilde{w}^m \in \mathcal{L}(P_m^K \mathfrak{X}^{c_m} \oplus \mathbb{R}, R_m^\Gamma \mathfrak{X}^{c_{m+1}}), q^m \in C^1([0, 1], \mathcal{X}_m^N). \end{cases} \quad (9.17)$$

The following theorem is a direct consequence of the equivalence established in Subsec. 6.3 and Thm. 5.1.

**Theorem 9.8.** (9.16) and (9.17) have minimizers. Minimal values of (9.16) and (9.17) are continuous in  $(X, Y)$ . Minimal values and  $F_{D+1}$  determined by minimizers of (9.16)

converge (in the sense of adherence values of Subsec. 4.7), as  $\min_m L_m \rightarrow \infty$  towards minimal values and  $F_{D+1}$  determined by minimizers of (9.17). At minima, the  $(\|w^m(t)\|_{\mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m)}^2 + \frac{1}{r} \|\dot{q}^m - v_m(q^m, t)\|_{\mathcal{X}_N^m}^2 dt)$  are constant over  $t \in [0, 1]$  and  $(\|w^{m,j}\|_{\mathcal{L}(P_m^\Gamma \mathcal{X}_m \oplus \mathbb{R}, R_m^\Gamma \mathcal{X}_m)}^2 + \frac{1}{r} \|q^{m,j+1} - q^{m,j} - v^{m,j}(q^{m,j})\|_{\mathcal{X}_N^m}^2)$  fluctuates by at most  $\mathcal{O}(1/L_m)$  over  $j \in \{1, \dots, L_m\}$ . The maps  $y = F_{D+1}(x)$  obtained from (9.16) and (9.17) are equal to the output  $y$  produced by the block diagrams (5.10) and (5.12), and the initial momenta  $p_0^m$  and  $Z^m$  are identified as minimizers of the total loss (5.11). In particular the results of Thm. 5.1, Prop. 5.2 and Prop. 5.3 hold true for (9.16) and (9.17).

**9.7. The algorithm.** The practical minimization of (9.17) is as follows. Introduce the slack variables (see Fig. 2)  $\tilde{z}^0 = q^{1,1} - f_0(X)$ ,  $z^{m,j} = q^{m,j+1} - q^{m,j} - v_{m,j}(q^{m,j})$  and  $\tilde{z}^m = q^{m+1,1} - f_m(q^{m,L_m+1})$ . Let  $\ell_Y$  be an arbitrary empirical loss (e.g.,  $\ell_Y(Y', Y) = \|Y' - Y\|_{\mathcal{Y}_N}^2$ ). Replace the minimization over the variables  $q^{m,j}$  by the minimization over the slack variables (note that  $q^{D+1,1}$  is a function of  $X$ , weights, biases, and slack variables). Use minibatching (as commonly practiced in ML) to form an unbiased estimate of the gradient of the total loss with respect to the weights, biases, and slack variables. The exact averages  $\mathbb{E}_{\mathcal{G}_m}$  in (9.15) can be replaced<sup>31</sup> by Monte-Carlo averages (by sampling the Haar measure over  $\mathcal{G}_m$ ). Modify the weights, biases, and slack variables in the gradient descent direction (note that the only slack variables impacted are those indexed by the minibatch). Repeat.

## 10. Related work

**10.1. Deep kernel learning.** The deep learning approach to Problem 1 is to approximate  $f^\dagger$  with the composition  $f := f_L \circ \dots \circ f_1$  of parameterized nonlinear functions  $f_k : \mathcal{X}_k \rightarrow \mathcal{X}_{k+1}$  (with  $\mathcal{X}_1 := \mathcal{X}$  and  $\mathcal{X}_{L+1} := \mathcal{Y}$ ) identified by minimizing the discrepancy between  $f(X)$  and  $Y$  via Stochastic Gradient Descent. [11] proposes to generalize this approach to the nonparametric setting by introducing a representer theorem for the identification of  $(f_1, \dots, f_L)$  as minimizers in  $\mathcal{H}_1 \times \dots \times \mathcal{H}_L$  (writing  $\mathcal{H}_k$  for a given reproducing kernel Hilbert space of functions  $f_k : \mathcal{X}_k \rightarrow \mathcal{X}_{k+1}$ ) of a loss of the form

$$\sum_{k=1}^L \ell_k(\|f_k\|_{\mathcal{H}_k}) + \ell_{L+1}(f_L \circ \dots \circ f_1(X), Y) \quad (10.1)$$

[11, Thm. 1] reduces (10.1) to a finite-dimensional optimization problem.

**10.2. Computational anatomy and image registration.** Applying concepts from mechanics to classification/regression problems can be traced back to computational anatomy [26] (and more broadly to image registration [12] and shape analysis [96]) where ideas from elasticity and visco-elasticity are used to represent biological variability and create algorithms for the alignment of anatomical structures. Joshi and Miller [39, 40] discovered that minimizers of (1.9) admit a representer formula of the form (3.20) which can be then used to produce computationally tractable algorithms for shape analysis/regression by (1) minimizing a reduced loss of the form (3.8) via gradient descent

<sup>31</sup>Since REM feature maps are expressed as expected values with respect to a randomization of the action of the group, their simulation can be randomized as in [16].

[40, 14] or (2) via (geodesic) shooting algorithms obtained from the Hamiltonian perspective [56, 89]. Therefore idea registration could be seen as a natural generalization of image registration in which image spaces are replaced by abstract feature spaces, material points are replaced by data points, and smoothing kernels (Green’s functions of differential operators) are replaced by REM kernels. Although discretizing the material space is a viable and effective strategy [18] in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) model [24, 8] of image registration, the curse of dimensionality renders it prohibitive for general abstract spaces, which is why idea registration must (for efficiency) be implemented with feature maps. Our regularization strategy for idea registration is a generalization of that of image registration [53]. The sparsity of idea registration in momentum map representation is akin to the sparsity of image deformations in momentum map representation [13, 89].

### 10.3. Interplays between learning, inference, and numerical approximation.

The error estimates discussed in Sec. 8 are instances of interplays between numerical approximation, statistical inference, and learning, which are intimately connected through the common purpose of making estimations/predictions with partial information. These confluences (which are not new, see [35, 21, 64, 63] for reviews) are not just objects of curiosity but constitute a pathway to simple solutions to fundamental problems in all three areas (e.g., solving PDEs as an inference/learning problem [59, 60, 73] facilitates the discovery of efficient solvers with some degree of universality [80, 79]). We also observe that the generalization properties of kernel methods (which, as stressed in [9], are intimately related to the generalization properties of ANNs [98]) can be quantified in a game-theoretic setting [63] through the observations [63] that (1) regression with the kernel  $K$  is minimax optimal when relative errors in RKHS norm  $\|\cdot\|_K$  are used as a loss, and (2)  $\mathcal{N}(0, K)$  is an optimal mixed strategy for the underlying adversarial game.

**10.4. ODE interpretations of ResNets.** The dynamical systems [90], ODE [29, 19], and diffeomorphism [76, 68] interpretations of ResNets are not new and have inspired the application of numerical approximation methods to the design and training of ANNs. Motivated by the stability of very deep networks [29] proposes to derive ANN architectures from the symplectic integration of the Hamiltonian system

$$\begin{cases} \dot{Y} &= \mathbf{a}(WZ + b) \\ \dot{Z} &= -\mathbf{a}(WY + b) \end{cases} \quad (10.2)$$

where,  $Y$  and  $Z$  are a partition of the features,  $\mathbf{a}$  is an activation function and  $W(t)$  and  $b(t)$  are time-dependent matrices and vectors acting as control parameters. Motivated by the reversibility of the network, [17] proposes to replace (10.2) by a Hamiltonian system of the form

$$\begin{cases} \dot{Y} &= W_1^T \mathbf{a}(W_1 Z + b_1) \\ \dot{Z} &= -W_2^T \mathbf{a}(W_2 Y + b_2) \end{cases} \quad (10.3)$$

where  $W_1$  and  $W_2$  are time dependent convolution matrices acting as control parameters (in addition to  $b_1$  and  $b_2$ ). Motivated by memory efficiency and an explicit control of the speed vs. accuracy tradeoff [19] proposes to use Pontryagin’s adjoint sensitivity method for computing gradients with respect to the parameters of the Network. While ResNets have been interpreted as solving an ODE of the form  $\dot{x} = \mathbf{a}(Wx + b)$  [90, 29], the feature

space formulation of idea registration (Subsec. 6.3) suggests using ODEs of the form  $\dot{x} = W\mathbf{a}(x) + b$ .

**10.5. Warping kernels.** Kernels of the form  $K(\phi(x), \phi(x'))$  defined by a warping of the space  $\phi$  have been employed in numerical homogenization [69] (where they enable upscaling with non separated scales), and in spatial statistics [78, 70, 81, 97] where they enable the nonparameteric estimation of nonstationary and anisotropic spatial covariance structures.

**10.6. Kernel Flows and deep learning without back-propagation.** In the setting of Sec. 2, given an operator-valued kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$ , the Kernel Flows [68, 94, 20, 32] solution to Problem 1 is, in its nonparametric version [68], to approximate  $f^\dagger$  via ridge regression with a kernel of the form  $\mathcal{K}_n(x, x') = K(\phi_n(x), \phi_n(x'))$ . where  $\phi_n : \mathcal{X} \rightarrow \mathcal{X}$  is a discrete flow learned from data via induction on  $n$  with  $\phi_0(x) = x$ . This induction can be described as follows. Let  $\Gamma : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{X})$  be a scalar operator-valued kernel. Set  $q_n = \phi_n(X)$  and

$$\phi_{n+1}(x) = \phi_n(x) + \Delta t \Gamma(\phi_n(x), q_n) p_n, \quad (10.4)$$

which is evidently a discretization of (1.7) with  $v$  of the form (4.10). To identify  $p_n$  let  $q_{n+1} = q_n + \Delta t \Gamma(q_n, q_n) p_n$ . Select  $(q'_{n+1}, Y')$  as a random subset of the (deformed) data  $(q_{n+1}, Y)$ , write  $u_{q_{n+1}}$  for the  $\Gamma$ -interpolant of  $(q_{n+1}, Y)$ ,  $u'_{q_{n+1}}$  for the  $\Gamma$ -interpolant of  $(q'_{n+1}, Y)$ , write  $\rho(q_{n+1}) = \|u_{q_{n+1}} - u'_{q_{n+1}}\|_{\Gamma}^2 / \|u_{q_{n+1}}\|_{\Gamma}^2$  and identify  $p_n$  in the gradient descent direction of  $\rho(q_{n+1}) = \rho(q_n + \Delta t \Gamma(q_n, q_n) p_n)$ . Since no backpropagation is used to identify  $p_n$ , the numerical evidence of the efficacy of this strategy [68] (interpretable as a variant of cross validation [20]) suggests that deep learning could be performed by replacing backpropagation with forward cross-validation.

## 11. The elephant in the dark deep learning room and the shape of ideas

Seeking to develop a theoretical understanding of deep learning can be compared to attempting to describe an elephant in a dark room [5]. Rephrasing [5], ResNets [34] look like discretized ODEs [90, 29, 19, 68], the generalization properties of ANNs [98] feel like those of kernel methods [9, 38, 68], the functional form of ANNs is akin to that of deep kernels [92], there seems to be a natural relation between ANNs and deep Gaussian processes [23]. Backpropagation seems to be solving an optimal control problem [46]. The identification of ANNs, CNNs, and ResNets as algorithms obtained from the discretization of idea/abstraction registration/formation variational problems suggests that (1) ANNs are essentially image registration/computational anatomy algorithms generalized to abstract high dimensional spaces (2) ideas do have shape and forming ideas can be expressed as manipulating their form in abstract RKHS spaces. Evidently, this identification opens the possibility of (1) analyzing deep learning the perspective of shape analysis [96] (2) identifying good architectures by programming good kernels [67]. Although it is difficult to visualize shapes in high dimensional spaces, we suspect that deep learning breaks the curse of dimensionality by (implicitly) employing kernels (such as

REM kernels) exploiting<sup>32</sup> universal patterns/structures in the shape of the data (e.g., the compositional nature of the world and its invariants under transformations). Therefore understanding interplays between learning and shapes/forms in high dimensional spaces may help us see “*the whole of the beast*” [5]. A *beast* that bears some intriguing similarities with Plato’s theory of forms<sup>33</sup> [71].

**Acknowledgments.** The author gratefully acknowledges support by the Air Force Office of Scientific Research under award number FA9550-18-1-0271 (Games for Computation and Learning). Thanks to Clint Scovel for a careful readthrough with detailed comments and feedback.

## References

- [1] Stéphanie Allasonnière, Alain Trouvé, and Laurent Younes. Geodesic shooting and diffeomorphic matching via textured meshes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 365–381. Springer, 2005.
- [2] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.
- [3] Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.
- [4] Julien Arino. Fundamental theory of ordinary differential equations. *Lecture Notes. University of Manitoba*, 2006.
- [5] Coleman Barks. The Essential Rumi. In *Elephant in the Dark*, volume 252. HarperSanFrancisco, 1995.
- [6] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.
- [7] Peter Baxendale. Brownian motions in the diffeomorphism group i. *Compositio Mathematica*, 53(1):19–50, 1984.
- [8] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.
- [9] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.
- [10] Sergio Blanes and Fernando Casas. *A concise introduction to geometric numerical integration*. CRC press, 2017.
- [11] Bastian Bohn, Christian Rieger, and Michael Griebel. A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, 20(64):1–32, 2019.
- [12] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.
- [13] Martins Bruveris, François Gay-Balmaz, Darryl D Holm, and Tudor S Ratiu. The momentum map representation of images. *Journal of nonlinear science*, 21(1):115–150, 2011.

---

<sup>32</sup>The corresponding RKHS norm of the target function should be small. Although Barron space error estimates [6, 25] and RKHS error estimates (Thm. 8.4) do not depend on dimension, they rely on bounding the Barron/RKHS norm of the target function which is the difficulty to be addressed.

<sup>33</sup>According to Plato’s theory of forms, (1) “*Ideas*” or “*Forms*”, are the non-physical essences of all things, of which, objects and matter, in the physical world, are merely imitations ([https://en.wikipedia.org/wiki/Theory\\_of\\_forms](https://en.wikipedia.org/wiki/Theory_of_forms)) and (2) *The world can be divided into two worlds, the visible and the intelligible. We grasp the visible world with our senses. The intelligible world we can only grasp with our mind, it is comprised of the forms . . . Only the forms are objects of knowledge because only they possess the eternal, unchanging truth that the mind, not the senses, must apprehend.* (Randy Aust, <https://www.youtube.com/watch?v=A7xjoHruQfY>).

- [14] Vincent Camion and Laurent Younes. Geodesic interpolating splines. In *International workshop on energy minimization methods in computer vision and pattern recognition*, pages 513–527. Springer, 2001.
- [15] Lapo Casetti, Cecilia Clementi, and Marco Pettini. Riemannian theory of Hamiltonian chaos and Lyapunov exponents. *Physical Review E*, 54(6):5969, 1996.
- [16] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.
- [17] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. Reversible architectures for arbitrarily deep residual neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [18] Nicolas Charon, Benjamin Charlier, and Alain Trounev. Metamorphoses of functional shapes in Sobolev spaces. *Foundations of Computational Mathematics*, 18(6):1535–1596, 2018.
- [19] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.
- [20] Yifan Chen, Houman Owhadi, and Andrew M Stuart. Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation. *arXiv preprint arXiv:2005.11375*, 2020.
- [21] Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.
- [22] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.
- [23] Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.
- [24] Paul Dupuis, Ulf Grenander, and Michael I Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of applied mathematics*, pages 587–600, 1998.
- [25] Weinan E, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.
- [26] Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.
- [27] Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *International conference on machine learning*, pages 3059–3068, 2016.
- [28] Bernard Haasdonk, A Vossen, and Hans Burkhardt. Invariance in kernel methods by Haar-integration kernels. In *Scandinavian Conference on Image Analysis*, pages 841–851. Springer, 2005.
- [29] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.
- [30] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the Störmer–Verlet method. *Acta numerica*, 12:399–450, 2003.
- [31] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.
- [32] Boumediene Hamzi and Houman Owhadi. Learning dynamical systems from data: a simple cross-validation perspective. *arXiv preprint arXiv:2007.05074*, 2020.
- [33] Gabriel L Hart, Christopher Zach, and Marc Niethammer. An optimal control approach for deformable registration. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–16. IEEE, 2009.
- [34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [35] Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.
- [36] Lasse Holmstrom and Petri Koistinen. Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1):24–38, 1992.



- [37] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [38] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
- [39] Sarang C Joshi. *Large deformation diffeomorphisms and Gaussian random fields for statistical characterization of brain sub-manifolds*. PhD thesis, Washington University, 1998.
- [40] Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.
- [41] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.
- [42] Hiroshi Kunita. *Stochastic flows and stochastic differential equations*, volume 24. Cambridge university press, 1997.
- [43] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.
- [44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [45] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.
- [46] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988.
- [47] Tien-Yien Li. Existence of solutions for ordinary differential equations in Banach spaces. *Journal of Differential Equations*, 18(1):29–40, 1975.
- [48] Jerrold E Marsden and Tudor S Ratiu. *Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems*, volume 17. Springer Science & Business Media, 2013.
- [49] Jerrold E Marsden and Matthew West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.
- [50] Mike McKerns. Mystic: a framework for predictive science; SciPy 2013 presentation; <https://www.youtube.com/watch?v=o-nwSnLC6DU&feature=youtu.be&t=74>.
- [51] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- [52] Charles A Micchelli and Massimiliano Pontil. Kernels for multi-task learning. In *Advances in neural information processing systems*, pages 921–928, 2005.
- [53] Mario Micheli. *The differential geometry of landmark shape manifolds: metrics, geodesics, and curvature*. PhD thesis, Brown University, 2008.
- [54] Mario Micheli, Peter W Michor, and David Mumford. Sectional curvature in terms of the cometric, with applications to the riemannian manifolds of landmarks. *SIAM Journal on Imaging Sciences*, 5(1):394–433, 2012.
- [55] Michael I Miller, Alain Trouvé, and Laurent Younes. On the metrics and Euler-Lagrange equations of computational anatomy. *Annual review of biomedical engineering*, 4(1):375–405, 2002.
- [56] Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006.
- [57] Stefan Müller and Michael Ortiz. On the  $\gamma$ -convergence of discrete dynamics and variational integrators. *Journal of Nonlinear Science*, 14(3):279–296, 2004.
- [58] Nicholas H Nelsen and Andrew M Stuart. The random feature model for input-output maps between Banach spaces. *arXiv preprint arXiv:2005.10224*, 2020.
- [59] Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015.
- [60] Houman Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Review*, 59(1):99–149, 2017.



- [61] Houman Owhadi and Clint Scovel. Brittleness of Bayesian inference and new Selberg formulas. *Communications in Mathematical Sciences*, 14(1):83–145, 2016.
- [62] Houman Owhadi and Clint Scovel. Qualitative robustness in Bayesian inference. *ESAIM: Probability and Statistics*, 21:251–274, 2017.
- [63] Houman Owhadi and Clint Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, volume 35. Cambridge University Press, 2019.
- [64] Houman Owhadi, Clint Scovel, and Florian Schäfer. Statistical numerical approximation. *Notices of the AMS*, 2019.
- [65] Houman Owhadi, Clint Scovel, and Tim Sullivan. On the brittleness of Bayesian inference. *SIAM Review*, 57(4):566–582, 2015.
- [66] Houman Owhadi, Clint Scovel, Tim Sullivan, et al. Brittleness of bayesian inference under finite information in a continuous world. *Electronic Journal of Statistics*, 9(1):1–79, 2015. arXiv:1304.6772 (April 2013).
- [67] Houman Owhadi, Clint Scovel, and Gene Ryan Yoo. Kernel mode decomposition and programmable/interpretable regression networks. *arXiv preprint arXiv:1907.08592*, 2019.
- [68] Houman Owhadi and Gene Ryan Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.
- [69] Houman Owhadi and Lei Zhang. Metric-based upscaling. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 60(5):675–723, 2007.
- [70] O Perrin and P Monestiez. Modelling of non-stationary spatial structure using parametric radial basis deformations. In *GeoENV II—Geostatistics for Environmental Applications*, pages 175–186. Springer, 1999.
- [71] Plato. *The Republic*, volume VII. 375 BCE.
- [72] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.
- [73] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.
- [74] Marco Reiser and Hans Burkhardt. Learning equivariant functions with matrix valued kernels. *Journal of Machine Learning Research*, 8(Mar):385–408, 2007.
- [75] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *Stat*, 1050:22, 2018.
- [76] François Rousseau and Ronan Fablet. Residual networks as geodesic flows of diffeomorphisms. *arXiv preprint arXiv:1805.09585*, 2018.
- [77] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.
- [78] Paul D Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.
- [79] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse cholesky factorization by Kullback-Leibler minimization. *arXiv preprint arXiv:2004.14455*, 2020.
- [80] Florian Schäfer, Timothy John Sullivan, and Houman Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *arXiv preprint arXiv:1706.02205*, 2017.
- [81] Alexandra M Schmidt and Anthony O’Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758, 2003.
- [82] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

- [83] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- [84] Georg Still. Lectures on parametric optimization: An introduction. *Optimization Online*, 2018.
- [85] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [86] Molei Tao. Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance. *Physical Review E*, 94(4):043303, 2016.
- [87] Eduardo V Teixeira. Strong solutions for differential equations in abstract spaces. *Journal of Differential Equations*, 214(1):65–91, 2005.
- [88] Alain Trounev. Diffeomorphisms groups and pattern matching in image analysis. *International journal of computer vision*, 28(3):213–221, 1998.
- [89] François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241, 2012.
- [90] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.
- [91] Matthew West. *Variational integrators*. PhD thesis, California Institute of Technology, 2004.
- [92] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.
- [93] Zong-min Wu and Robert Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA journal of Numerical Analysis*, 13(1):13–27, 1993.
- [94] Gene Ryan Yoo and Houman Owhadi. Deep regularization and direct training of the inner layers of neural networks with kernel flows. *arXiv preprint arXiv:2002.08335*, 2020.
- [95] Laurent Younes. Computable elastic distances between shapes. *SIAM Journal on Applied Mathematics*, 58(2):565–586, 1998.
- [96] Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer, 2010.
- [97] Andrew Zammit-Mangion, Tin Lok James Ng, Quan Vu, and Maurizio Filippone. Deep compositional spatial models. *arXiv preprint arXiv:1906.02840*, 2019.
- [98] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.