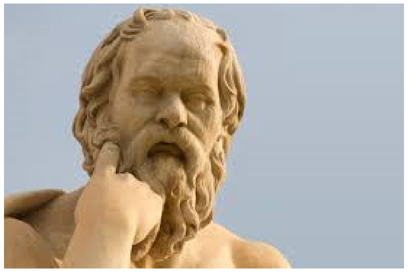
Do ideas have shape? Plato's theory of forms as the continuous limit of artificial neural networks

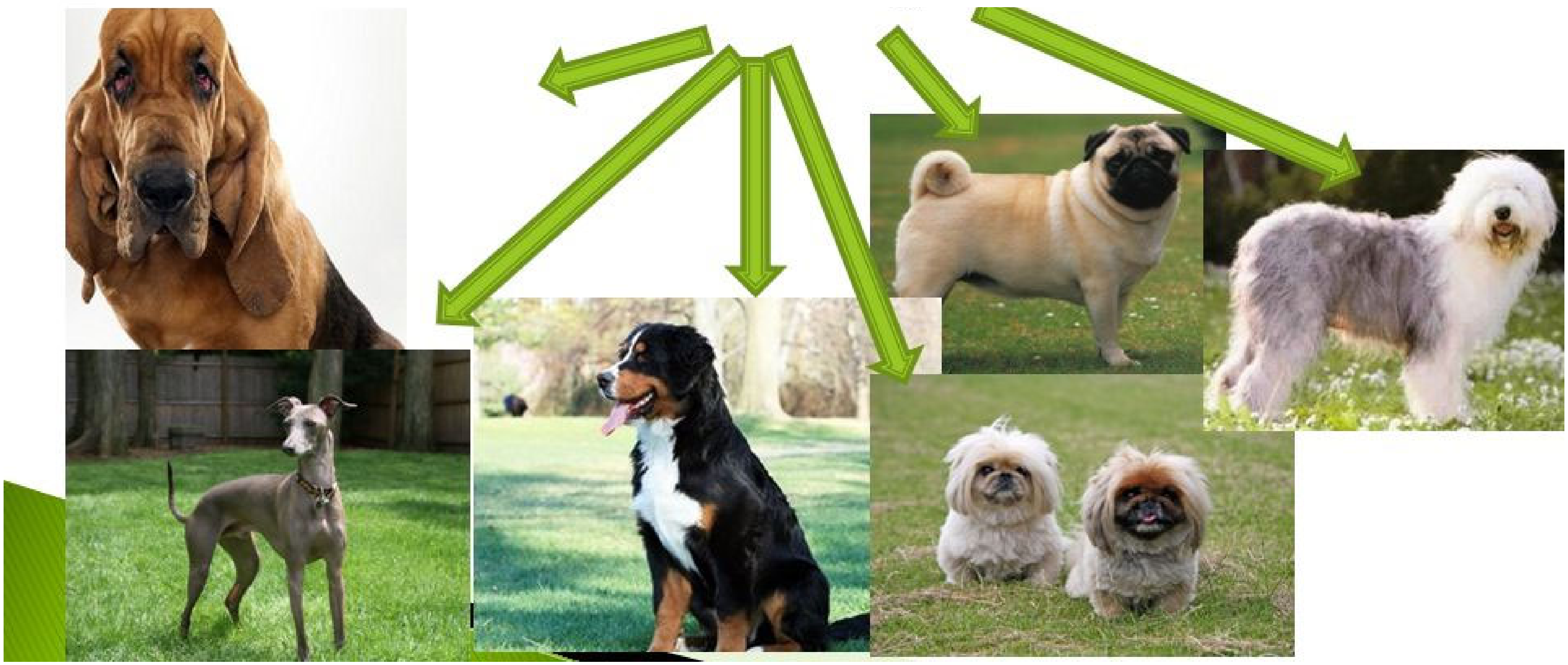
Houman Owhadi

AFOSR. Grant number FA9550-18-1-0271.
Games for Computation and Learning, 2018-2021.



Socrates

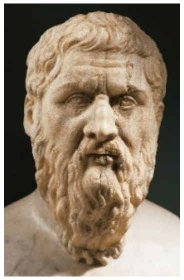
How do we know that these are all dogs?



https://slideplayer.com/slide/10637983/

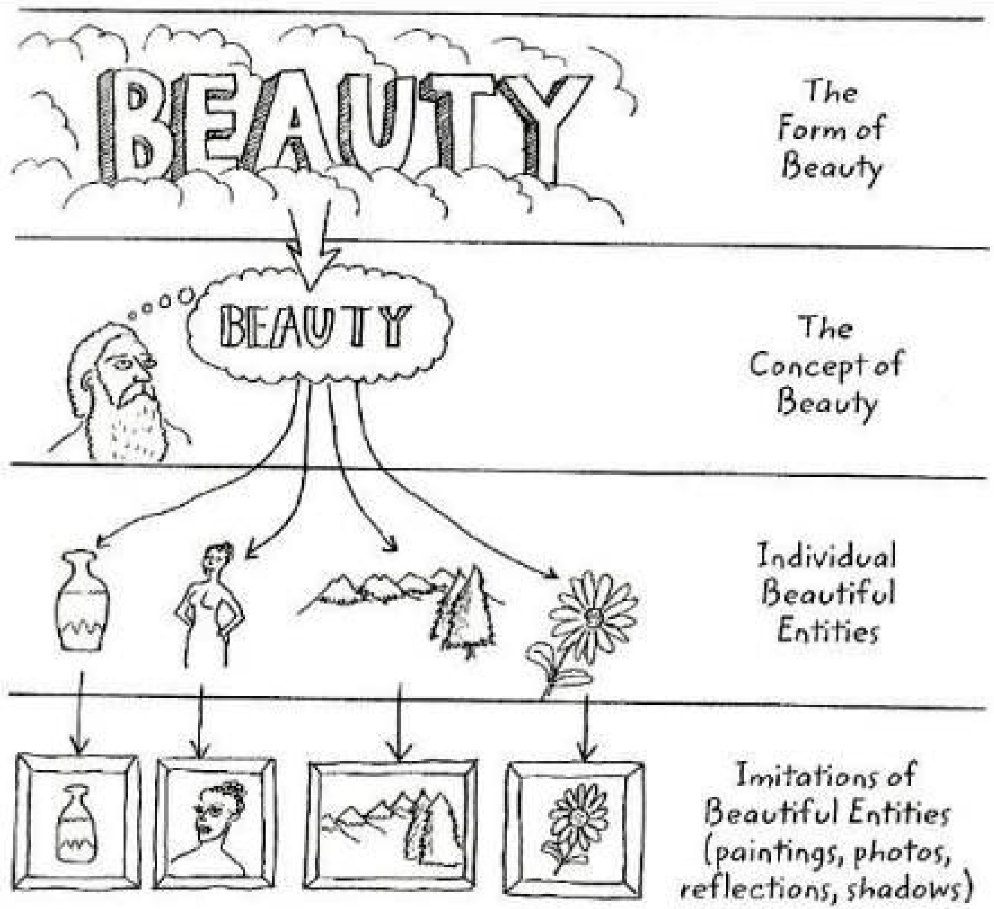
Plato's allegory of the cave

https://www.studiobinder.com/blog/platos-allegory-of-the-cave/

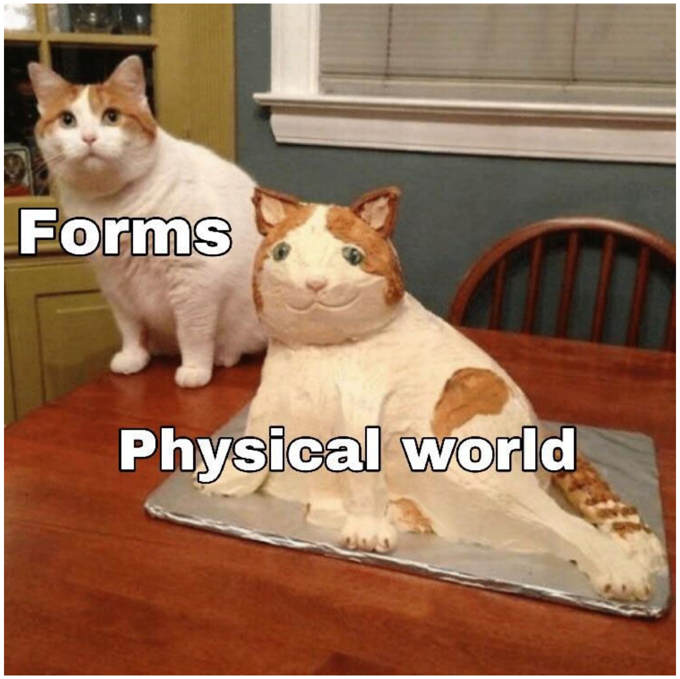


The world can be divided into two worlds, the visible and the intelligible. We grasp the visible world with our senses. The intelligible world we can only grasp with our mind, it is the world of abstractions or ideas

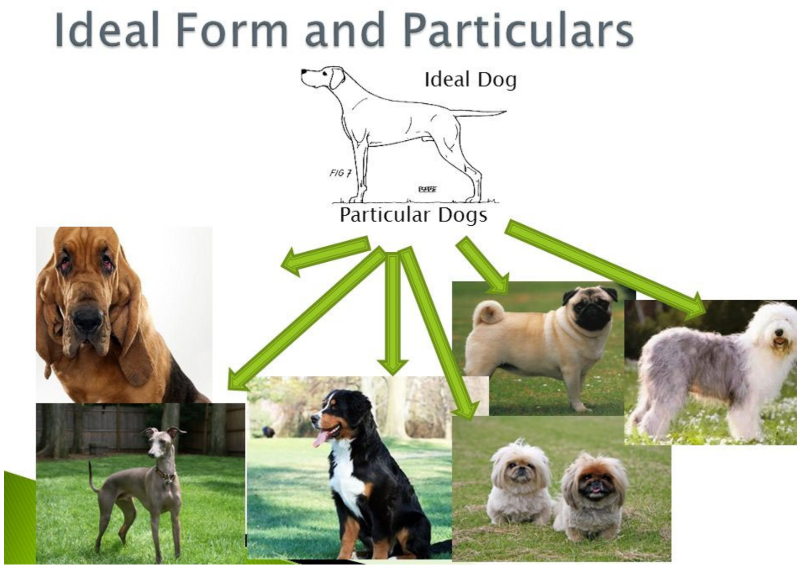
Plato’s theory of forms



https://twitter.com/PhilosophyMttrs



reddit/PhilosophyMemes

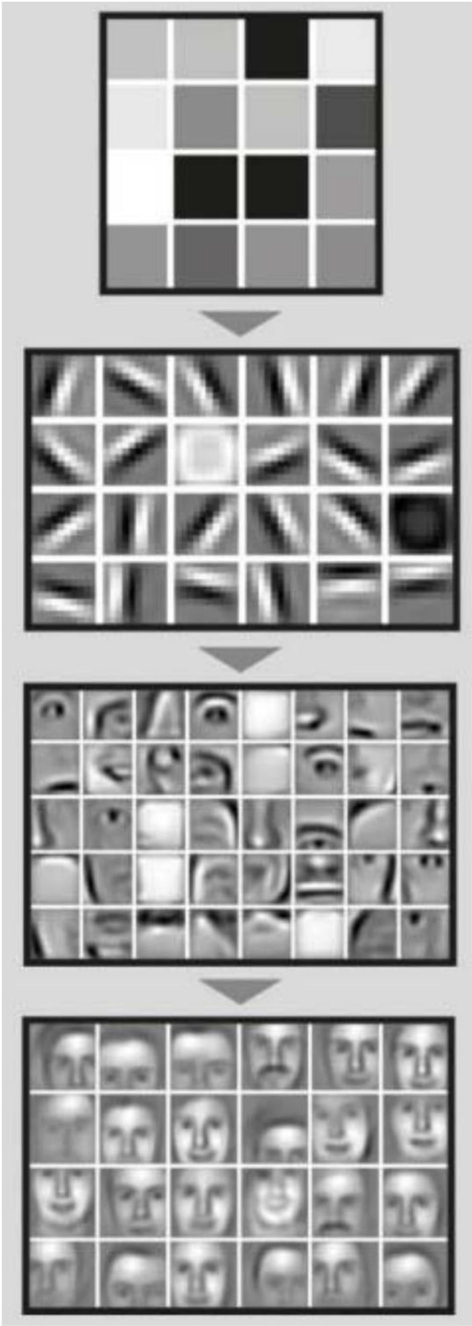


https://slideplayer.com/slide/10637983/

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”*

`https://www.etymonline.com/word/idea`

What does that have to do with Deep Learning?



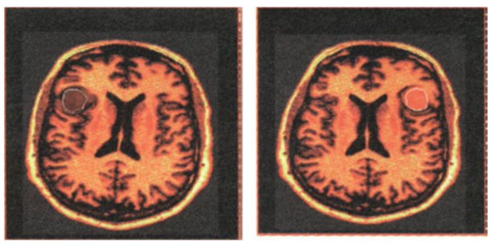
Andrew Ng Source: http://www.nature.com/news/computer-science-the-learning-machines-1.14481

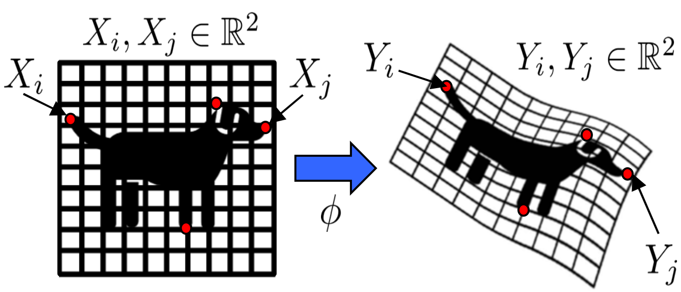
Main message

ANNs are are essentially discretized solvers for a generalization of image registration/computational anatomy variational problems.

Image registration

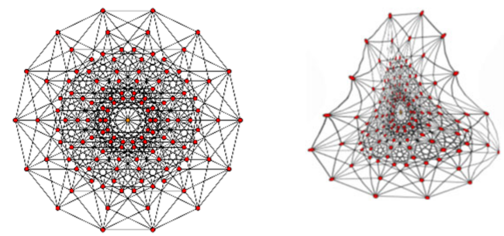
Images

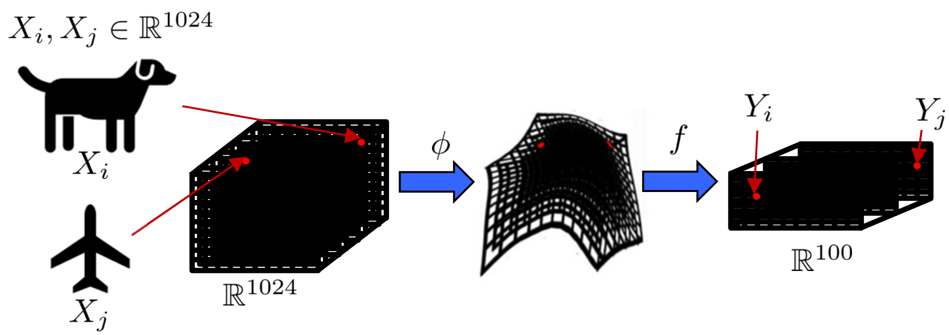




Generalization

High dimensional RKHS





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.

Problem

$$\mathcal{X} \xrightarrow{\quad f^{\dagger} \quad} \mathcal{Y}$$

$f^{\dagger}$: Unknown

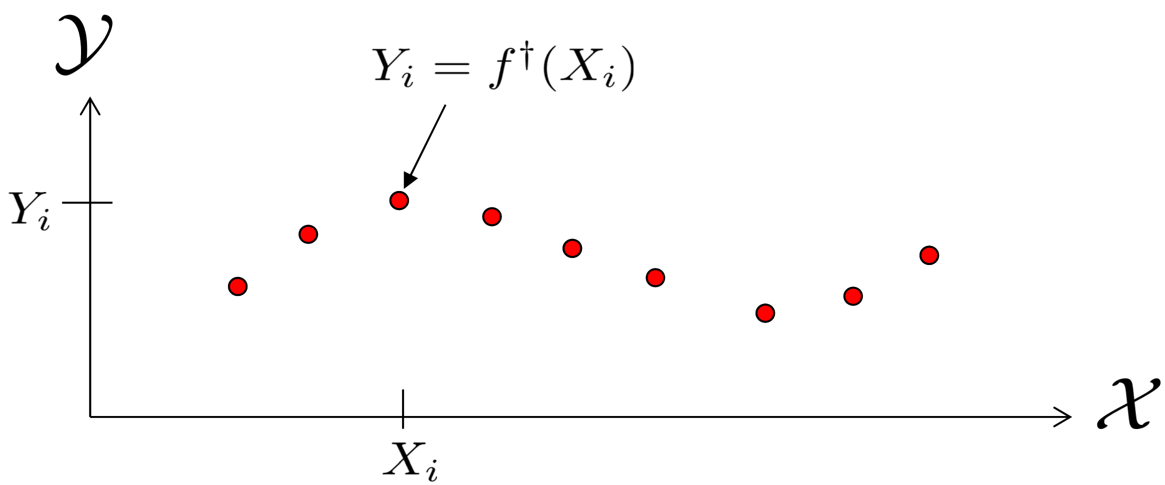
Given $f^{\dagger}(X) = Y$ with $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^{\dagger}$

$\mathcal{X}$, $\mathcal{Y}$: Finite-dimensional Hilbert spaces

$X := (X_1, \ldots, X_N) \in \mathcal{X}^N$

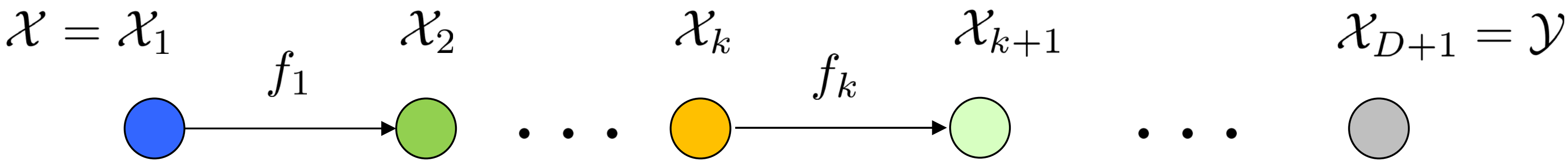
$f^{\dagger}(X) := \big(f^{\dagger}(X_1), \ldots, f^{\dagger}(X_N)\big) \in \mathcal{Y}^N$

$Y := (Y_1, \ldots, Y_N) \in \mathcal{Y}^N$



Artificial neural network solution Approximate $f^\dagger$ with

$$f = f_D \circ \cdots \circ f_1$$



$$f_k(x) = \mathbf{a}(W_k x + b_{k+1})$$

$\mathbf{a}$: Activation function / Elementwise nonlinearity

$\mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})$: Set of bounded linear operators from $\mathcal{X}_k$ to $\mathcal{X}_{k+1}$

$W_k \in \mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})$, $b_{k+1} \in \mathcal{X}_{k+1}$ identified as minimizers of

$$\min_{W_k, b_k} \quad \|f(X) - Y\|^2_{\mathcal{Y}^N}$$

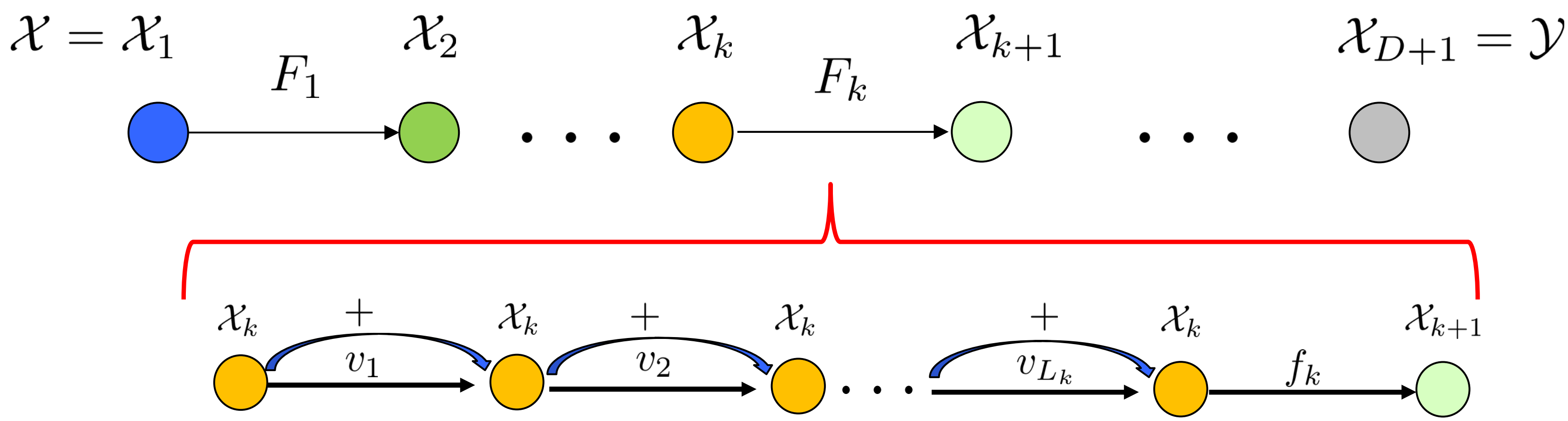
$$\|Y\|^2_{\mathcal{Y}^N} := \sum_{i=1}^N \|Y_i\|^2_{\mathcal{Y}}$$

Residual neural network solution

Approximate $f^\dagger$ with

[He et Al, 2016]

$$f = F_D \circ \cdots \circ F_1$$



$$F_k = f_k \circ (I + v^k_{L_k}) \circ \cdots \circ (I + v^k_1)$$

$$f_k : \mathcal{X}_k \to \mathcal{X}_{k+1} \qquad f_k(x) = \mathbf{a}(W_k x + b_{k+1})$$

$$v^k_s : \mathcal{X}_k \to \mathcal{X}_k \qquad v^s_k(x) = \mathbf{a}(W^s_k x + b^s_k)$$

$$\min_{W_k, b_k, W^s_k, b^s_k} \quad \|f(X) - Y\|^2_{\mathcal{Y}^N}$$

ODE/Dynamical system interpretation of ResNets

[E, 2017], [Haber, Ruthotto, 2017], [Chen, Rubanova, Bettencourt, Duvenaud, 2018], [Chang, Meng, Haber, Ruthotto, Begert, Holtham, 2018]

$(I + v^k_{L_k}) \circ \cdots \circ (I + v^k_1)(x_0)$ is a discrete approximation of $x(1)$

$$\begin{cases} \dot{x} = \mathbf{a}(Wx + b) \\ x(0) = x_0 \end{cases}$$

for some $t \to W(t)$, $b(t)$

[Haber, Ruthotto, 2017]: Use a Hamiltonian ODE and discretize with a symplectic integrator to ensure stability

$$\begin{cases} \dot{y} = \mathbf{a}(Wz + b) \\ \dot{z} = -\mathbf{a}(Wy + b) \end{cases}$$

[Chang et Al, 2018]: The following Hamiltonian system ensures stability + reversibility

$$\begin{cases} \dot{y} = W_1^T \mathbf{a}(W_1 z + b_1) \\ \dot{z} = -W_2^T \mathbf{a}(W_2 y + b_2) \end{cases}$$

Problem

$$\mathcal{X} \xrightarrow{\quad f^\dagger \quad} \mathcal{Y}$$

$f^\dagger$: Unknown

Given $f^\dagger(X) = Y$ with $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^\dagger$

Kernel method solutions

Approximate $f^\dagger$ with

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

$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is an **operator valued kernel**

Operator valued kernels

[Kadri et Al, 2016]: Operator-valued kernels
[Alvarez et Al, 2012]: Vector-valued kernels

$\mathcal{X}$, $\mathcal{Y}$: Separable Hilbert spaces
$\mathcal{L}(\mathcal{Y})$: Set of bounded linear operators on $\mathcal{Y}$.

Definition

$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is an **operator valued kernel** if

(1) $K(x, x') = K(x', x)^T$ where A^T is transpose of A w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$

(2) $\sum_{i,j=1}^{m} \langle y_i, K(x_i, x_j) y_j \rangle_{\mathcal{Y}} \geq 0$ for $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Definition

$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is **scalar** if

$$K(x, x') = k(x, x')\, I_{\mathcal{Y}}$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ scalar valued kernel

$I_{\mathcal{Y}}$: Identity operator on $\mathcal{Y}$

Reproducing Kernel Hilbert Space

$$\mathcal{H} := \text{Closure}\,\text{Span}\{z \to K(z,x)y \mid (x,y) \in \mathcal{X} \times \mathcal{Y}\}$$

Hilbert space of continuous functions mapping $\mathcal{X}$ to $\mathcal{Y}$

RKHS norm

$$\|\textstyle\sum_i K(\cdot, x_i) y_i\|_{\mathcal{H}}^2 = \sum_{i,j} \left\langle y_i, K(x_i, x_j) y_j \right\rangle_{\mathcal{Y}}$$

Reproducing identity

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

Write $\|f\|_K^2 := \|f\|_{\mathcal{H}}^2$

Feature map $\mathcal{F}$: Separable Hilbert space

$$\psi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{F})$$

Definition

$\mathcal{F}$ and ψ are a **feature space** and a **feature map** for the kernel K if

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

$$\Updownarrow$$

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

$$\psi^T : \mathcal{X} \to \mathcal{L}(\mathcal{F}, \mathcal{Y})$$

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

Theorem

$$\mathcal{H} := \mathrm{Span}\{\psi^T \alpha \mid \alpha \in \mathcal{F}\}$$

$$\|\psi^T \alpha\|_{\mathcal{H}}^2 = \|\alpha\|_{\mathcal{F}}^2$$

Problem

$$\mathcal{X} \xrightarrow{\quad f^\dagger \quad} \mathcal{Y}$$

$f^\dagger$: Unknown

Given $f^\dagger(X) = Y$ with $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^\dagger$

Optimal recovery solution

Approximate $f^\dagger$ with minimizer of

$$\begin{cases} \text{Minimize} & \|f\|_K \\ \text{subject to} & f(X) = Y \end{cases}$$

$$f(x) = K(x, X) K(X, X)^{-1} Y$$

$K(X, X)$: $N \times N$ block matrix with blocks $K(X_i, X_j)$

$K(x, X)$: $1 \times N$ block vector with blocks $K(x, X_i)$

Theorem [Micchelli and Rivlin, 1977] [O. and Scovel, 2019]

f is minimax optimal if loss = relative error in $\|\cdot\|_K$-norm

$$f = \operatorname{argmin} \min_f \max_{f^\dagger | f^\dagger(X)=Y} \frac{\|f^\dagger - f\|_K^2}{\|f^\dagger\|_K^2}$$

Theorem [Myers, 1992] [O., 2005]

$$\big\|f^\dagger(x) - f(x)\big\|_{\mathcal{Y}} \leq \sigma(x)\|f^\dagger\|_K$$

$$\sigma^2(x) := \operatorname{Trace}\big[K(x,x) - K(x,X)K(X,X)^{-1}K(X,x)\big]$$

Does not depend on dimension!
But need to bound $\|f^\dagger\|_K$ to be useful

Problem

$$\mathcal{X} \xrightarrow{\quad f^\dagger \quad} \mathcal{Y}$$

$f^\dagger$: Unknown

Given $f^\dagger(X) = Y$ with $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^\dagger$

Ridge regression solution

Approximate $f^\dagger$ with minimizer of

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

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

Theorem [O., Scovel and Yoo 2019]

f is minimax optimal in the setting of Tikhonov regularization/mode decomposition

$$f = \operatorname{argmin} \min_f \max_{f^\dagger} \frac{\lambda\|f^\dagger - f\|_K^2 + \|f^\dagger(X) - f(X)\|_{\mathcal{Y}^N}^2}{\lambda\|f^\dagger\|_K^2 + \|f^\dagger(X) - Y\|_{\mathcal{Y}^N}^2}$$

Theorem [O. 2020]

$$\left\|f^\dagger(x) - f(x)\right\|_{\mathcal{Y}} \leq \sigma(x)\|f^\dagger\|_K$$

$$\sigma^2(x) := \operatorname{Trace}\left[K(x,x) - K(x,X)\big(K(X,X) + \lambda I\big)^{-1} K(X,x)\right]$$

Mechanical regression

Approximate $f^\dagger$ with

$$f^\ddagger = f \circ \phi_L$$

$$\phi_L : \mathcal{X} \to \mathcal{X}$$
$$\phi_L = (I + v_L) \circ \cdots \circ (I + v_1)$$

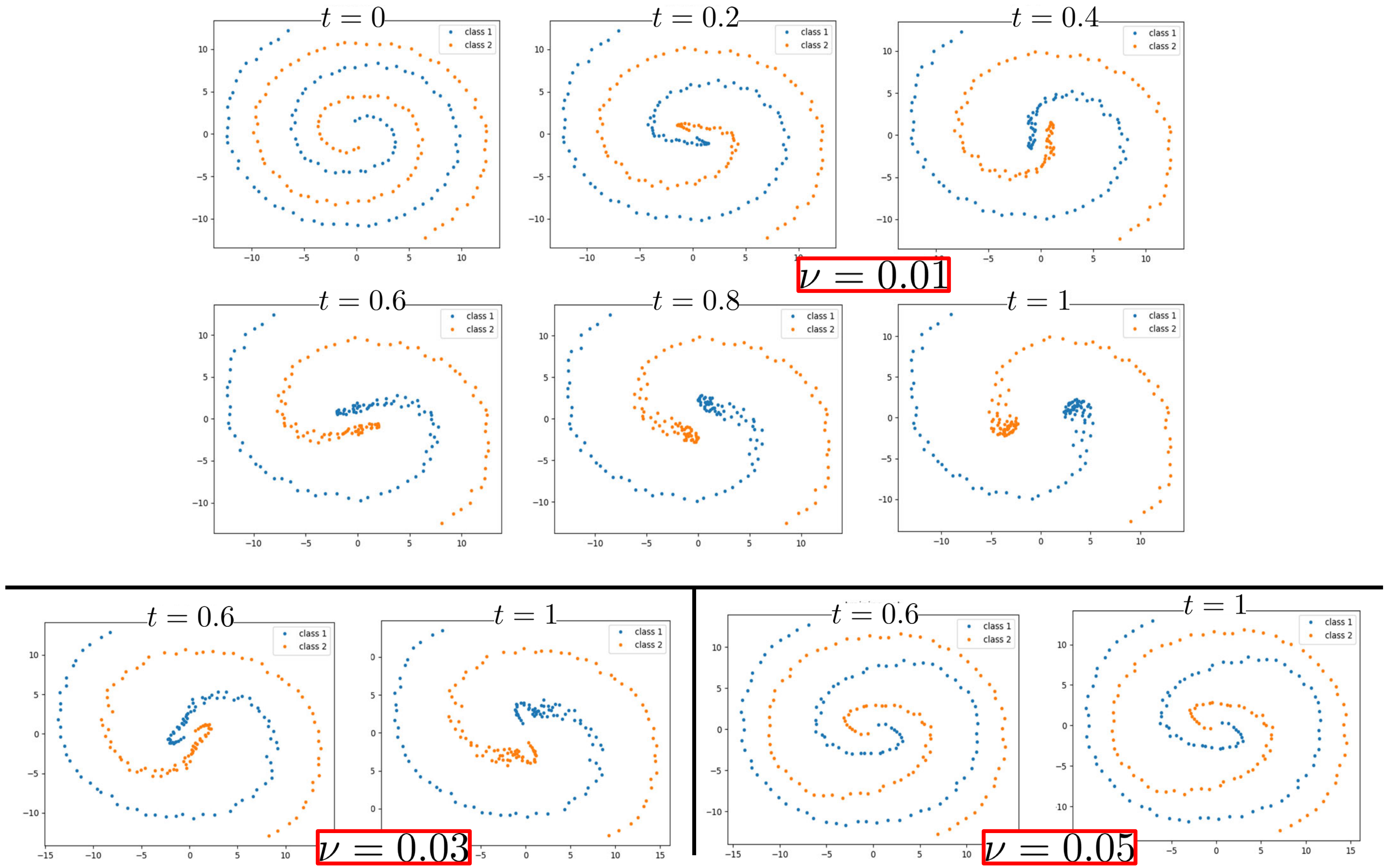
$f : \mathcal{X} \to \mathcal{Y}$ and $v_s : \mathcal{X} \to \mathcal{X}$ identified as minimizers of

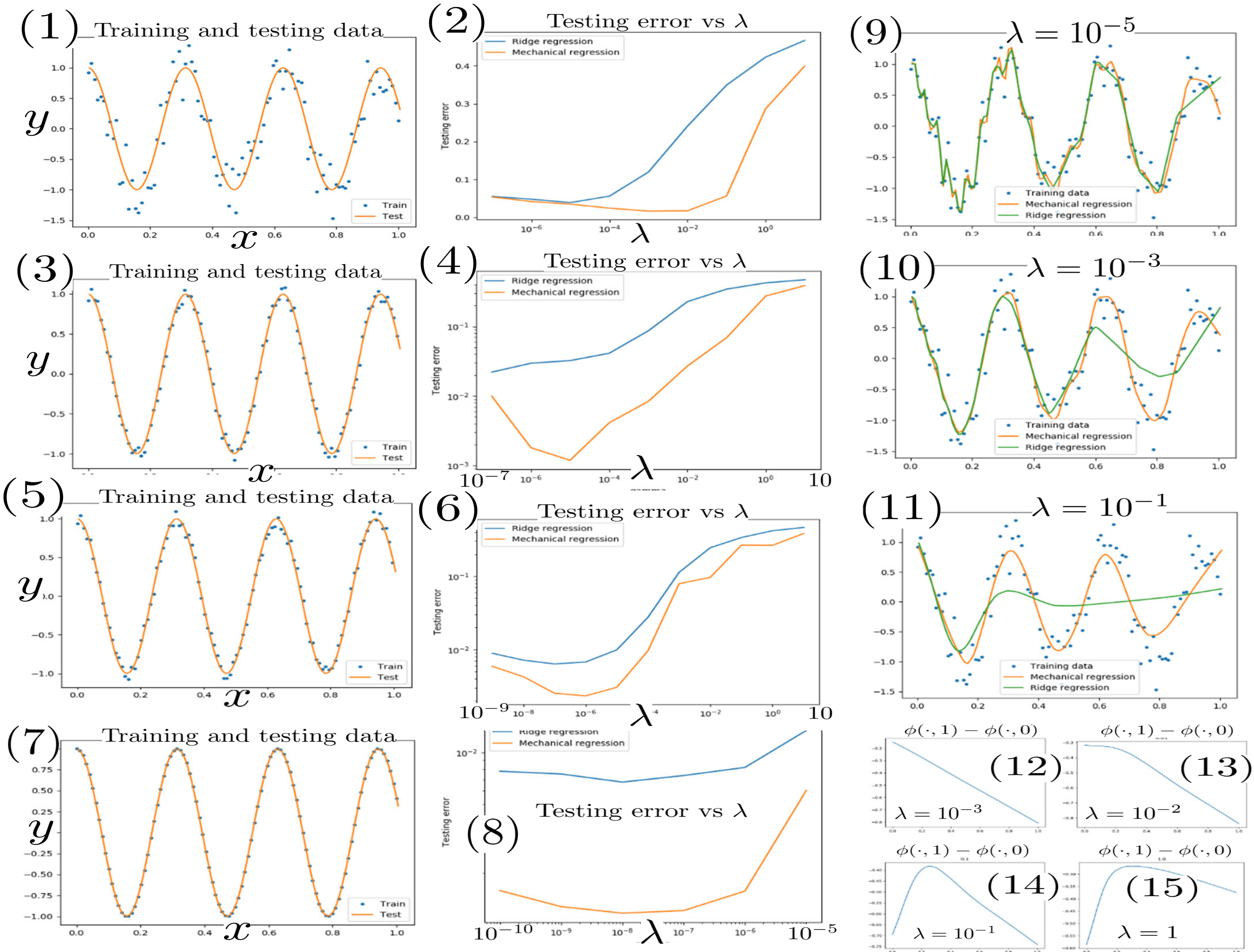
$$\min_{f, v_1, \ldots, v_L} \frac{\nu L}{2} \sum_{s=1}^{L} \|v_s\|_\Gamma^2 + \lambda \|f\|_K^2 + \|f \circ \phi_L(X) - Y\|_{\mathcal{Y}^N}^2$$

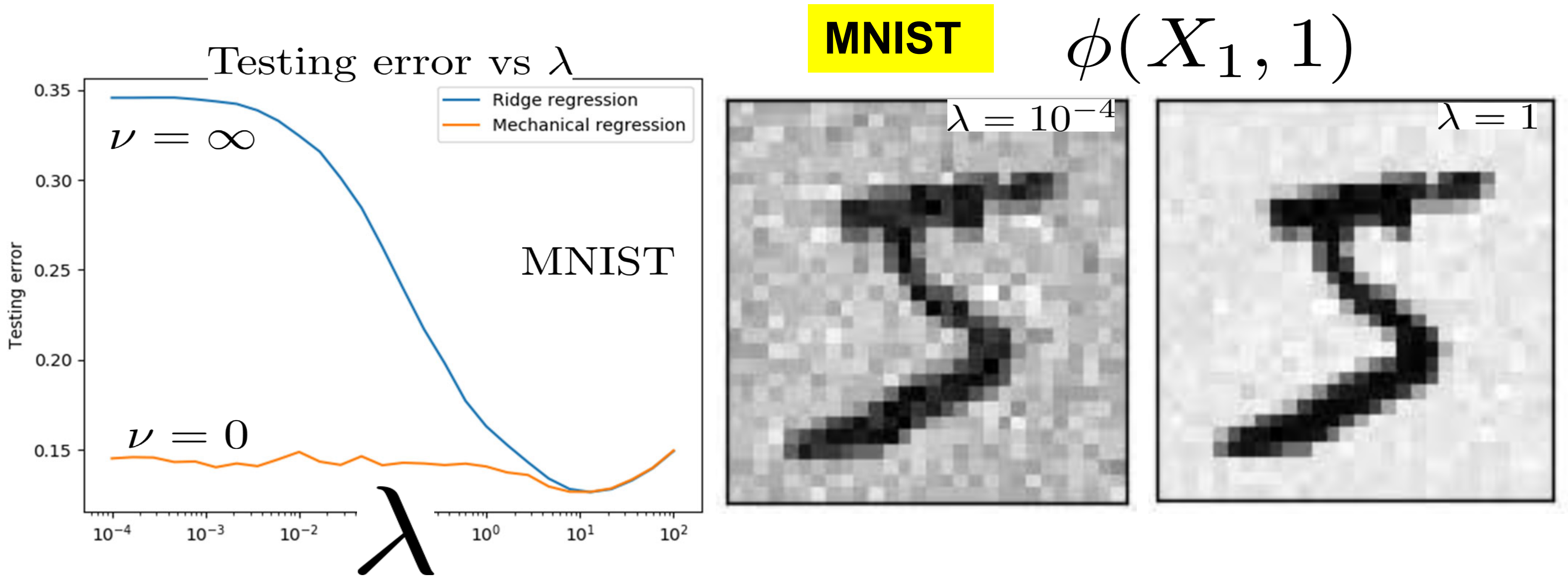
$$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$$
$$\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{X})$$

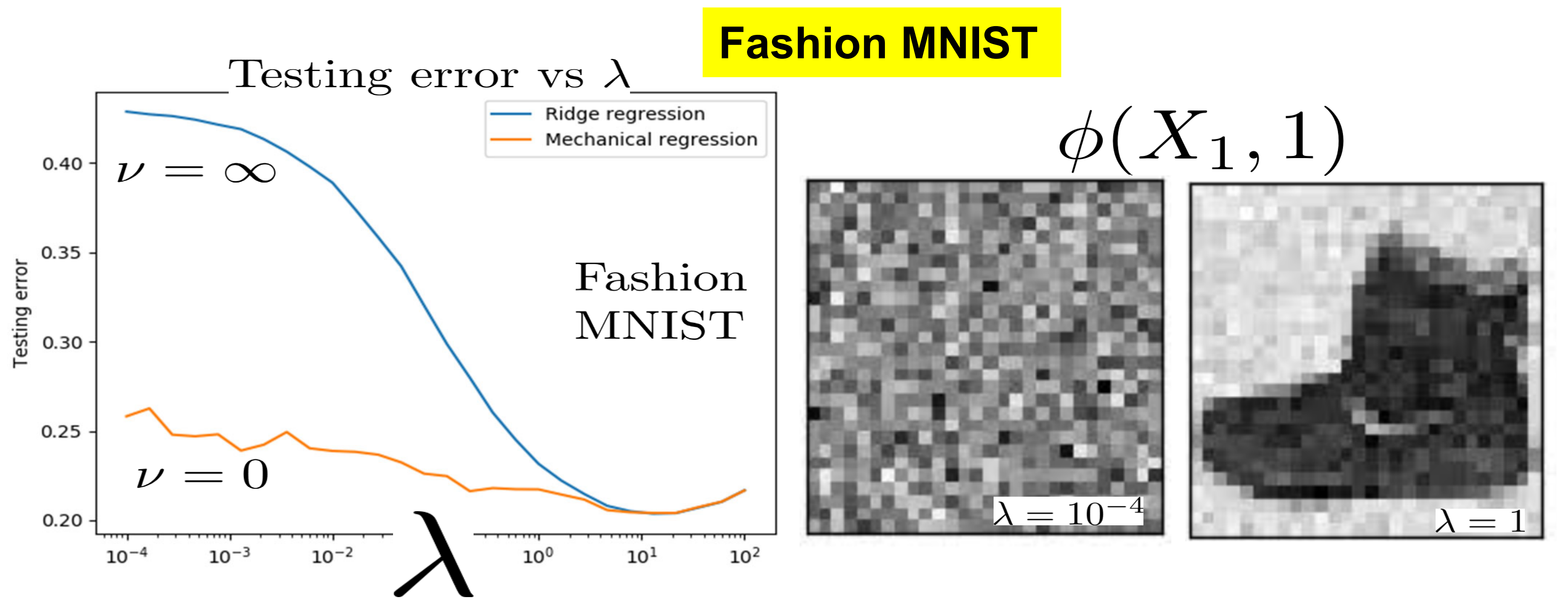
Numerical experiments

$$\phi_{[tL]}(X) = (I + v_{[Lt]}) \circ \cdots \circ (I + v_1)(X)$$




(1) Training and testing data
Train
Test
(2) Testing error vs λ
Ridge regression
Mechanical regression
Testing error
(3) Training and testing data
(4) Testing error vs λ
(5) Training and testing data
(6) Testing error vs λ
(7) Training and testing data
(8) Testing error vs λ
(9) λ = 10^{-5}
Training data
(10) λ = 10^{-3}
(11) λ = 10^{-1}
(12) φ(·, 1) − φ(·, 0) λ = 10^{-3}
(13) φ(·, 1) − φ(·, 0) λ = 10^{-2}
(14) φ(·, 1) − φ(·, 0) λ = 10^{-1}
(15) φ(·, 1) − φ(·, 0) λ = 1


MNIST
$\phi(X_1, 1)$
Testing error vs λ
Ridge regression
Mechanical regression
Testing error
$\nu = \infty$
$\nu = 0$
MNIST
0.35
0.30
0.25
0.20
0.15
10^{-4}
10^{-3}
10^{-2}
10^{0}
10^{1}
10^{2}
λ
$\lambda = 10^{-4}$
$\lambda = 1$


Fashion MNIST
$\phi(X_1, 1)$
Testing error vs λ
Ridge regression
Mechanical regression
Testing error
$\nu = \infty$
$\nu = 0$
Fashion
MNIST
0.40
0.35
0.30
0.25
0.20
10^{-4}
10^{-3}
10^{-2}
10^{0}
10^{1}
10^{2}
λ
$\lambda = 10^{-4}$
$\lambda = 1$

Mechanical regression

Approximate $f^\dagger$ with

$$f^\ddagger = f \circ \phi_L$$

$$\phi_L : \mathcal{X} \to \mathcal{X}$$

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

$f : \mathcal{X} \to \mathcal{Y}$ and $v_s : \mathcal{X} \to \mathcal{X}$ identified as minimizers of

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

$$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$$

$$\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{X})$$

Particular case

Let Γ and K be scalar operator valued kernels defined by the scalar kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\Gamma(x, x') = k(x, x')\, I_{\mathcal{X}} \qquad K(x, x') = k(x, x')\, I_{\mathcal{Y}}$$

Let k have feature space $\mathcal{X} \oplus \mathbb{R}$ and feature map $\boldsymbol{\varphi}$.

$$k(x, x') = \boldsymbol{\varphi}^T(x)\boldsymbol{\varphi}(x') \qquad \boldsymbol{\varphi} : \mathcal{X} \to \mathcal{X} \oplus \mathbb{R}$$

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

$\tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})$ and $w_s \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})$

Particular case: ResNet block

$$\boldsymbol{\varphi} : \mathcal{X} \to \mathcal{X} \oplus \mathbb{R}$$

Let $\boldsymbol{\varphi}(x) = (\mathbf{a}(x), 1) \leftarrow$ always active neuron

$\mathbf{a} : \mathcal{X} \to \mathcal{X}$ $\quad \mathbf{a}(x)$: Activation function

$\tilde{w}\boldsymbol{\varphi}(x) = W\mathbf{a}(x) + b$ $\quad W \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$: weights

$b \in \mathcal{Y}$: bias

$w_s\boldsymbol{\varphi}(x) = W_s\mathbf{a}(x) + b_s$ $\quad W_s \in \mathcal{L}(\mathcal{X})$: weights

$b_s \in \mathcal{X}$: bias

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

$\Updownarrow$ $\quad \tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})$ and $w_s \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})$

$$f \circ \phi_L(x) = (W\mathbf{a}(\cdot) + b) \circ (I + W_L\mathbf{a}(\cdot) + b_L) \circ \cdots \circ (I + W_1\mathbf{a}(\cdot) + b_1)$$

$$\Gamma(x,x') = \boldsymbol{\varphi}^T(x)\boldsymbol{\varphi}(x')I_{\mathcal{X}}$$

$$K(x,x') = \boldsymbol{\varphi}^T(x)\boldsymbol{\varphi}(x')I_{\mathcal{Y}}$$

$$\boldsymbol{\varphi}(x) = \big(\mathbf{a}(x), 1\big) \qquad \boldsymbol{\varphi} : \mathcal{X} \to \mathcal{X} \oplus \mathbb{R}$$

$\mathbf{a}(x)$: Activation function $\qquad \mathbf{a} : \mathcal{X} \to \mathcal{X}$

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

$\tilde{w} \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{Y})$ and $w_s \in \mathcal{L}(\mathcal{X} \oplus \mathbb{R}, \mathcal{X})$ minimizers of

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

This is one ResNet block with L2 regularization on weights and biases!

Mechanical regression

Approximate $f^\dagger$ with

$$f^\ddagger = f \circ \phi_L$$

$$\phi_L : \mathcal{X} \to \mathcal{X}$$
$$\phi_L = (I + v_L) \circ \cdots \circ (I + v_1)$$

$f : \mathcal{X} \to \mathcal{Y}$ and $v_s : \mathcal{X} \to \mathcal{X}$ identified as minimizers of

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

$$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$$
$$\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{X})$$

Theorem

As $L \to \infty$, adherence values of $f \circ \phi_L(x)$ are

$$f \circ \phi^v(x)$$

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

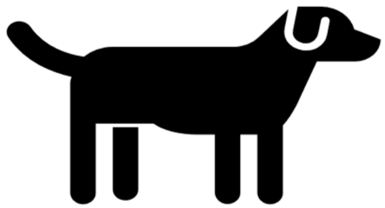
$v : \mathcal{X} \times [0,1] \to \mathcal{X}$ and $f : \mathcal{X} \to \mathcal{Y}$ are minimizers of

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

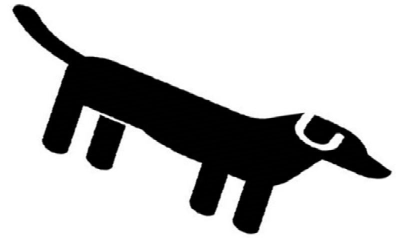
Looks like an image registration/computational anatomy variational problem

Image registration

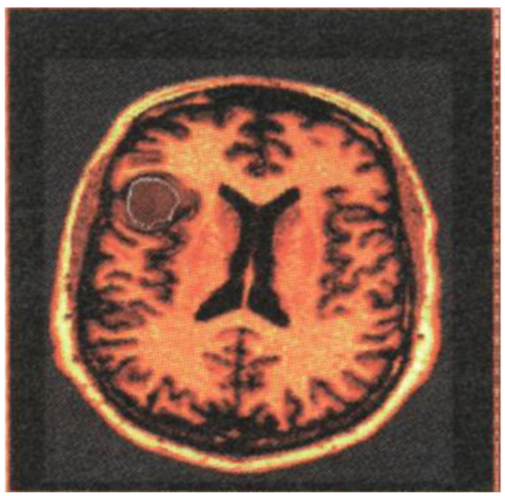
How to best align image I and image I'?



I



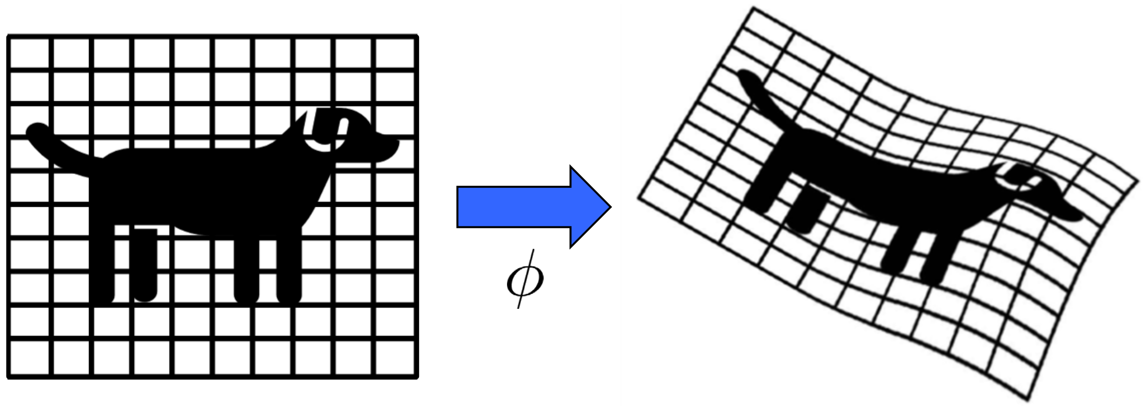
I'



[Grenander, Miller, 1998]: Computational anatomy

[Joshi, Miller, 2000], [Micheli, 2008], [Beg, Miller, Trouvé, Younes, 2005], [Dupuis, Grenander, Miller, 1998], [Vialard, Risser, Rueckert, Cotter, 2012].

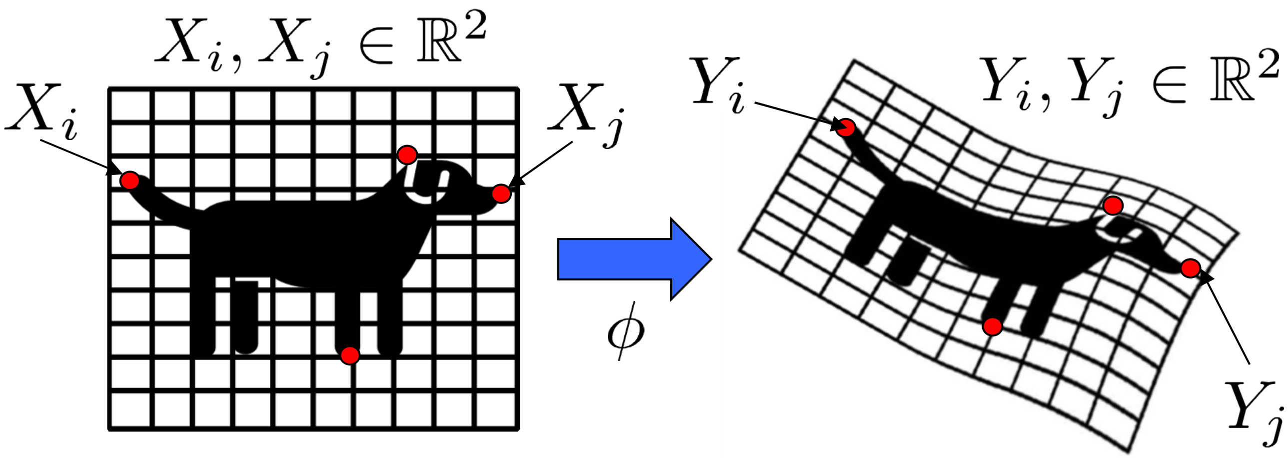
Image registration



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

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

Image registration with landmarks

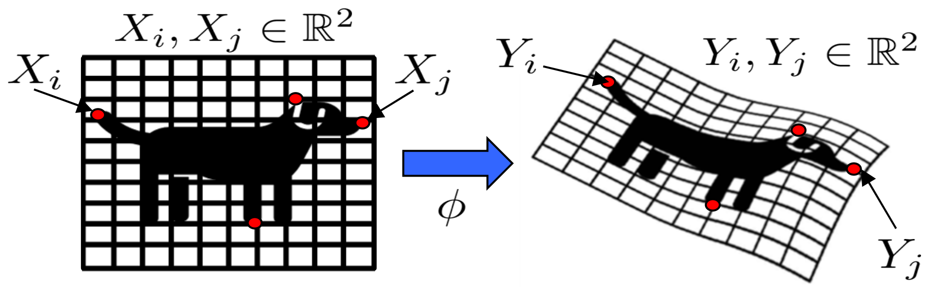


$$\min_v \lambda \int_0^1 \|\Delta v\|^2_{L^2([0,1]^2)}\, dt + \sum_i |\phi^v(X_i,1) - Y_i|^2$$

$$\begin{cases} \dot\phi(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

[Joshi, Miller, 2000]: Landmark matching

Image registration with landmark matching



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

$$\begin{cases} \dot{\phi}(x,t) = v(\phi(x,t), t) \\ \phi(x,0) = x \end{cases}$$

Idea registration with data matching

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

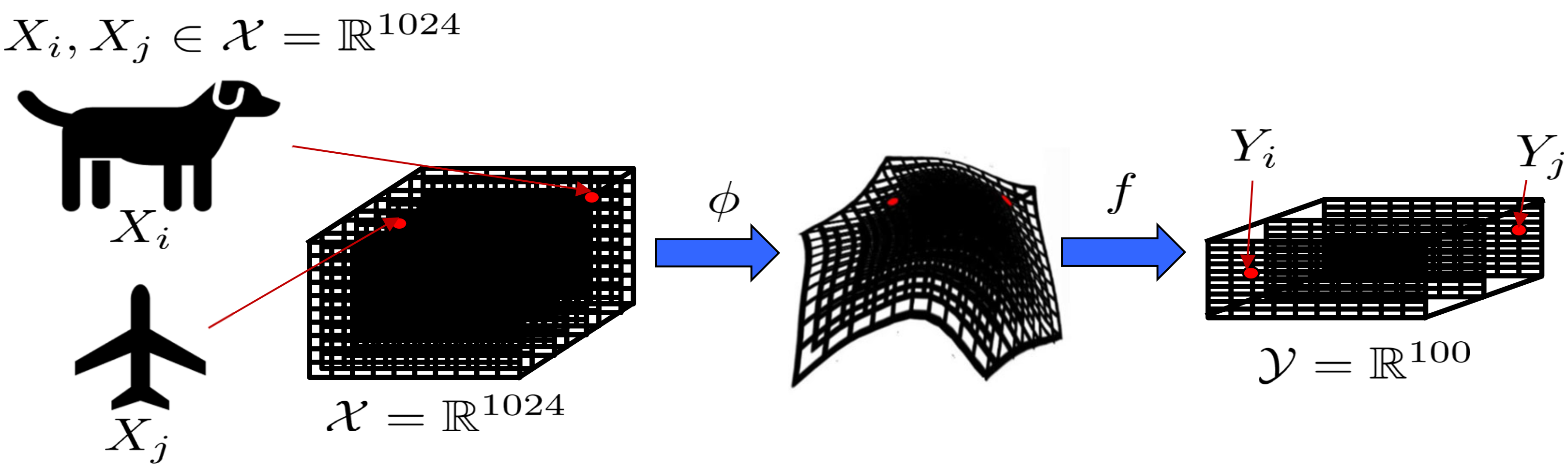


Image registration with landmark matching

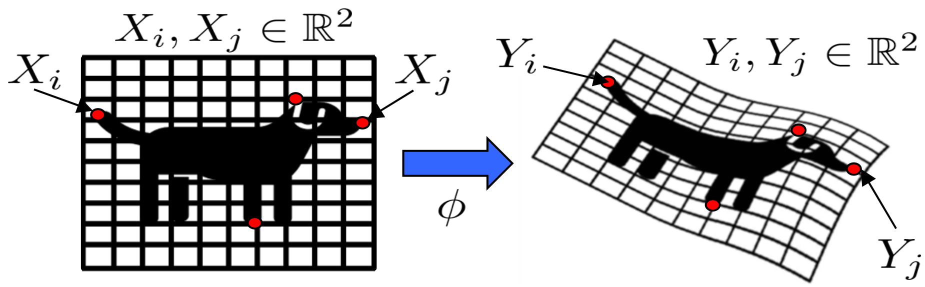
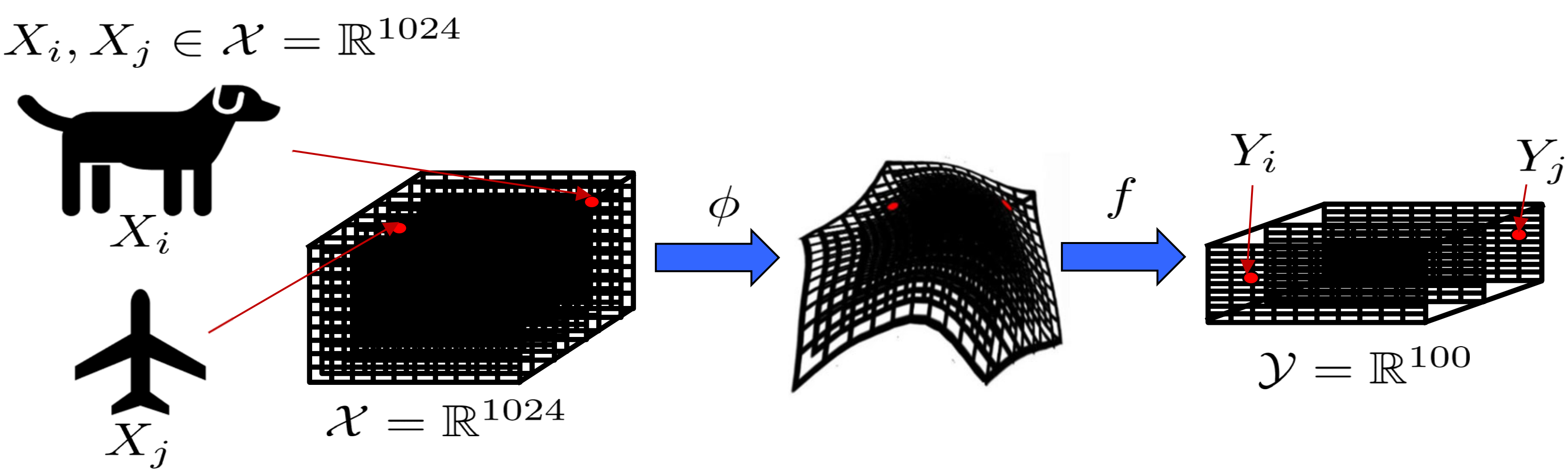
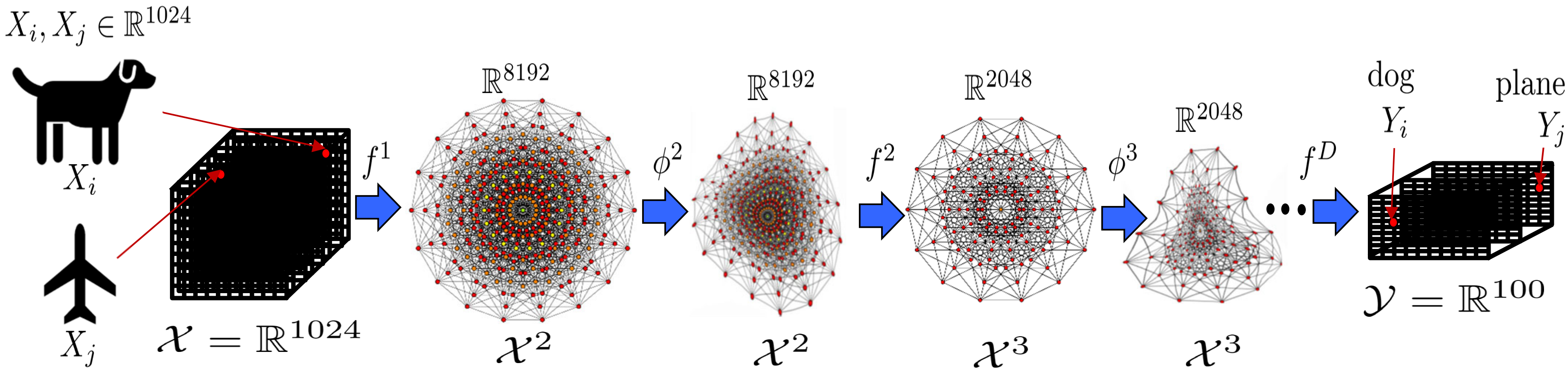


	Image registration	Idea registration
	Image $I : [0,1]^2 \to \mathbb{R}$ $I' : [0,1]^2 \to \mathbb{R}$	Idea/abstraction $I : \mathcal{X} \to \mathcal{Y}$ $I' : \mathcal{Y} \to \mathcal{Y}$
X_i, Y_i	Landmark/material points $X_i \in [0,1]^2$, $Y_i \in [0,1]^2$	Data points $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$
ϕ	Deforms $[0,1]^2$ and $I : [0,1]^2 \to \mathbb{R}$	Deforms $\mathcal{X}$ and $I : \mathcal{X} \to \mathcal{Y}$



Composed idea registration



Composed idea registration blocks $\rightarrow$ idea *form*ation

ANNs and ResNets are solvers for discretized idea *form*ation problems!

CNNs are solvers for discretized idea *form*ation problems
defined with a particular choice of kernels for Γ and K! (REM kernels)

Composed mechanical regression blocks $\rightarrow$ ANNs and their generalization

Idea registration

Approximate $f^\dagger$ with

$$f \circ \phi^v(x)$$

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

$v : \mathcal{X} \times [0,1] \to \mathcal{X}$ and $f : \mathcal{X} \to \mathcal{Y}$ are minimizers of

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

Bayesian interpretation

Theorem

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

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

$\xi \sim \mathcal{N}(0,K)$

$\phi^{\zeta}(x,t)$: solution of

$$\begin{cases} \dot{z} & = \zeta(z,t) \\ z(0) & = x \end{cases}$$

ζ centered GP defined by norm $\int_0^1 \|v(\cdot,t)\|_{\Gamma}^2\,dt$ (independent from ξ)

$Z = (Z_1,\ldots,Z_N)$: centered random Gaussian vector,
independent from ζ and ξ, with i.i.d. $\mathcal{N}(0, I_{\mathcal{Y}})$ entries

Bayesian interpretation

ζ centered GP defined by norm $\int_0^1 \|v(\cdot,t)\|_\Gamma^2 \, dt$

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

$$\Gamma = \psi^T \psi$$

e_i: orthonormal basis of $\mathcal{F}$

$$\psi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{F})$$

Deep residual Gaussian process

$\phi^\zeta(x,t)$: solution of

$$\begin{cases} \dot{z} & = \zeta(z,t) \\ z(0) & = x \end{cases}$$

Related:

[Baxendale, 1984]: Brownian motion in the diffeomorphism group

[Kunita, 1997]: Stochastic flows.

[Damianou and Lawrence, 2013]: Deep gaussian processes.

Idea registration is ridge regression with a warped kernel

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

$$f^{\mathrm{IR}} = f \circ \phi^v(x)$$

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

(RR) $\min_f \lambda \|f\|_{K^v}^2 + \|f(X) - Y\|_{\mathcal{Y}^N}^2$ $K^v(x,x') = K\big(\phi^v(x,1), \phi^v(x',1)\big)$

$$f^{RR} = f$$

Theorem $f^{\mathrm{IR}} = f^{\mathrm{RR}}$

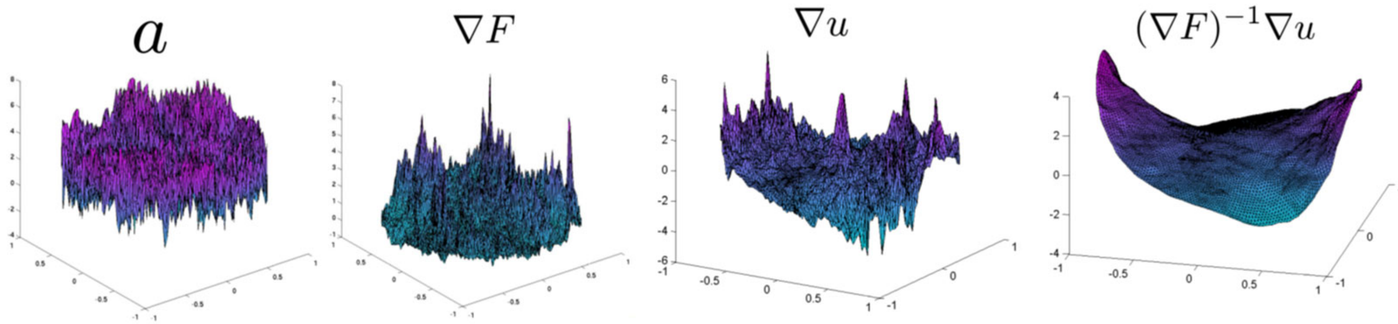
Warping kernels

Spatial statistics
[Sampson, Guttorp, 1992], [Perrin, Monestiez, 1999], [Schmidt, O'Hagan, 2003]
Enable the nonparameteric estimation of nonstationary and anisotropic spatial covariance structures

Numerical homogenization: [O., Zhang, 2005]

$$\begin{cases} -\operatorname{div}(a\nabla u) = g, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases}$$

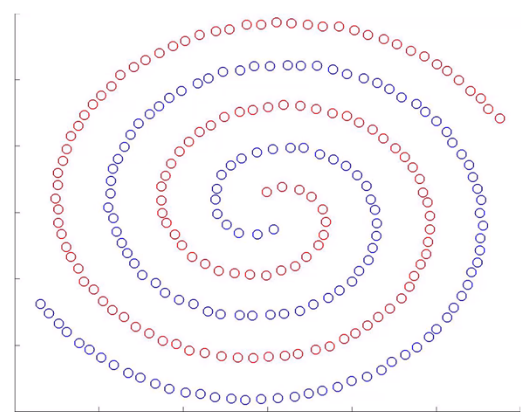
$$\begin{cases} -\operatorname{div}(a\nabla F_i) = 0 & \Omega \\ F_i(x) = x_i & \partial\Omega \end{cases}$$



Kernel Flows:
[O., Yoo, 2018], [Chen, O., Stuart, 2020], [Hamzi, O., 2020], [Yoo, O., 2020]

Kernel Flows learns a kernel of the form $K\big(\phi^v(x,1), \phi^v(x',1)\big)$ without backpropagration (via cross-validation)

back-propagation could be replaced by forward cross-validation in DL



Diffeomorphic learning: [Younes, 2019], [Rousseau, Fablet, 2018], [Zammit-Mangion et al, 2019]

Idea registration is ridge regression with a prior learned from data

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

$$f^{\mathrm{IR}} = f\circ\phi^v(x)$$

$$\begin{cases}\dot\phi(x,t) = v\big(\phi(x,t),t\big)\\ \phi(x,0) = x\end{cases}$$

(RR) $\min_f \lambda\|f\|_{K^v}^2 + \|f(X) - Y\|_{\mathcal{Y}^N}^2$ $\quad K^v(x,x') = K\big(\phi^v(x,1),\phi^v(x',1)\big)$

$$f^{RR} = f$$

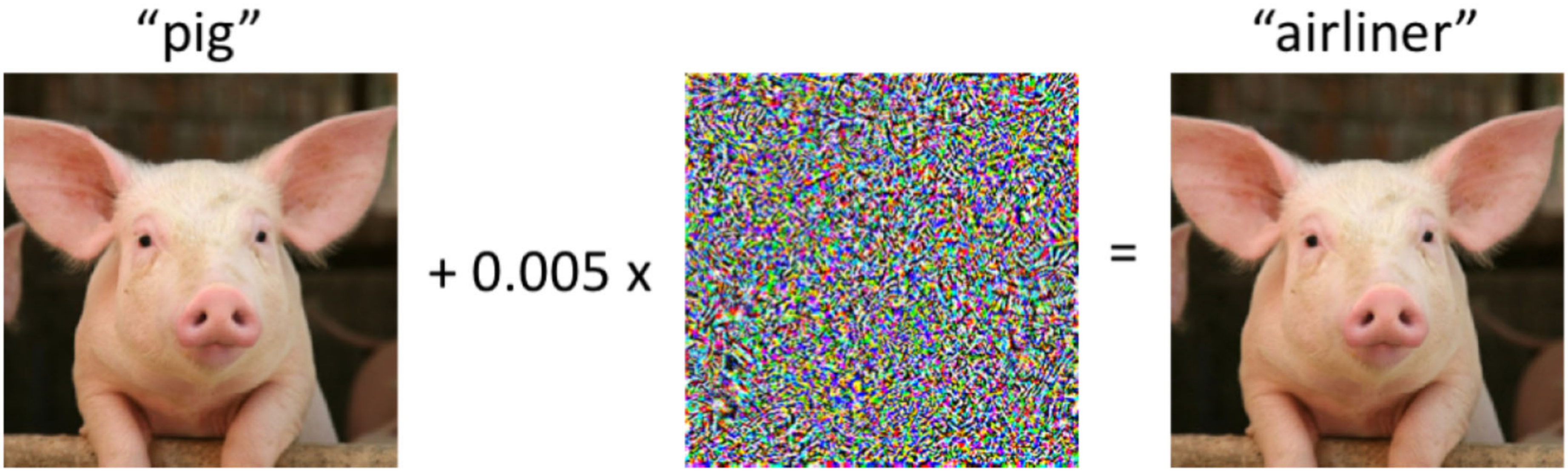
Theorem

$$f^{\mathrm{IR}} = f^{\mathrm{RR}}$$

$$f^{\mathrm{IR}}(x) = \mathbb{E}_{\xi\sim\mathcal{N}(0,K^v)}\big[\xi(x)\mid\xi(X)=Y\big]$$

[Biggio et al, 2012-2018], [Moisejevs et al, 2019]:
ANNs are brittle to data poisining

[Szegedy et al, Dec 2013]: ANNs are brittle to adversarial noise



[Madry, Schmidt, 2018]

Why?

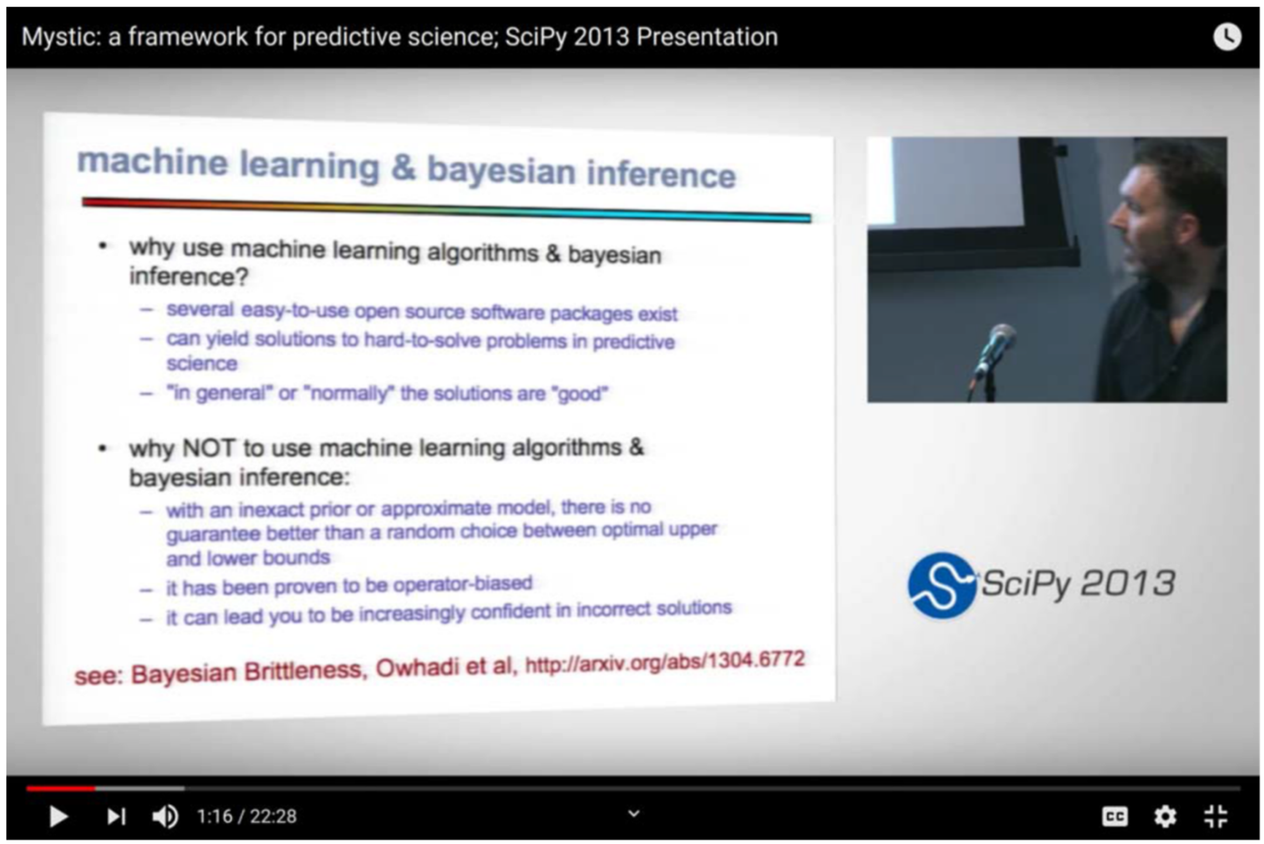
$$f^{\mathrm{IR}}(x) = \mathbb{E}_{\xi \sim \mathcal{N}(0,K^{v})}\left[\xi(x) \mid \xi(X) = Y\right]$$

[O., Scovel, Sullivan, Apr 2013]: Bayesian inference is brittle w.r. to perturbations of the prior

[McKerns, SyiPy, June 2013]: Bayesian brittleness can lead machine learning algorithms to be increasingly confident in incorrect solutions

https://youtu.be/o-nwSnLC6DU?t=74

Brittleness of Bayesian inference implies the brittleness of ANNs



Other causes?

$$f^{\mathrm{IR}}(x) = \mathbb{E}_{\xi\sim\mathcal{N}(0,K)}\left[\xi(\phi^v(x,1)) \mid \xi(\phi^v(X,1)) = Y\right]$$

Hamiltonian Chaos ⇒ Brittleness

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

Instability is inherent to Deep Learning

[Antun, Renna, Poon, Adcock, Hansen, 2020]

Can we fix it?

Not without giving up some accuracy because accuracy and robustness are conflicting requirements ([O., Scovel, 2017, qualitative robustness of Bayesian inference])

How do we fix it?

$$f^{\mathrm{IR}} = f \circ \phi^v(x)$$

Training without regularization

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

$$\begin{cases} \dot\phi(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

Training with regularization

$$\Gamma \Longleftrightarrow \Gamma + \underbrace{rI}$$

nugget

$$K \Longleftrightarrow K + \overbrace{\rho I}$$

$$\min_{v,f,q,Y'} \frac{\nu}{2}\int_0^1 \|v(\cdot,t)\|_\Gamma^2\,dt + \frac{1}{r}\int_0^1 \|\dot q - v(q(t))\|_{\mathcal{X}^N}^2\,dt$$
$$+\lambda\|f\|_K^2 + \frac{\lambda}{\rho}\|f(q(1)) - Y'\|_{\mathcal{Y}^N}^2 + \|Y' - Y\|_{\mathcal{Y}^N}^2$$

$$q : [0,1] \to \mathcal{X}^N \qquad q(0) = X$$

Unregularized ANN

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

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

Regularized ANN

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

Theorem

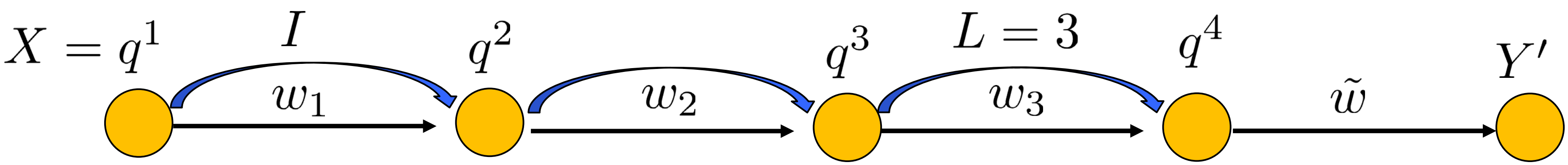
$f\circ\phi_L$ obtained from regularized ANN is continuous in x, X, Y

➡ Provides a principled alternative to Dropout

One ResNet block with and without regularization

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

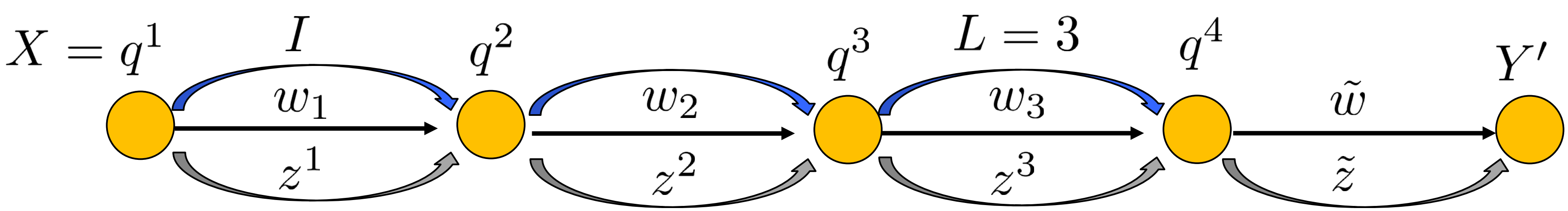
$$f \circ \phi_L(x) = (\tilde{w}\boldsymbol{\varphi}) \circ (I + w_L \boldsymbol{\varphi}) \circ \cdots \circ (I + w_1 \boldsymbol{\varphi})$$



$q^1 + w_1 \boldsymbol{\varphi}(q^1) = q^2$ $\quad q^2 + w_2 \boldsymbol{\varphi}(q^2) = q^3$ $\quad q^3 + w_3 \boldsymbol{\varphi}(q^3) = q^4$ $\quad \tilde{w}\boldsymbol{\varphi}(q^4) = Y'$

Without regularization

$$\min_{w^s, \tilde{w}} \frac{\nu L}{2} \sum_{s=1}^{L} \|w^s\|^2_{\mathcal{L}(\mathcal{X}\oplus\mathbb{R},\mathcal{X})} + \lambda \|\tilde{w}\|^2_{\mathcal{L}(\mathcal{X}\oplus\mathbb{R},\mathcal{Y})} + \|Y' - Y\|^2_{\mathcal{Y}^N}$$



With regularization

$z^2 + q^2 + w_2 \boldsymbol{\varphi}(q^2) = q^3$ $\quad \tilde{z} + \tilde{w}\boldsymbol{\varphi}(q^4) = Y'$

$z^1 + q^1 + w_1 \boldsymbol{\varphi}(q^1) = q^2$ $\quad z^3 + q^3 + w_3 \boldsymbol{\varphi}(q^3) = q^4$

$$\min_{w^s, \tilde{w}, z^s, \tilde{z}} \frac{\nu L}{2} \sum_{s=1}^{L} \left(\|w^s\|^2_{\mathcal{L}(\mathcal{X}\oplus\mathbb{R},\mathcal{X})} + \frac{1}{r}\|z^s\|^2_{\mathcal{X}^N}\right) + \lambda \left(\|\tilde{w}\|^2_{\mathcal{L}(\mathcal{X}\oplus\mathbb{R},\mathcal{Y})} + \frac{1}{\rho}\|\tilde{z}\|^2_{\mathcal{Y}^N}\right) + \|Y' - Y\|^2_{\mathcal{Y}^N}$$

What are the minimizers of mechanical regression or idea registration variational problems?

Mechanical regression

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

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

Idea registration

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

$$\begin{cases} \dot\phi(x,t) = v(\phi(x,t),t) \\ \phi(x,0) = x \end{cases}$$

Mechanical regression

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

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

Theorem

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

$q^s \in \mathcal{X}^N$

$\Gamma(q^s, q^s)$: $N \times N$ block matrix with blocks $\Gamma(q_i^s, q_j^s)$

$\Gamma(\cdot, q^s)$: $1 \times N$ block matrix with blocks $\Gamma(\cdot, q_i^s)$

$q^1 = X$, $q^2, \ldots, q^{L+1}$ minimizers of

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

Discrete least action principle $\qquad \Delta t = \frac{1}{L}$

Corollary Introducing momentum variables

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

(q^s, p^s) follows Hamiltonian dynamic

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

$$q^1 = X$$

$v_1, \ldots, v_L, f$ uniquely determined by p^1

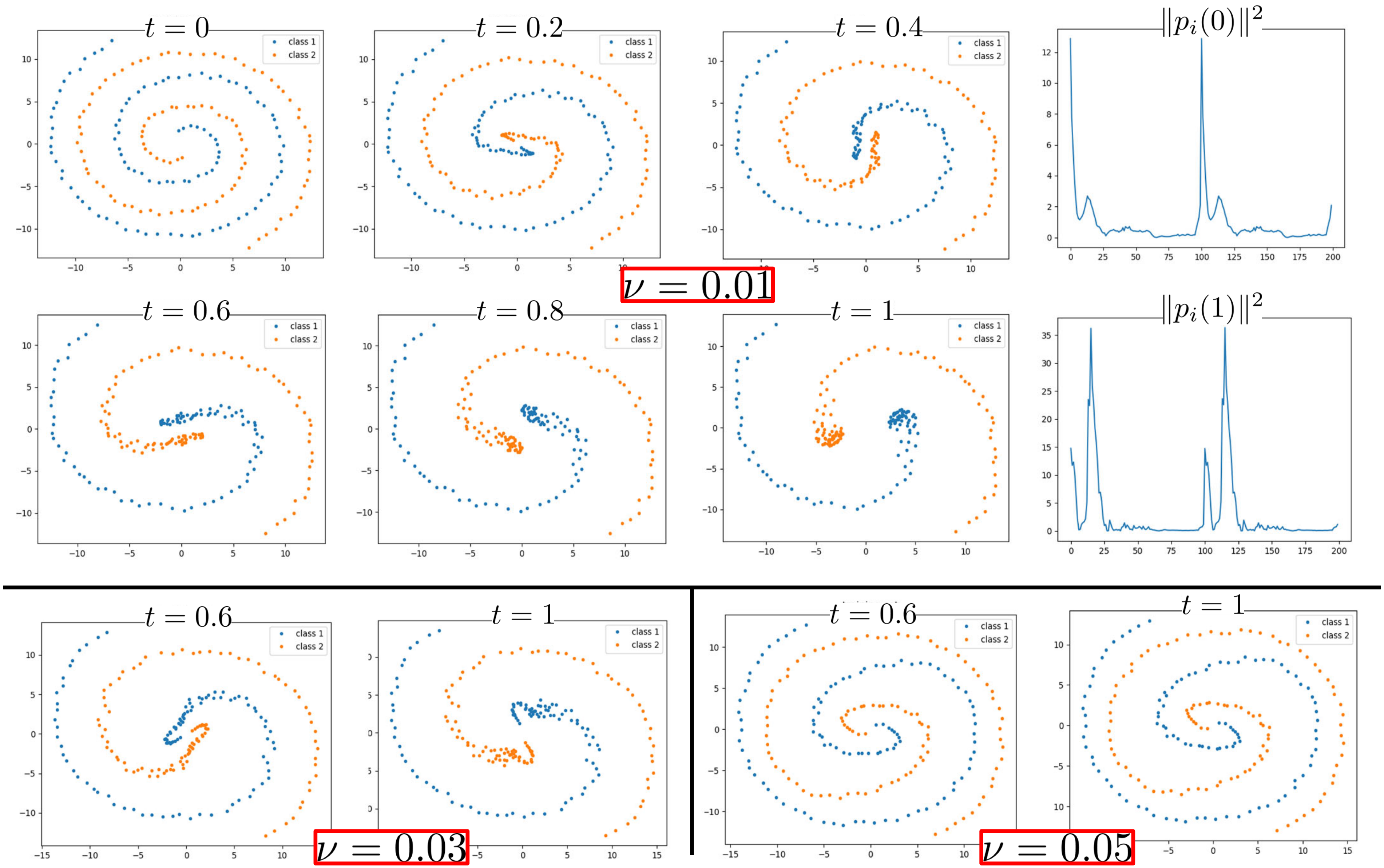
$w_1, \ldots, w_L, \tilde{w}$ uniquely determined by p^1

Weights and biases of ANN determined by initial momentum p^1

Geodesic shooting: [Allassonière, Trouvé, Younes, 2005], [Vialard et Al, 2020]

As in image registration: [Bruveris et Al 2011], [Vialard, 2012]

The momentum representation of the regressor is sparse



Corollary Near energy preservation

↓

The norms $\|v_s\|_\Gamma^2$ and $\|w_s\|^2_{\mathcal{L}(\mathcal{X}\oplus\mathbb{R},\mathcal{X})}$ fluctuate by at most $\mathcal{O}(1/L)$

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

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

Idea registration

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

$$\begin{cases} \dot{\phi}(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

Theorem

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

q position variable in $\mathcal{X}^N$ started from $q(0) = X$, minimizing the least action principle

$$\min_{f,q} \frac{\nu}{2} \int_0^1 \dot{q}^T \Gamma(q,q)^{-1} \dot{q} + \lambda \|f\|_K^2 + \|f(q(1)) - Y\|_{\mathcal{Y}^N}^2$$

Idea registration

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

$$\begin{cases} \dot\phi(x,t) = v\big(\phi(x,t),t\big) \\ \phi(x,0) = x \end{cases}$$

Corollary

$$v(x,t) = \Gamma(x,q)p \qquad p = \Gamma(q,q)^{-1}\dot q$$

(q,p) position and momentum variables in $\mathcal{X}^N$ started from $q(0) = X$

$$\begin{cases} \dot q_i &= \partial_{p_i}\mathfrak{H}(q,p) \\ \dot p_i &= -\partial_{q_i}\mathfrak{H}(q,p) \end{cases} \qquad \mathfrak{H}(q,p) = \frac{1}{2} p^T \Gamma(q,q) p$$

v, f uniquely determined by $p(0)$

$\|v(\cdot,t)\|_\Gamma^2$ constant over $t \in [0,1]$

Mean field limit $\Gamma(x,x') = \psi^T(x)\psi(x')$

Rescale momentum variables $p_j = \frac{1}{N}\bar{p}_j$

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

$$v(x,t) = \psi^T(x)\,\alpha(t)$$

Theorem

If $\mu_N := \frac{1}{N}\sum_{i=1}\delta_{(q_i,\bar{p}_i)}$ converges (weakly) then its limit is

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

Ensemble analysis of gradient descent

[Mei et al, 2018]

[Rotsko, Vanden-Eijnden, 2018]

Existence, uniqueness and convergence of minimizers

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

(IR) $$\min_{v,f} \frac{\nu}{2}\int_0^1 \|v(\cdot,t)\|_\Gamma^2\,dt + \lambda\|f\|_K^2 + \|f\circ\phi^v(X,1) - Y\|_{\mathcal{Y}^N}^2$$ $$\begin{cases}\dot\phi(x,t) = v(\phi(x,t),t)\\ \phi(x,0) = x\end{cases}$$

Theorem

Minimizers of (MR) and (IR) exist

Minimizers of (MR) and (IR) are unique given initial momentum (p^1 or $p(0)$)

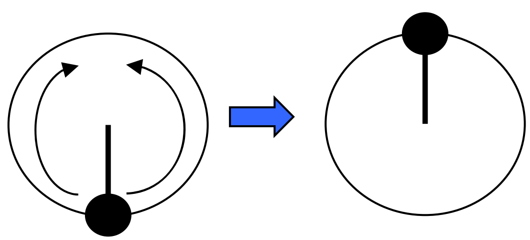
Minimal value of (MR) converges (as $L\to\infty$) to the minimal value of (IR)

Adherence values of ϕ_L minimizing (MR) are the ϕ^v minimizing (IR)

Remark

Minimizers of (MR) and (IR) are (for pathological examples) non unique

[Marsden, Ratiu, 2013]: Conjugate points in mechanics



Deterministic Error estimates

Theorem

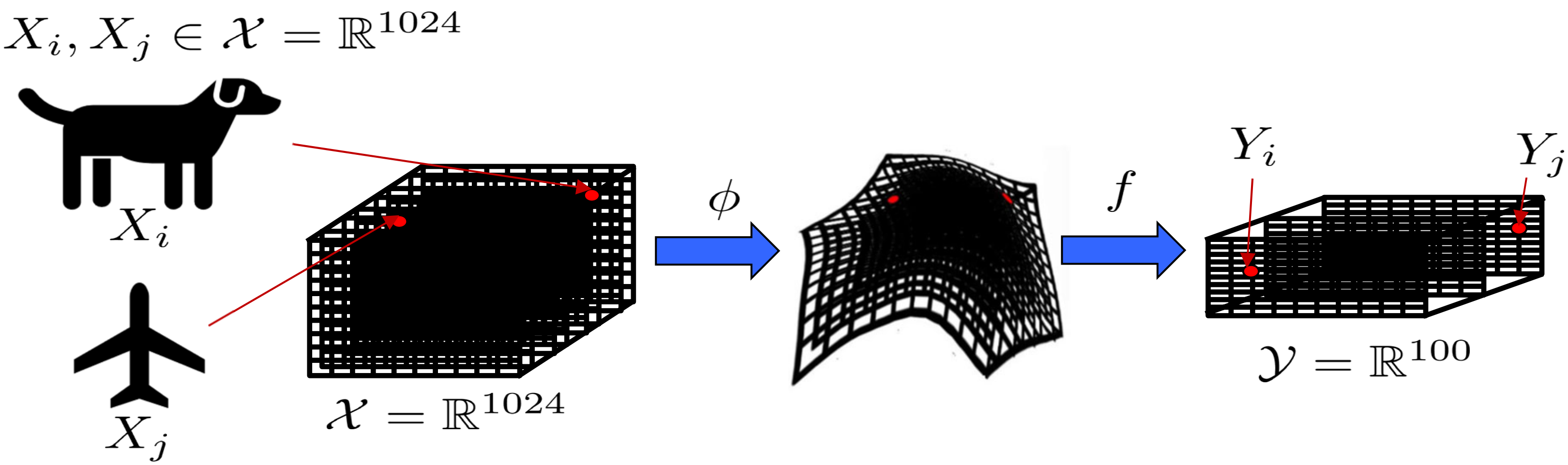
$$\left\|f^\dagger(x) - f \circ \phi^v(x,1)\right\|_{\mathcal{Y}} \leq \sigma(x) \|f^\dagger\|_{K^v}$$

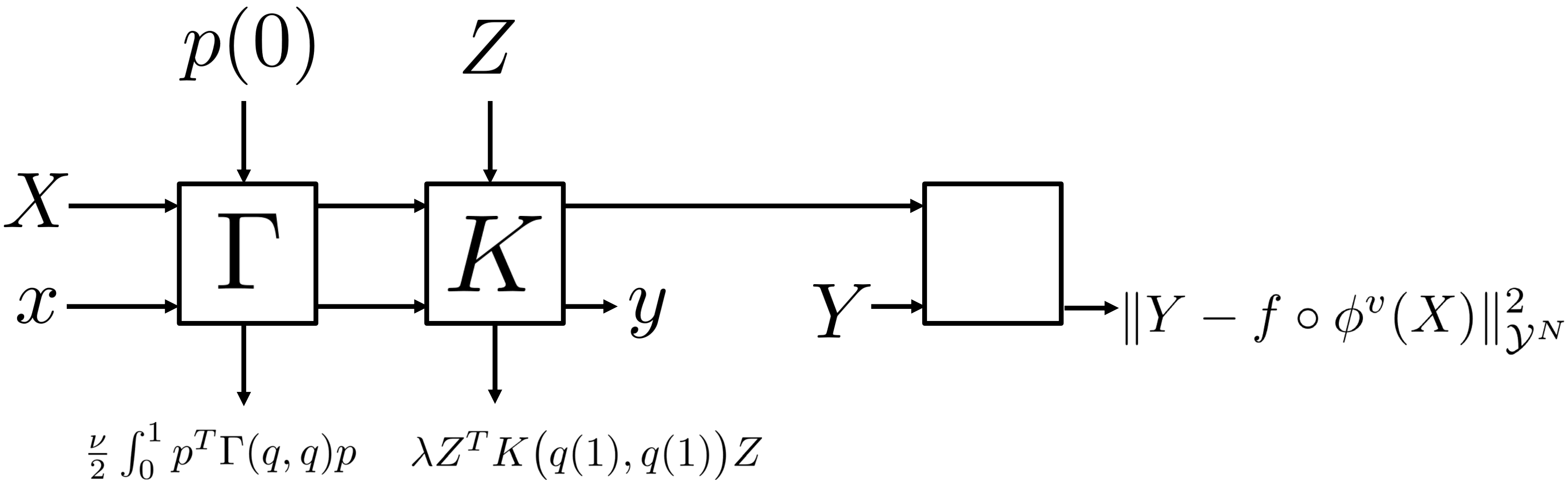
$$\sigma^2(x) := \text{Trace}\left[K^v(x,x) - K^v(x,X)\big(K^v(X,X) + \lambda I_{\mathcal{Y}}\big)^{-1} K^v(X,x)\right]$$

$$\boxed{K^v(x,x') = K\big(\phi^v(x,1), \phi^v(x',1)\big)}$$

Does not depend on dimension!
But need to bound $\|f^\dagger\|_{K^v}$ to be useful

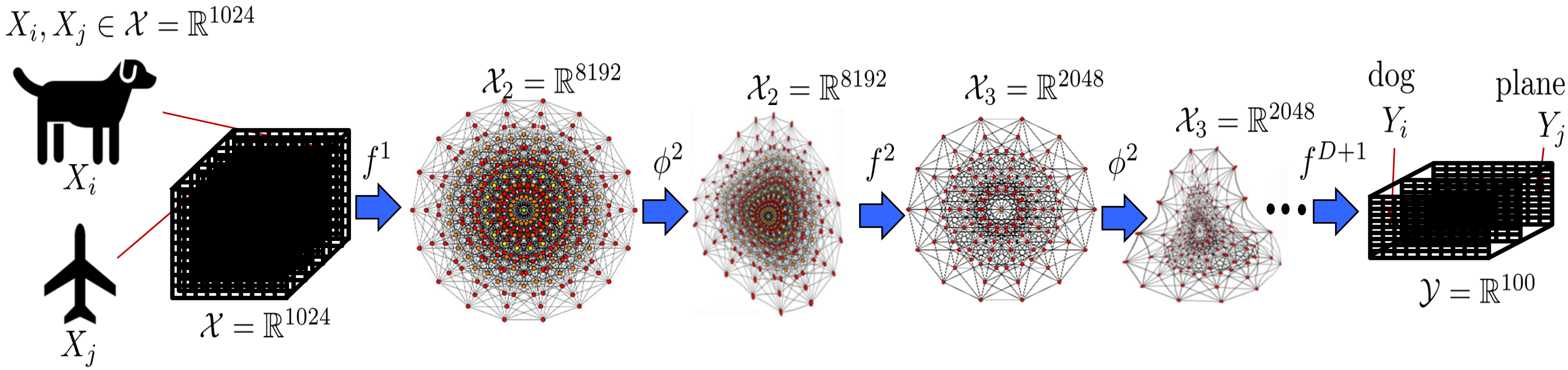
One mechanical regression / idea registration block





Total loss= $\frac{\nu}{2}\int_0^1 p^T \Gamma(q,q)p + \lambda Z^T K\big(q(1), q(1)\big) Z + \|Y - f \circ \phi^v(X)\|^2_{\mathcal{Y}^N}$

Composing mechanical regression / idea registration blocks



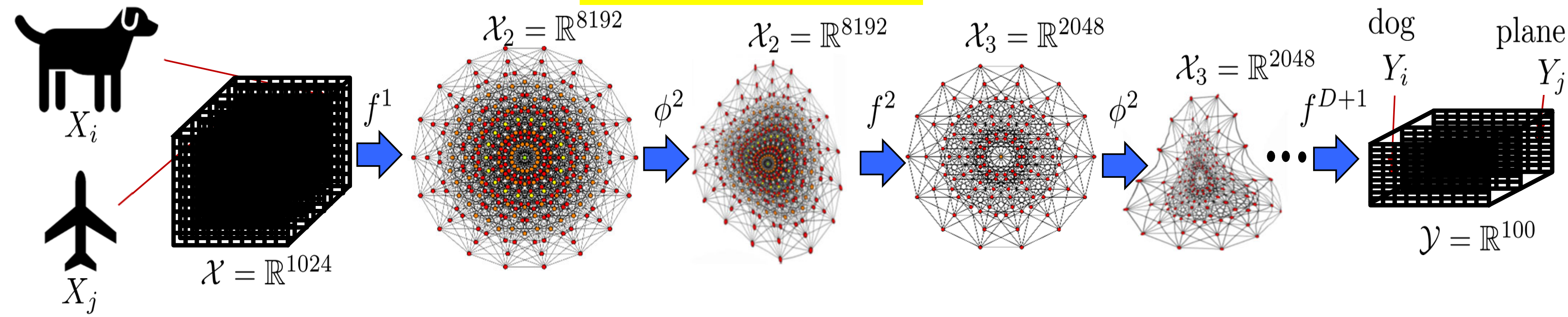
Composed mechanical regression blocks $\rightarrow$ ANNs and ResNets

Composed idea registration blocks $\rightarrow$ idea *form*ation

ANNs and ResNets are solvers for discretized idea *form*ation problems!

CNNs are solvers for discretized idea *form*ation problems defined with REM kernels!

Idea *formation*



Theorem

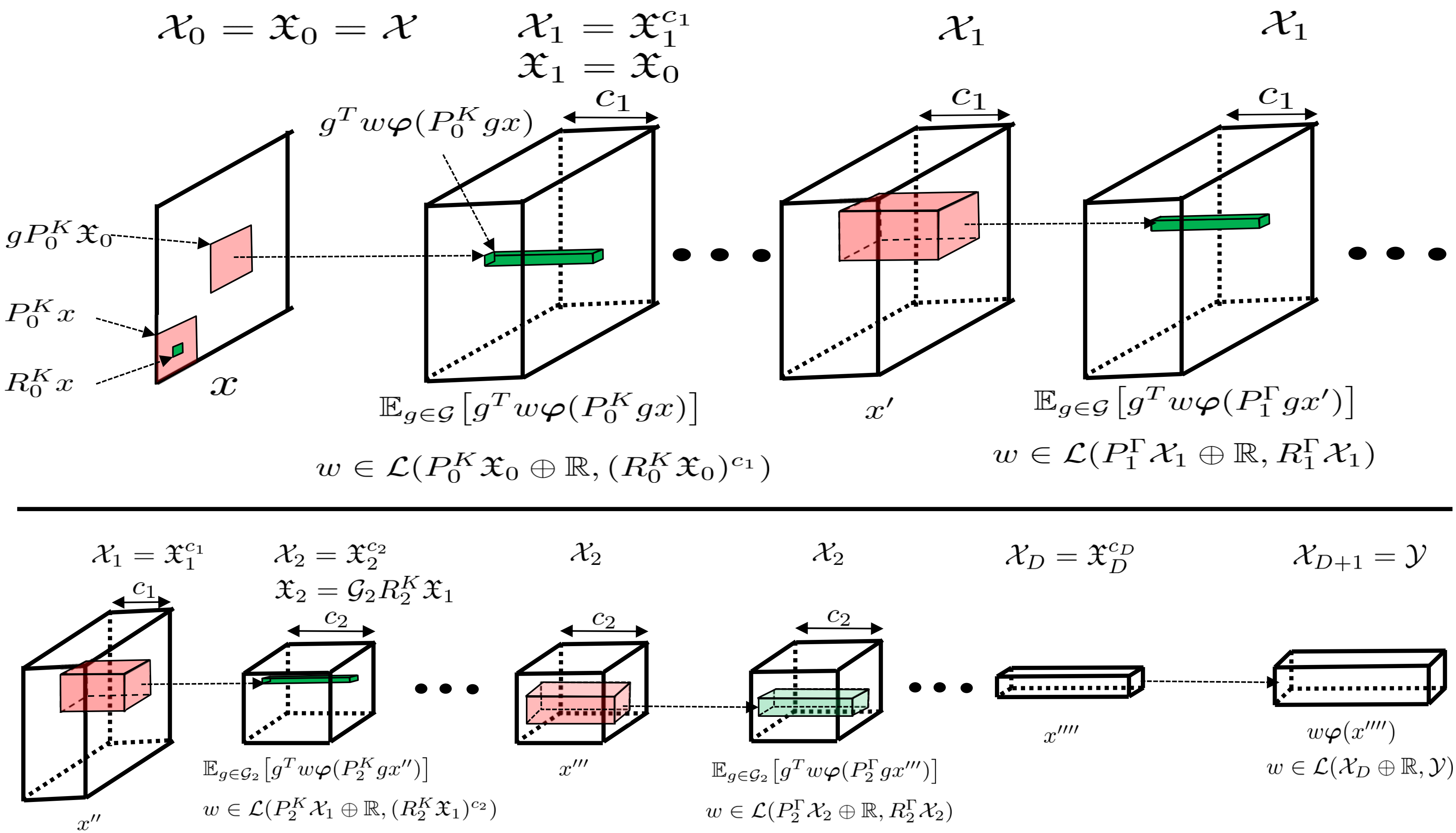
L^2 regularized ANNs/ResNets/CNNs have minimizers

Uniquely determined by initial momentum (weights and biases of first layer)

Norms of weights and biases of ResNet blocks are nearly preserved

ResNets converge to nested idea formation
(in the sense of adherence values as depth of ResNet blocks goes to infinity)

CNN/ResNet are discretized idea formation solvers with REM kernels



Equivalent kernels [Reisert, Burkhardt, 2007]

$\mathcal{X}$: Hilbert space

$\mathcal{G}$: Group of linear unitary transformations on $\mathcal{X}$

Definition

An operator-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{X})$ is **$\mathcal{G}$-equivariant** if

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

Similarly a function $f : \mathcal{X} \to \mathcal{X}$ is $\mathcal{G}$-equivariant if

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

$\mathbb{E}_g$: Expectation with respect to Haar measure on $\mathcal{G}$

Proposition

Given a (possibly non-equivariant) kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{X})$,

$$K^{\mathcal{G}}(x, x') := \mathbb{E}_{g,g'}\left[g^T K(gx, g'x')g'\right],$$

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

Theorem [Reisert, Burkhardt, 2007]

If K is scalar and $K(x, x') = K(gx, gx')$ then the minimizer of

$$\begin{cases} \text{Minimize} & \|f\|_K \\ \text{subject to} & f(X) = Y \text{ and } f \text{ is } \mathcal{G} - \text{equivarient} \end{cases}$$

is $f^{\mathcal{G}}(\cdot) := K^{\mathcal{G}}(\cdot, X)K^{\mathcal{G}}(X, X)^{-1}Y$.

Reduced Equivariant Multichannel (REM) kernels [O., 2020]

$\mathfrak{X}$: Hilbert space

Linear projections

$$P : \mathfrak{X} \to \mathfrak{X}$$
$$R : \mathfrak{X} \to \mathfrak{X}$$

$\mathcal{G}$: Group of linear unitary transformations on $\mathfrak{X}$

Extend $\mathcal{G}$, P, R to $\mathfrak{X}^c$

$$g(x_1, \ldots, x_c) = (gx_1, \ldots, gx_c)$$
$$P(x_1, \ldots, x_c) = (Px_1, \ldots, Px_c)$$

$c_1, c_2 \in \mathbb{N}$

$$K : P\mathfrak{X}^{c_1} \times P\mathfrak{X}^{c_1} \to \mathcal{L}(R\mathfrak{X}^{c_2})$$

$\mathbb{E}_g$: Expectation with respect to Haar measure on $\mathcal{G}$

REM kernel

$$K^{\mathrm{REM}} : \mathfrak{X}^{c_1} \times \mathfrak{X}^{c_1} \to \mathcal{L}(\mathfrak{X}^{c_2})$$

$$K^{\mathrm{REM}}(x, x') = \mathbb{E}_{g,g'}\left[g^T R K(Pgx, Pg'x')Rg'\right]$$

With activation functions

$$K(x,x') = \boldsymbol{\varphi}^T(x)\boldsymbol{\varphi}(x') I_{R\mathfrak{X}^{c_2}}$$

$$\boldsymbol{\varphi}(x) = \big(\mathbf{a}(x), 1\big) \qquad \boldsymbol{\varphi} : P\mathfrak{X}^{c_1} \to P\mathfrak{X}^{c_1} \oplus \mathbb{R}$$

$\mathbf{a}(x)$: Activation function $\qquad \mathbf{a} : P\mathfrak{X}^{c_1} \to P\mathfrak{X}^{c_1}$

$$\boxed{K^{\mathrm{REM}}(x,x') = \Psi^T(x)\Psi(x)}$$

$$\Psi\colon\ \mathfrak{X}^{c_1} \to \mathcal{L}(P\mathfrak{X}^{c_1} \oplus \mathbb{R}, R\mathfrak{X}^{c_2})$$

For $w \in \mathcal{L}(P\mathfrak{X}^{c_1} \oplus \mathbb{R}, R\mathfrak{X}^{c_2})$

$$\boxed{\Psi^T(x)w = \mathbb{E}_g\left[g^T(w\boldsymbol{\varphi}(Pgx))\right]}$$

With activation functions

$$K(x,x') = \boldsymbol{\varphi}^T(x)\boldsymbol{\varphi}(x')I_{R\mathfrak{X}^{c_2}}$$

$$\boldsymbol{\varphi}(x) = \big(\mathbf{a}(x), 1\big) \qquad \boldsymbol{\varphi} : P\mathfrak{X}^{c_1} \to P\mathfrak{X}^{c_1} \oplus \mathbb{R}$$

$\mathbf{a}(x)$: Activation function $\qquad \mathbf{a} : P\mathfrak{X}^{c_1} \to P\mathfrak{X}^{c_1}$

$$K^{\mathrm{REM}}(x,x') = \Psi^T(x)\Psi(x)$$

$$\Psi\colon\ \mathfrak{X}^{c_1} \to \mathcal{L}(P\mathfrak{X}^{c_1} \oplus \mathbb{R}, R\mathfrak{X}^{c_2})$$

For $w \in \mathcal{L}(P\mathfrak{X}^{c_1} \oplus \mathbb{R}, R\mathfrak{X}^{c_2})$

$$\Psi^T(x)w = \mathbb{E}_g\left[g^T(w\boldsymbol{\varphi}(Pgx))\right]$$

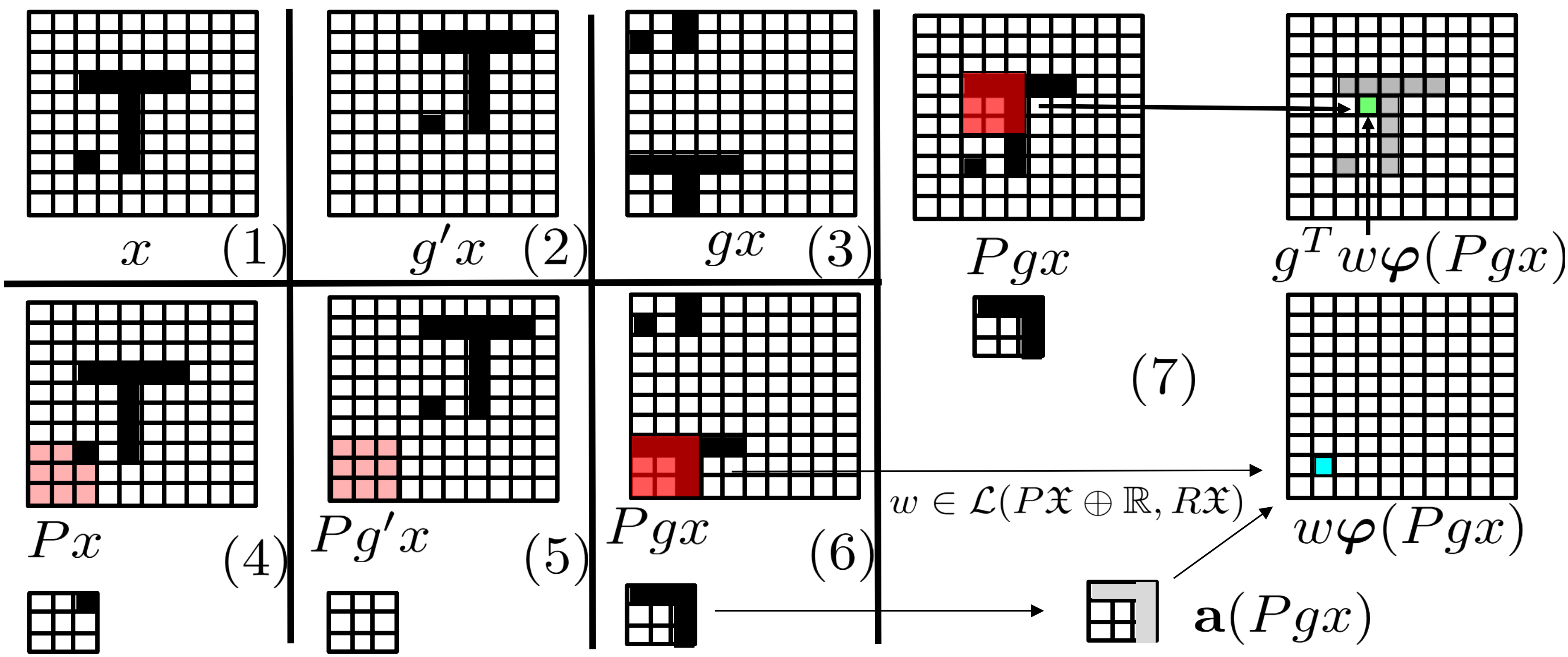
$\Psi^T w$ appears in the representation of any given layer of the regressor obtained from mechanical regression or from the composition of mechanical regression blocks

$$f \circ \phi_L(x) = (\Psi^T\tilde{w}) \circ (I + \Psi^T w_L) \circ \cdots \circ (I + \Psi^T w_1)$$

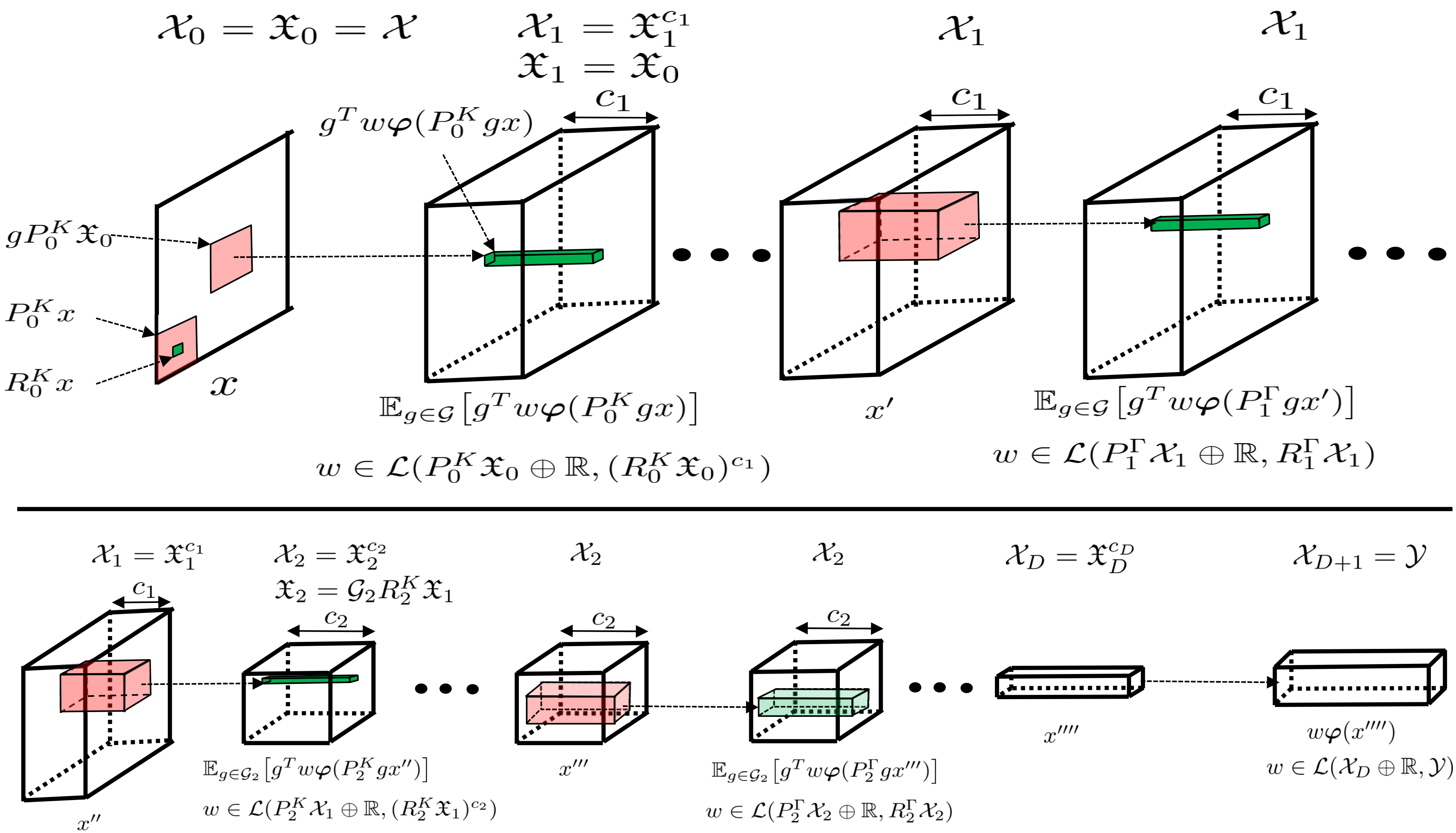
With activation functions

For $w \in \mathcal{L}(P\mathfrak{X}^{c_1} \oplus \mathbb{R}, R\mathfrak{X}^{c_2})$

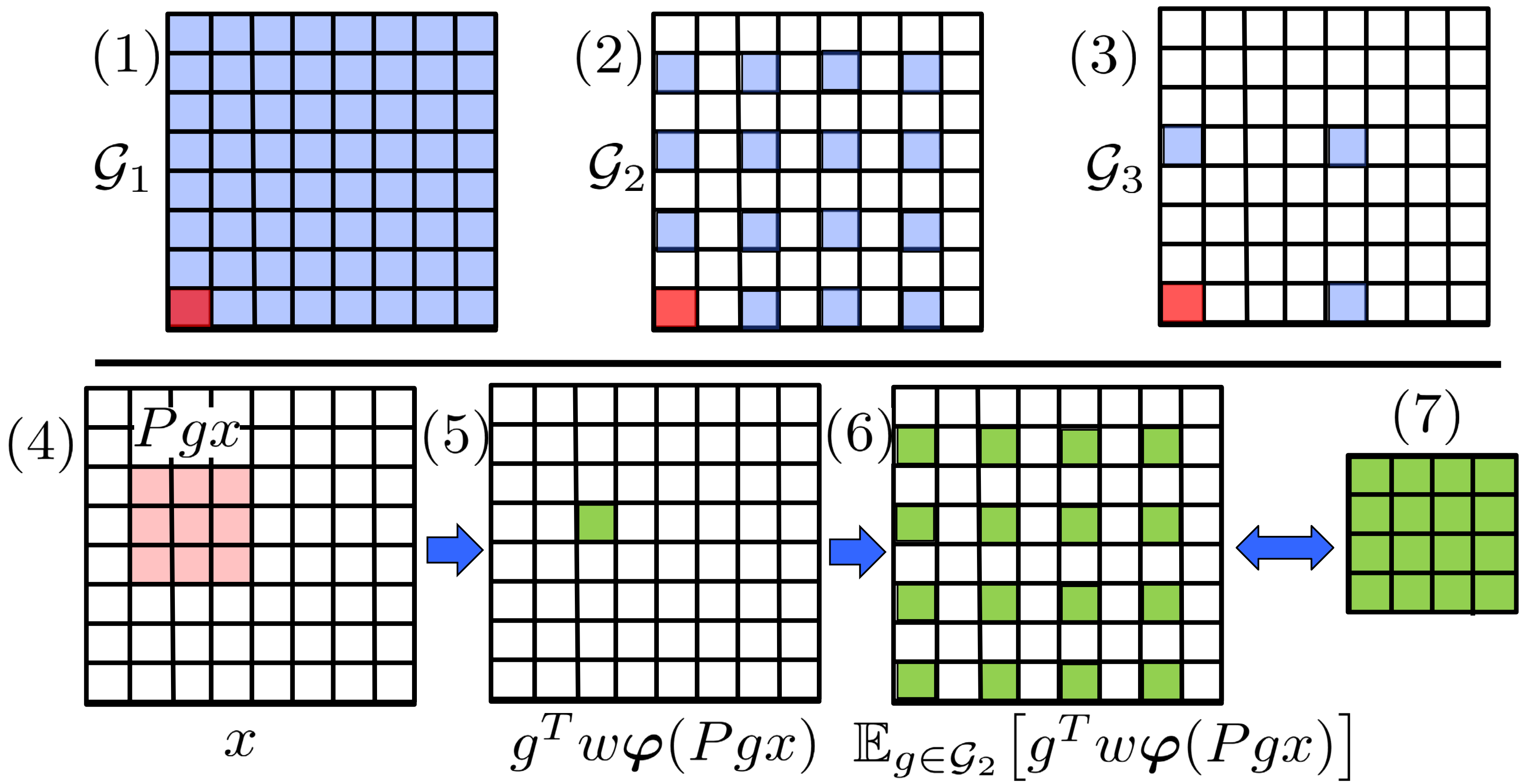
$$\Psi^T(x)w = \mathbb{E}_g\left[g^T(w\boldsymbol{\varphi}(Pgx))\right]$$



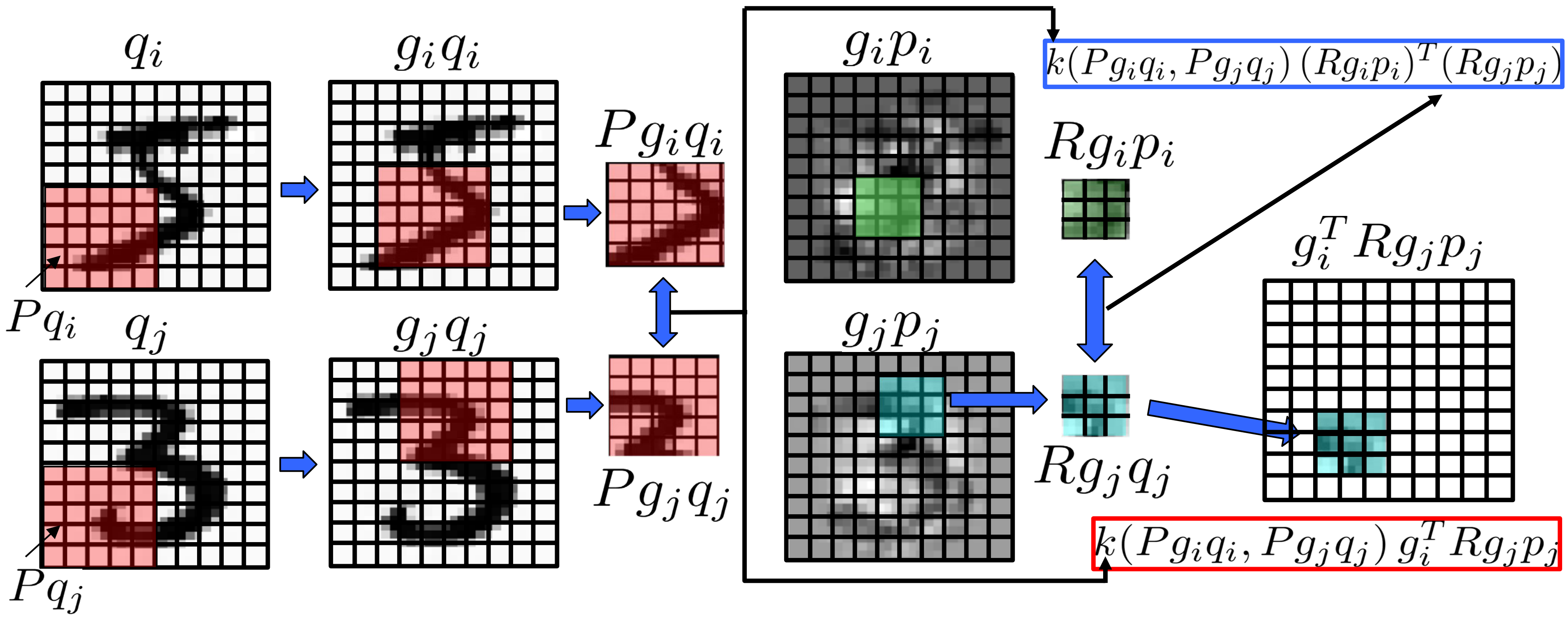
CNN/ResNet are discretized idea formation solvers with REM kernels



Pooling via striding is subgrouping



Hamiltonian Flow with REM kernels



Related work

- Deep kernel learning. [Wilson et al, 2016], [Bohn, Rieger, Griebel. 2019]
- Computational anatomy and image registration. [Joshi, Miller, 2000], [Micheli, 2008], [Beg, Miller, Trouvé, Younes, 2005], [Dupuis, Grenander, Miller, 1998], [Vialard, Risser, Rueckert, Cotter, 2012].
- Statistical numerical approximation. [O. 2015, 2017], [O., Scovel, 2019], [O., Scovel, Schäfer, 2019], [Raissi, Perdikaris, Karniadakis, 2019], [Cockayne, Oates, Sullivan, Girolami, 2019], [Hennig, Osborne, Girolami, 2015]
- ODE interpretations of ResNets. [E, 2017], [Haber, Ruthotto, 2017], [Chen, Rubanova, Bettencourt, Duvenaud, 2018], [Chang, Meng, Haber, Ruthotto, Begert, Holtham, 2018]
- Warping kernels [O., Zhang, 2005], [Sampson, Guttorp, 1992], [Perrin, Monestiez, 1999], [Schmidt, O'Hagan, 2003]
- Kernel Flows [O., Yoo, 2019], [Chen, O., Stuart, 2020], [Hamzi, O., 2020], [Yoo, O., 2020]
- Deep Gaussian processes. [Damianou, Lawrence, 2013]
- Brownian flow of diffeomorphisms [Kunita, 1997], [Baxendale., 1984]
- Equivariant kernels [Reisert, Burkhardt, 2007]
- Operator valued kernels [Kadri et al, 2016]
- Diffeomorphic learning: [Younes, 2019], [Rousseau, Fablet, 2018], [Zammit-Mangion et al, 2019]

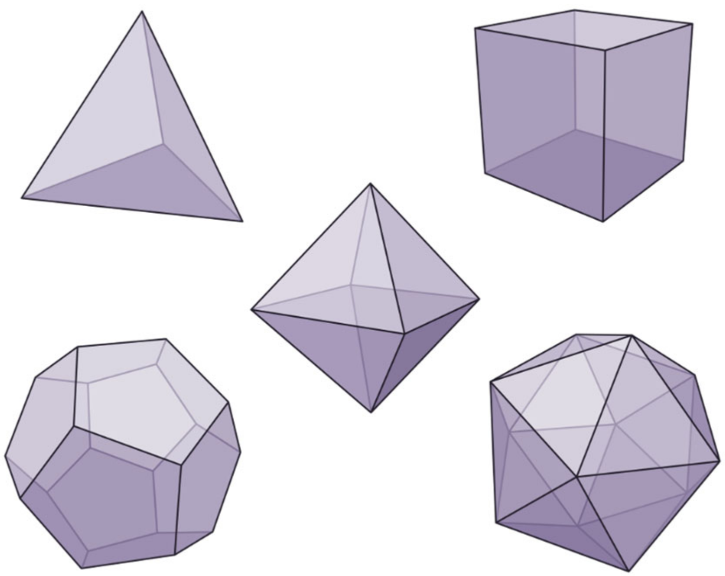
This work

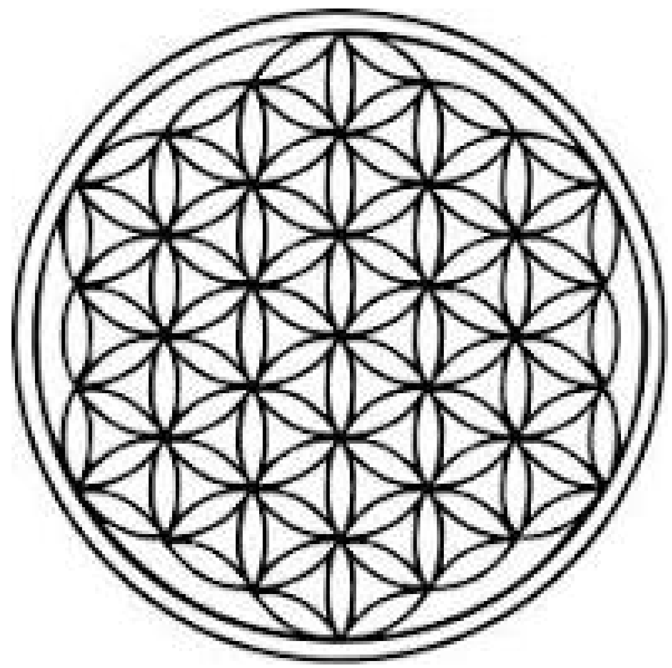
- Do ideas have shape? Plato's theory of forms as the continuous limit of artificial neural networks. [arXiv:2008.03920, O., 2020]

Do ideas have shape?

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"*

`https://www.etymonline.com/word/idea`

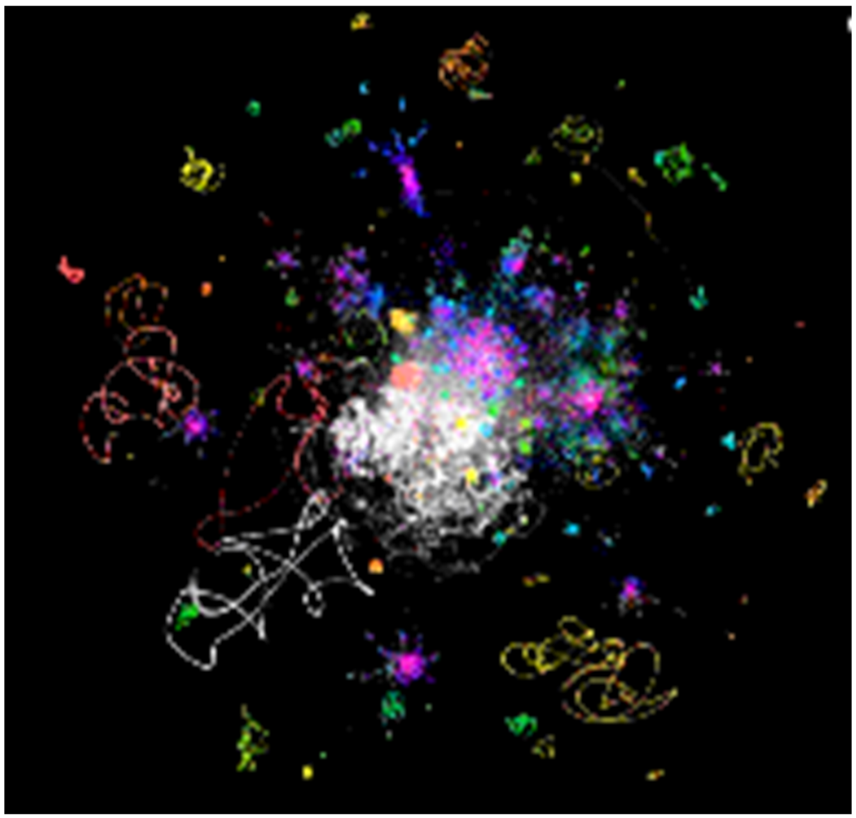




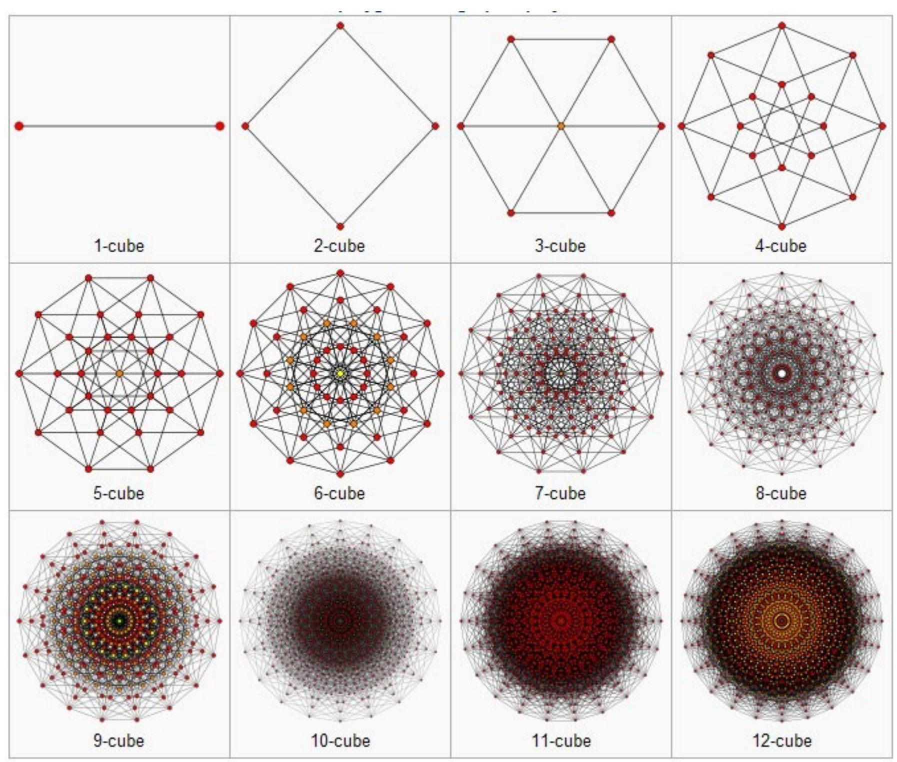
Conclusion

ANNs are 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.



https://johnhw.github.io/umap_primes/index.md.html



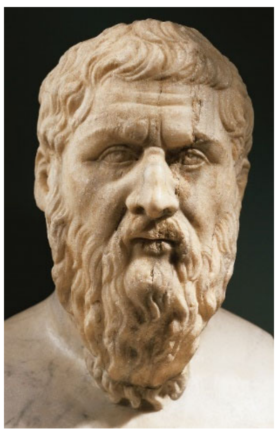
https://en.wikipedia.org/wiki/Hypercube

Do ideas have shape?

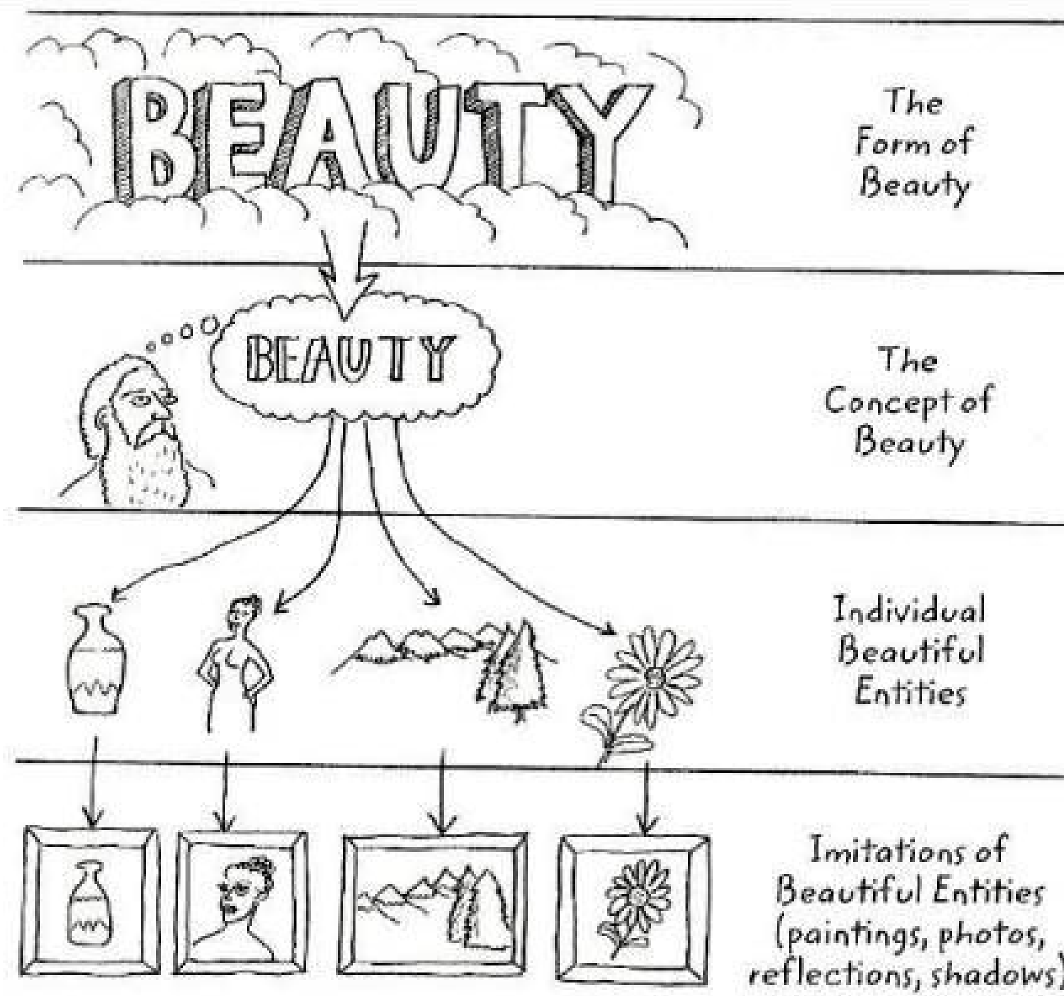
Idea: *"mental image or picture"... from Greek idea "form"... In Platonic philosophy, "an archetype, or pure immaterial pattern, of which the individual objects in any one natural class are but the imperfect copies"*

https://www.etymonline.com/word/idea

Plato's theory of forms



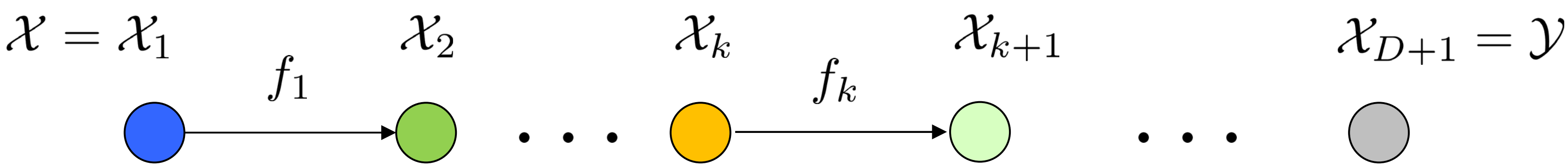
https://twitter.com/PhilosophyMttrs



Artificial neural network solution

Approximate $f^\dagger$ with

$$f = f_D \circ \cdots \circ f_1$$



$$f_k(x) = \mathbf{a}(W_k x + b_{k+1})$$

$\mathbf{a}$: Activation function / Elementwise nonlinearity

$\mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})$: Set of bounded linear operators from $\mathcal{X}_k$ to $\mathcal{X}_{k+1}$

$W_k \in \mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})$, $b_{k+1} \in \mathcal{X}_{k+1}$ identified as minimizers of

$$\min_{W_k, b_k} \quad \nu \sum_{k=1}^{D} \left(\|W_k\|^2_{\mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})} + \|b_{k+1}\|^2_{\mathcal{X}_{k+1}} \right) + \|f(X) - Y\|^2_{\mathcal{Y}^N}$$

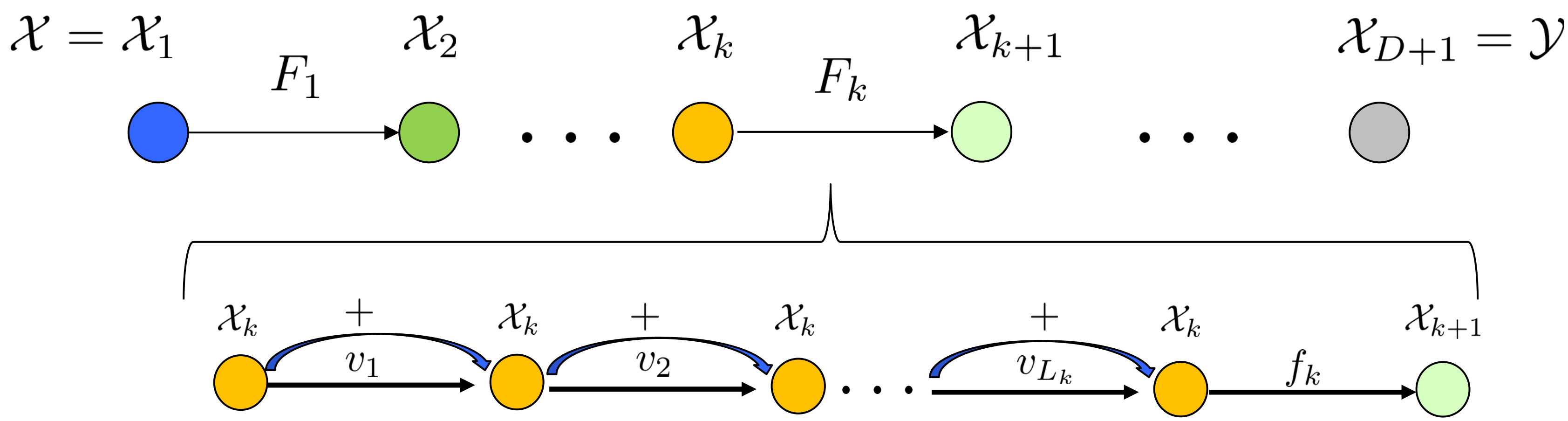
$$\|Y\|^2_{\mathcal{Y}^N} := \sum_{i=1}^{N} \|Y_i\|^2_{\mathcal{Y}}$$

Residual neural network solution

[He et Al, 2016]

Approximate $f^\dagger$ with

$$f = F_D \circ \cdots \circ F_1$$



$$F_k = f_k \circ (I + v^k_{L_k}) \circ \cdots \circ (I + v^k_1)$$

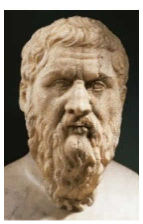
$$f_k \, : \, \mathcal{X}_k \to \mathcal{X}_{k+1} \qquad f_k(x) = \mathbf{a}(W_k x + b_{k+1})$$

$$v^k_s \, : \, \mathcal{X}_k \to \mathcal{X}_k \qquad v^s_k(x) = \mathbf{a}(W^s_k x + b^s_k)$$

$$\begin{aligned} \min_{W_k, b_k, W^s_k, b^s_k} \quad & \nu \sum_{k=1}^{D} \Big(\|W_k\|^2_{\mathcal{L}(\mathcal{X}_k, \mathcal{X}_{k+1})} + \|b_{k+1}\|^2_{\mathcal{X}_{k+1}} \\ & \qquad + \sum_{s=1}^{L_k} \|W^s_k\|^2_{\mathcal{L}(\mathcal{X}_k)} + \|b^s_k\|^2_{\mathcal{X}_k} \Big) \\ & + \|f(X) - Y\|^2_{\mathcal{Y}^N} \end{aligned}$$

Plato’s allegory of the cave

https://www.studiobinder.com/blog/platos-allegory-of-the-cave/



Plato

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 the world of abstractions or ideas