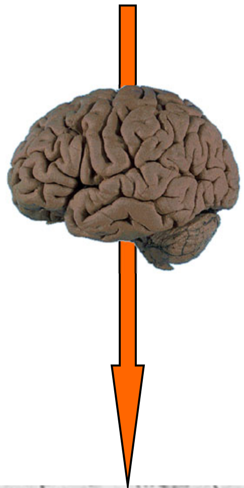
On learning adapted kernels for numerical approximation

Houman Owhadi

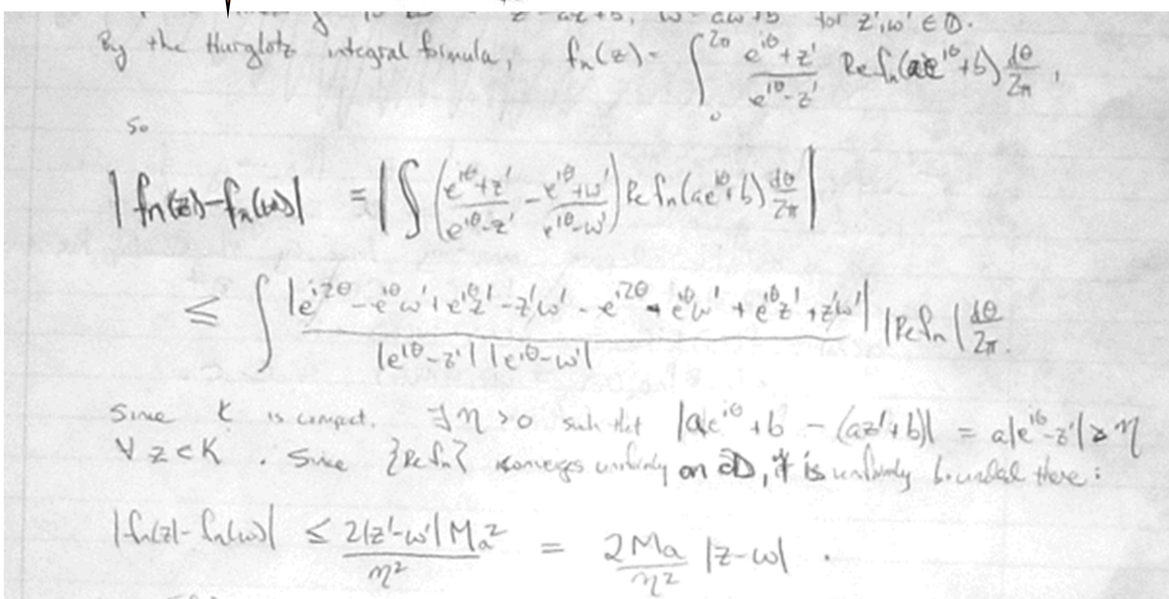
Solving PDEs: Two centuries ago

$$\Delta u = f$$



A. L. Cauchy (1789-1857)

S. D. Poisson (1781-1840)



By the Herglotz integral formula, $f_n(z) = \int_0^{2\pi} \frac{e^{i\theta}+z'}{e^{i\theta}-z'} \mathrm{Re} f_n(ae^{i\theta}+b) \frac{d\theta}{2\pi}$,

so

$$|f_n(z)-f_n(w)| = \left| \int \left(\frac{e^{i\theta}+z'}{e^{i\theta}-z'} - \frac{e^{i\theta}+w'}{e^{i\theta}-w'} \right) \mathrm{Re} f_n(ae^{i\theta}+b) \frac{d\theta}{2\pi} \right|$$

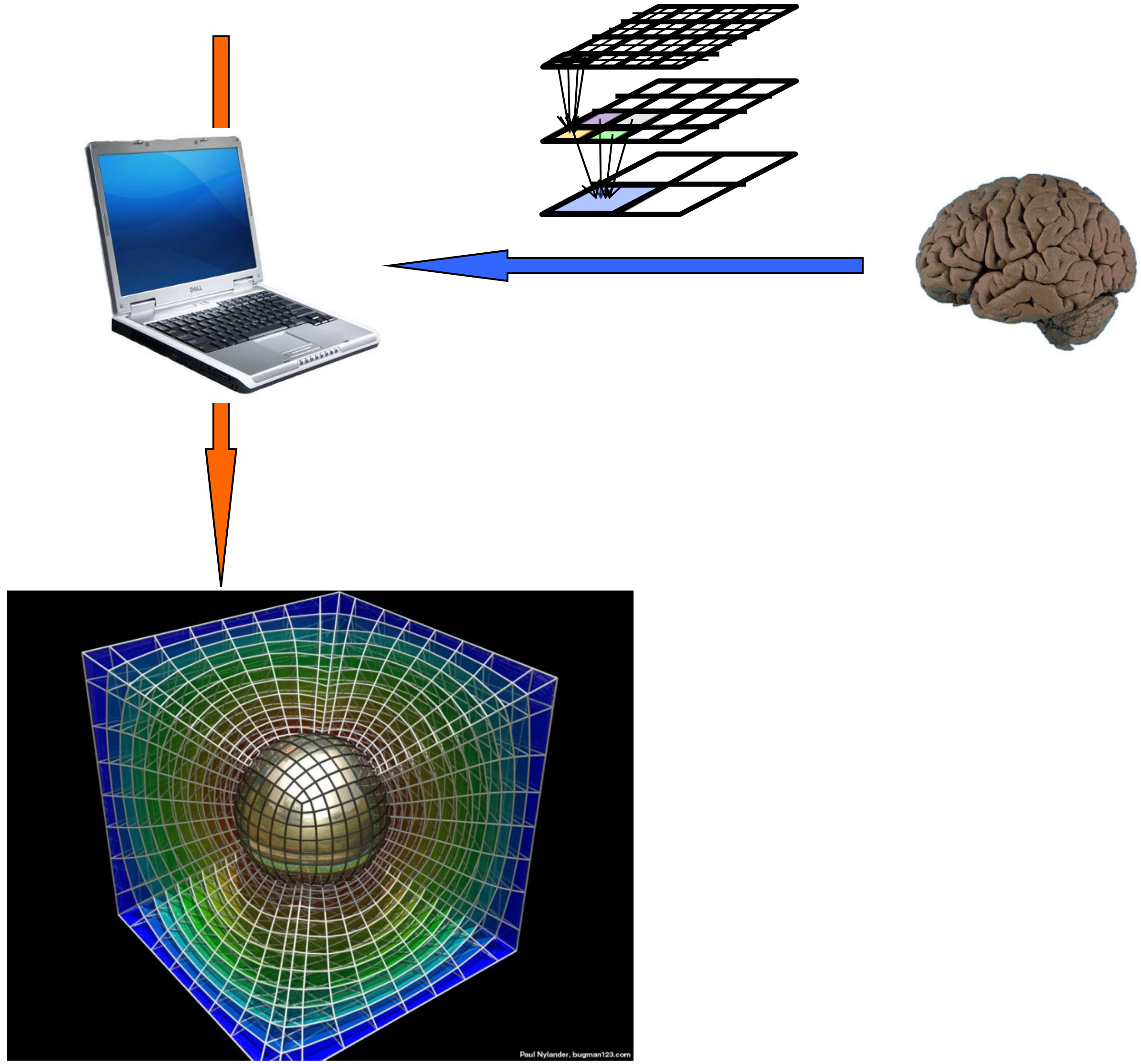
$$\le \int \frac{|e^{i2\theta} - e^{i\theta}w' + e^{i\theta}z' - z'w' - e^{i2\theta} + e^{i\theta}w' + e^{i\theta}z' + z'w'|}{|e^{i\theta}-z'|\,|e^{i\theta}-w'|} |\mathrm{Re} f_n| \frac{d\theta}{2\pi}.$$

Since K is compact, $\exists \eta > 0$ such that $|ae^{i\theta}+b - (az'+b)| = a|e^{i\theta}-z'| \ge \eta$ $\forall z \in K$. Since $\{\mathrm{Re} f_n\}$ converges uniformly on ∂D, it is uniformly bounded there:

$$|f_n(z)-f_n(w)| \le \frac{2|z'-w'|M_a^2}{\eta^2} = \frac{2M_a}{\eta^2}|z-w|.$$

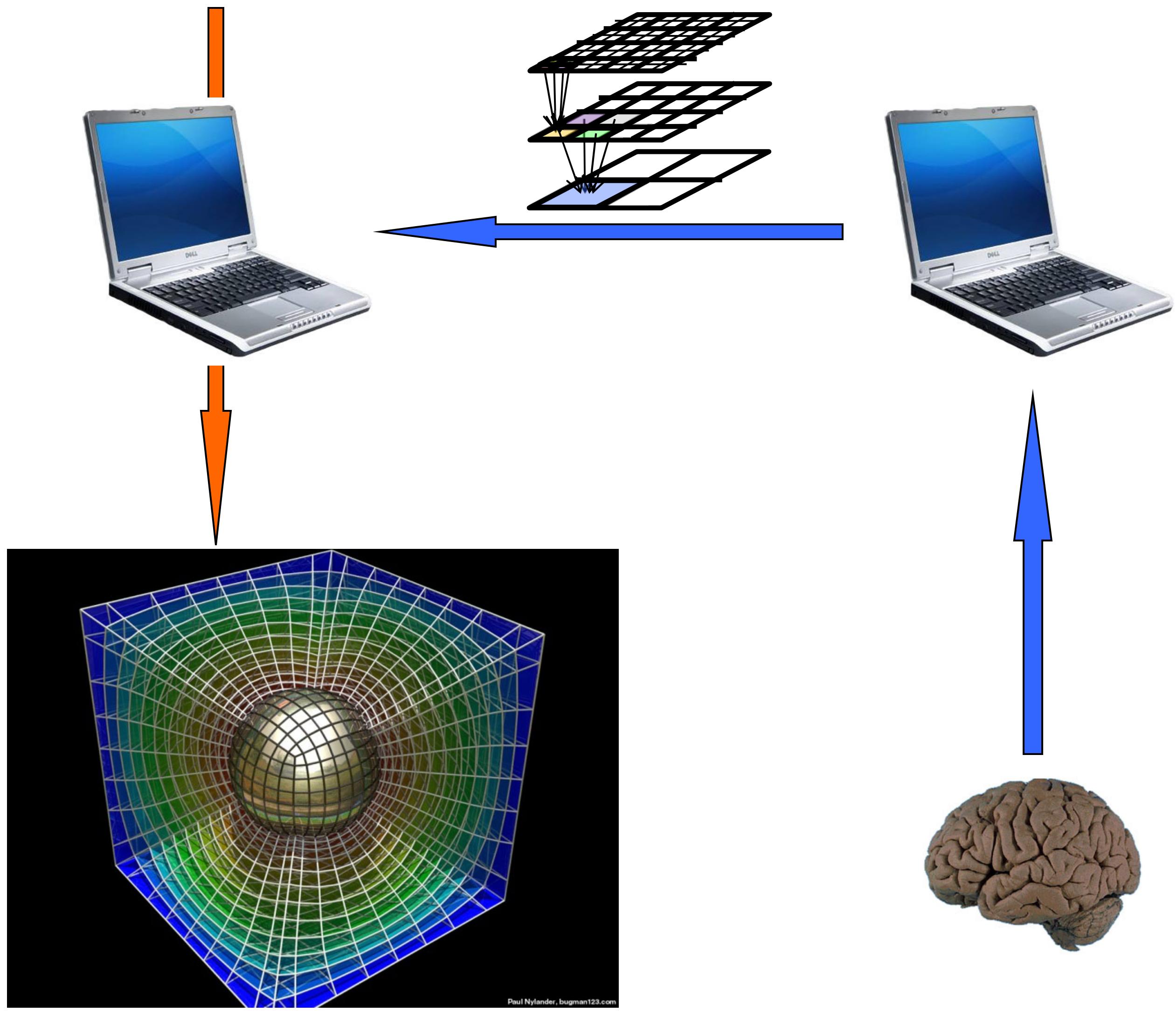
Solving PDEs: Now

$$\Delta u = f$$



Question: Can you program a computer not only solve the PDE but also find the method or the algorithm for solving the PDE?

$$\Delta u = f$$



O., SIAM CSE, March 2015,
https://www.pathlms.com/siam/courses/1043/sections/1259/thumbnail_video_presentations/9883

Kernel methods

- Numerical approximation and statistical inference and intimately connected through the process of making estimations with partial information.
- Well understood, strong theoretical foundations, but relies upon the prior selection of a kernel.

ANNs

- Stem from popularity of deep learning and the popularization of automatic differentiation in Python.
- Does not have much theoretical support yet, but can also be understood as a kernel method with an adapted learned from the data

This talk

- Learning of adapted kernels for numerical approximation

Problem

$$\mathcal{X} \xrightarrow{\quad f^\dagger \quad} \mathbb{R}$$

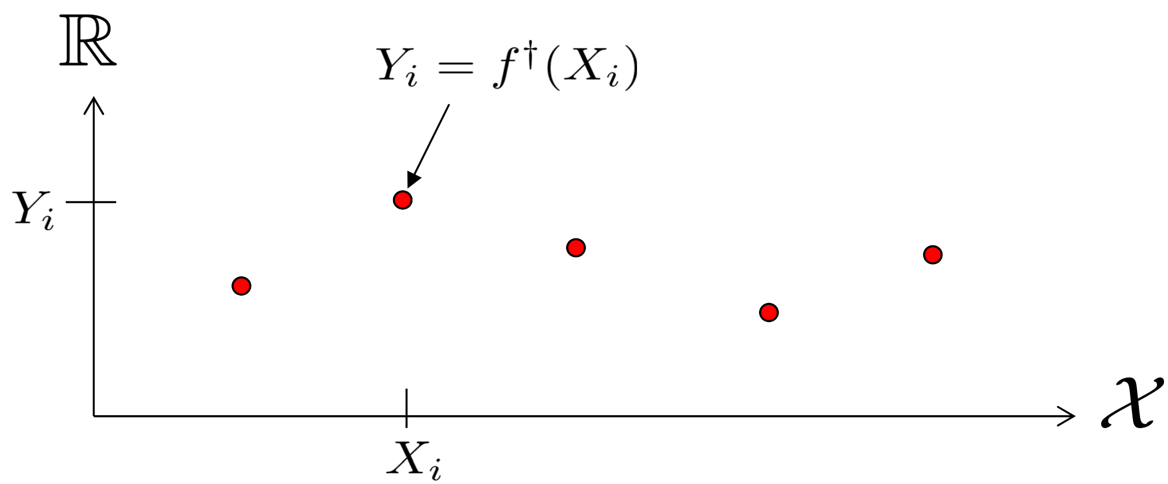
$f^\dagger$: Unknown

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

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

$f^\dagger(X) := \left(f^\dagger(X_1), \ldots, f^\dagger(X_N)\right) \in \mathbb{R}^N$

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



Kernel: $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

What is a kernel?

$\Updownarrow$ For all $m \geq 1$, $x_1, \ldots, x_m \in \mathcal{X}$ the $m \times m$ matrix with entries $K(x_i, x_j)$ is symmetric positive

Feature map:

$\Updownarrow$ $\exists$ a Hilbert space $\mathcal{F}$ and a map $\psi : \mathcal{X} \to \mathcal{F}$ such that

$$K(x, x') = \big\langle \psi(x), \psi(x') \big\rangle_{\mathcal{F}}$$

RKHS space: $\exists$ a Hilbert space $\mathcal{H} := \{f : \mathcal{X} \to \mathbb{R}\}$ such that

$\Updownarrow$

$$f(x) = \big\langle f, K(x, \cdot) \big\rangle_{\mathcal{H}} \text{ for } x \in \mathcal{X},\ f \in \mathcal{H}$$

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

GP: $\exists$ a Gaussian process, $\xi : \mathcal{X} \to$ Gaussian space, such that

$$K(x, x') = \mathbb{E}\big[\xi(x)\xi(x')\big]$$

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

Kernel method solution

Kernel: Approximate $f^\dagger$ with

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

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

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

Feature map: Approximate $f^\dagger$ with

$$\boxed{f(x) = \langle \psi(x), c \rangle_{\mathcal{F}}}$$

$c \in \mathcal{F}$ such that $f(X) = Y$ and $\|c\|_{\mathcal{F}}$ is minimal

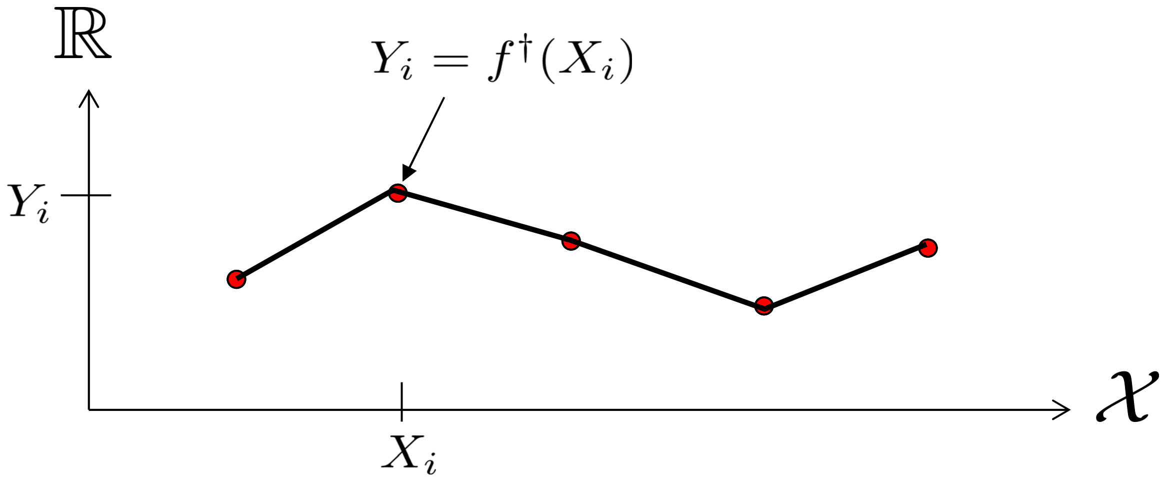
RKHS space: Approximate $f^\dagger$ with minimizer of

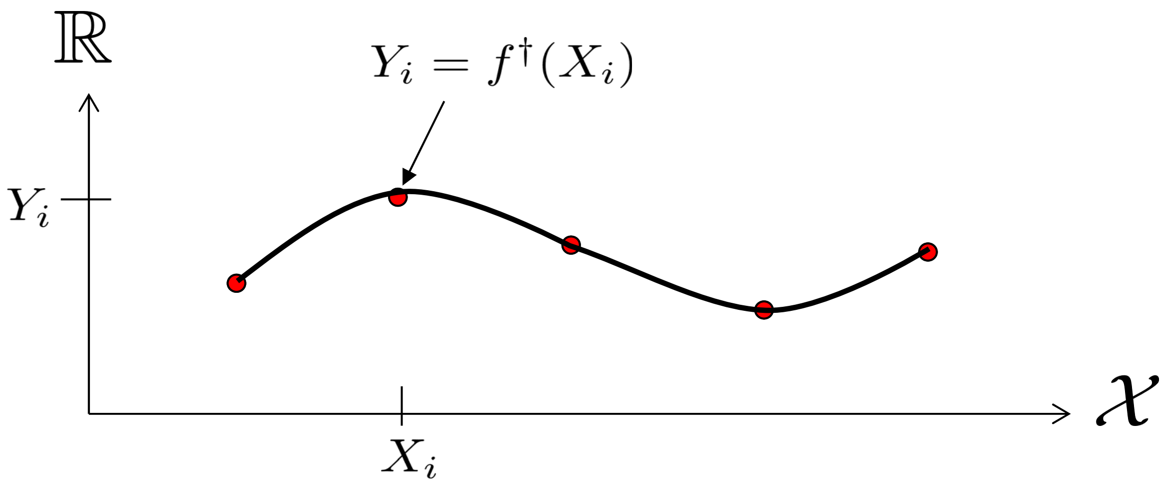
Optimal recovery

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

GPR: Approximate $f^\dagger$ with

$$\boxed{f(x) = \mathbb{E}\big[\xi(x) \big| \xi(X) = Y\big]}$$





Main question

Which kernel do we pick?

Main objectives of this talk

Show why this question is important

Cover 5 main answers from the perspective of numerical approximation

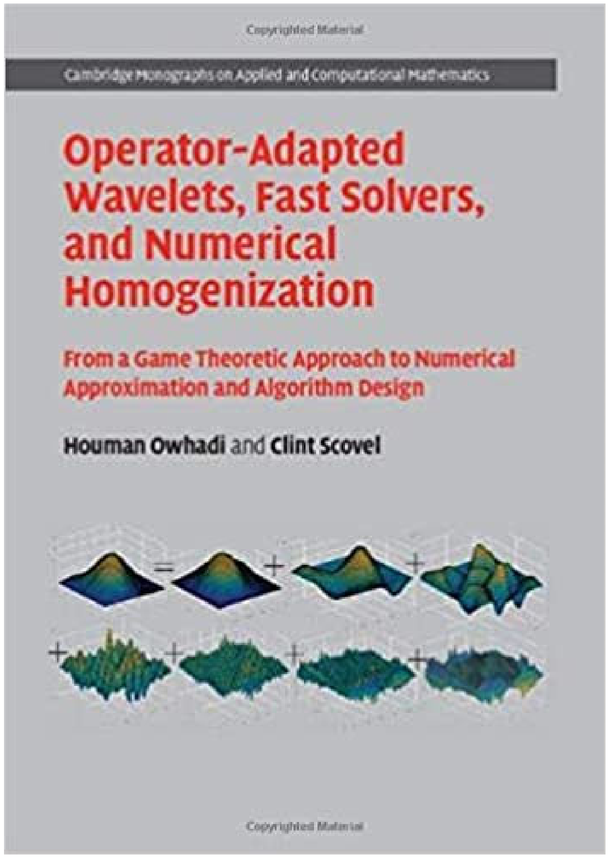
- Use prior informtation on the regularity of $f^\dagger$ (classical numerical approximation approach)
- Use the PDE solved by $f^\dagger$ (current numerical homogenization approach, well understood when the PDE is elliptic and linear)
- Bayesian (MLE, MAP)
- Cross validation
- Deep Learning (Bayesian, MAP)

Most numerical approximation methods are kernel interpolation methods

Sard (1963)

Larkin (1972)

Diaconis (1986)



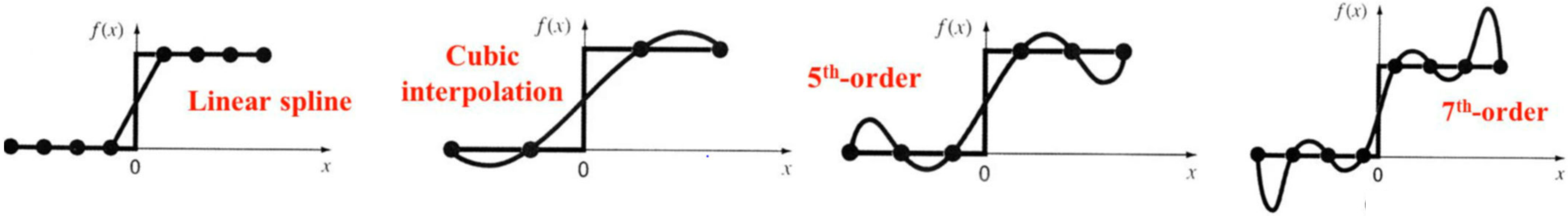
See also: Sul'din (1959). Kimeldorf and Wahba (1970).

Survey: "Statistical Numerical Approximation", O., Scovel, Schäfer, 2019

Book: Cambridge University Press, O., Scovel, 2019

Cardinal splines

[Schoenberg, 1973]

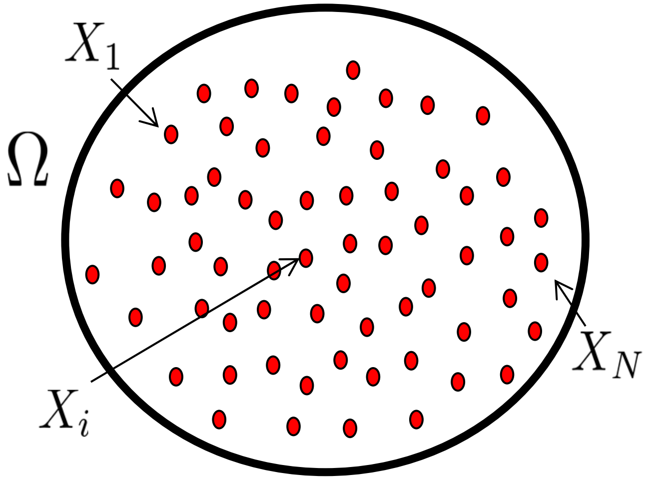


https://slideplayer.com/slide/4635359/

Cardinal spline interpolants are optimal recovery (kernel interpolants) splines

Polyharmonic splines [Harder and Desmarais, 1972], [Duchon, 1977]

$$\boxed{\begin{cases} -\Delta f^\dagger = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases}} \qquad g \in L^2(\Omega)$$



$$\Omega \subset \mathbb{R}^d$$
$$d \leq 3$$

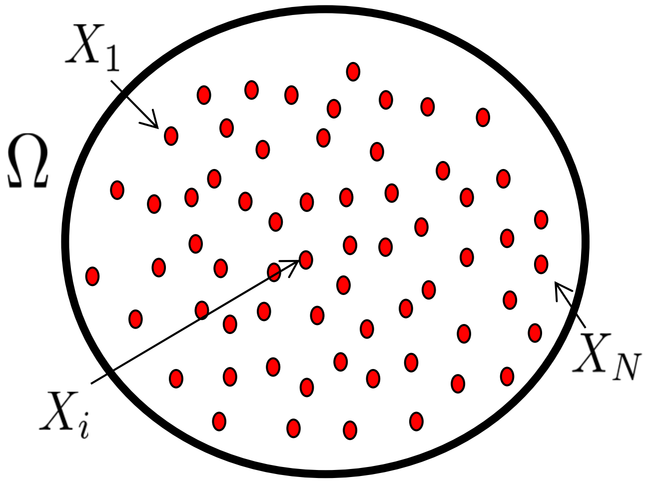
Problem: Given $f^\dagger(X)$ recover $f^\dagger$

$$\boxed{\begin{cases} \text{Minimize} & \int_\Omega |\Delta f|^2 \\ \text{subject to} & f(X) = Y \end{cases}} \qquad \|f^\dagger - f\|_{L^2(\Omega)} \lesssim N^{-\frac{2}{d}} \|g\|_L^2$$

The convergence can be arbitrarily bad if the kernel is not adapted

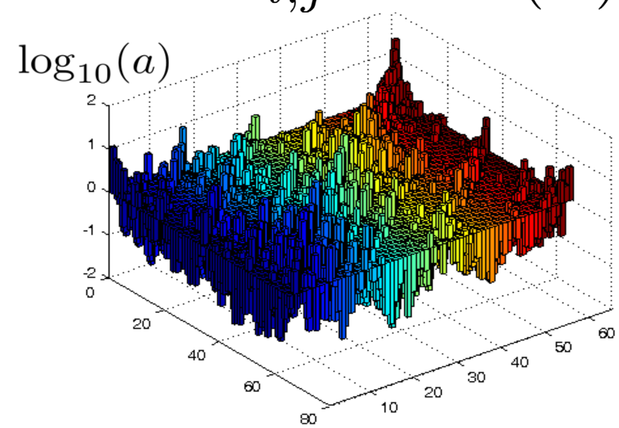
$$\begin{cases} -\operatorname{div}(a\nabla f^\dagger) = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases}$$

$$g \in L^2(\Omega)$$



$$\Omega \subset \mathbb{R}^d$$
$$d \le 3$$

$$a_{i,j} \in L^\infty(\Omega)$$



$$\begin{cases} \text{Minimize} & \int_\Omega |\Delta f|^2 \\ \text{subject to} & f(X) = Y \end{cases}$$

$$\|f^\dagger - f\|_{L^2(\Omega)} \ge \chi(N)$$

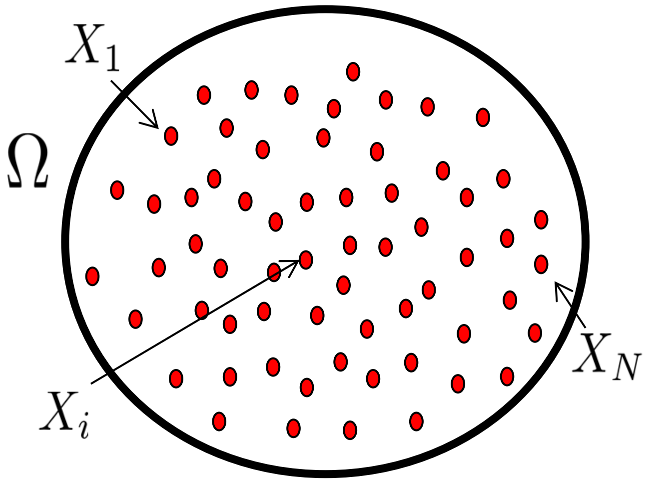
The convergence of $\chi(N)$ towards zero can be arbitrarily slow

[Babuška, Osborn, 2000]: Can a finite element method perform arbitrarily badly?

PDE adapted kernel

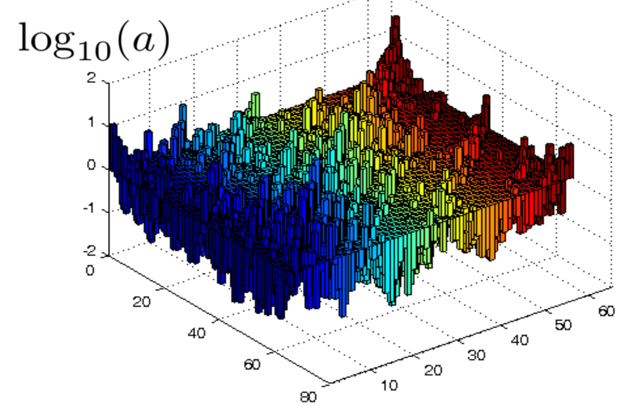
$$\begin{cases} -\operatorname{div}(a\nabla f^\dagger) = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases}$$

$$g \in L^2(\Omega)$$



$$\Omega \subset \mathbb{R}^d$$
$$d \leq 3$$

$$a_{i,j} \in L^\infty(\Omega)$$



$$\begin{cases} \text{Minimize} & \int_\Omega \left|\operatorname{div}(a\nabla f)\right|^2 \\ \text{subject to} & f(X) = Y \end{cases}$$

$$\|f^\dagger - f\|_{L^2(\Omega)} \lesssim N^{-\frac{2}{d}} \|g\|_L^2$$

[O., Berlyand, Zhang, 2014]: Rough polyharmonic splines

PDE adapted Gaussian prior

$$\begin{cases} -\operatorname{div}(a\nabla f^\dagger) = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases} \qquad g \in L^2(\Omega)$$

$$\Updownarrow$$

$$\begin{cases} -\operatorname{div}(a\nabla \xi) = \zeta, & x \in \Omega, \\ \xi = 0, & x \in \partial\Omega, \end{cases} \qquad \zeta \sim \mathcal{N}(0, \delta(x-x'))$$

$$\boxed{f(x) = \mathbb{E}\big[\xi(x)\big|\xi(X) = Y\big]} \qquad \|f^\dagger - f\|_{L^2(\Omega)} \lesssim N^{-\frac{2}{d}}$$

[O., 2014]: Bayesian Numerical Homogenization

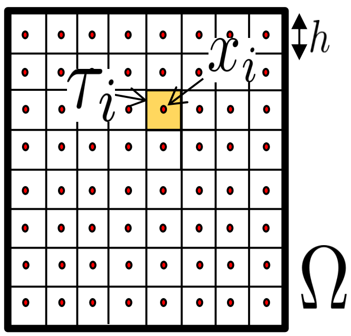
[O., 2015], [O., Zhang, 2016], [O., Scovel, 2019], [Schäfer, Sullivan, O., 2017]: Gamblets

Current popular kernel for the numerical homogenization of elliptic PDEs

$$(1)\quad \boxed{\begin{cases} -\operatorname{div}(a\nabla f^\dagger) = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases}}$$

G : Green's function

The solution of (1) is $f^\dagger(x) = \int_\Omega G(x,y)g(y)\,dy$



The numerical homogenization approximation of (1) is

$$\boxed{f(x) = \sum_i c_i \int_\Omega G(x,y)\phi_i(y)\,dy}$$

[Hughes, 1995]: Variational Multiscale Method.

[Malqvist, Peterseim, 2012-2014]: Local Orthogonal Decomposition.

[O., 2015]: Gamblets

Which kernel do we pick?

The answer is by now well understood if the regularity of $f^\dagger$ is known a priori or if $f^\dagger$ is the solution of a known linear elliptic PDE

What if the underlying regularity of $f^\dagger$ is unknown?

Kernel Flows: from learning kernels from data into the abyss.
H. Owhadi and G. R. Yoo, arXiv:1808.04475.
Journal of Computational Physics, 2019

Gene Ryan Yoo

Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation. Y. Chen, H. Owhadi, A. M. Stuart. 2020. arXiv:2005.11375

Yifan Chen

Andrew Stuart

Interpolation problem

Recover $f^\dagger : D \subset \mathbb{R}^d \to \mathbb{R}$

Given $f^\dagger(X_i)$, for $i = 1, \ldots, N$

Family of kernels

$K_\theta : D \times D \to \mathbb{R}$

θ: Hierarchical parameter

Kernel/GP interpolant

$$f(\cdot, \theta, X) = K_\theta(\cdot, X) K_\theta(X, X)^{-1} f^\dagger(X)$$

Question

Which θ do we pick?

Empirical Bayes answer

Place a prior on θ

Assume that $f^\dagger|\theta \sim \mathcal{N}(0, K_\theta)$

Select the θ maximizing the marginal probability of θ suject to conditionning on $f^\dagger(X)$

Uniformative prior on θ

$\Downarrow$

Maximum Likelihood Estimate

$$\boxed{\theta^{EB} = \underset{\theta}{\operatorname{argmin}}\, L^{EB}(\theta, X, f^\dagger)}$$

$$\boxed{L^{EB}(\theta, X, f^\dagger) = f^\dagger(X)^T K_\theta(X,X)^{-1} f^\dagger(X) + \log\det K_\theta(X,X)}$$

Kernel Flow answer (Variant of cross-validation, O., Yoo, 2019)

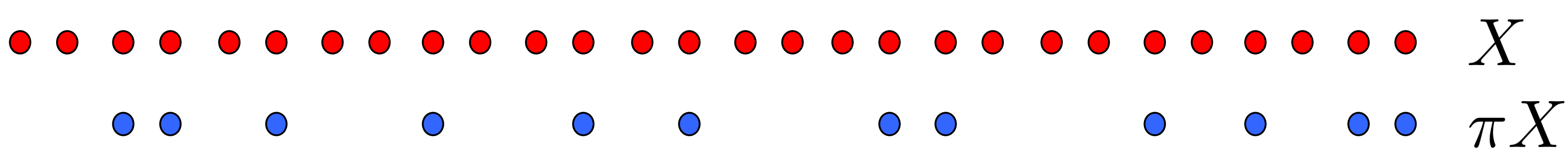
Pick a θ such that subsampling the data does not influence the interpolant much

$$\boxed{\theta^{KF} = \underset{\theta}{\operatorname{argmin}}\, L^{KF}(\theta, X, \pi X, f^\dagger)}$$

$$\boxed{L^{KF}(\theta, X, \pi X, f^\dagger) = \frac{\left\| f(\cdot,\theta,X) - f(\cdot,\theta,\pi X) \right\|_{K_\theta}^2}{\left\| f(\cdot,\theta,X) \right\|_{K_\theta}^2}}$$

$$f(\cdot,\theta,X) = K_\theta(\cdot,X) K_\theta(X,X)^{-1} f^\dagger(X)$$

π: subsampling operator, πX is a subvector of X



$\|\cdot\|_{K_\theta}$: RKHS norm determined by K_θ

A kernel is good if subsampling the data does not influence the interpolant much

Question

How do θ^{EB} and θ^{KF} behave as # of data $\to \infty$

Model

- Domain $D = \mathbb{T}^d = [0,1]^d_{\text{per}}$
- Lattice data $X_q = \{j \cdot 2^{-q}, j \in J_q\}$

  where $J_q = \{0, 1, \ldots, 2^q - 1\}^d$, # of data 2^{qd}
- Kernel $K_\theta = (-\Delta)^{-\theta}$
- Subsampling in KF: $\pi X_q = X_{q-1}$

Theorem

(Chen, O., Stuart, 2020)

If $f^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$ for some $s > d/2$, then as $q \to \infty$

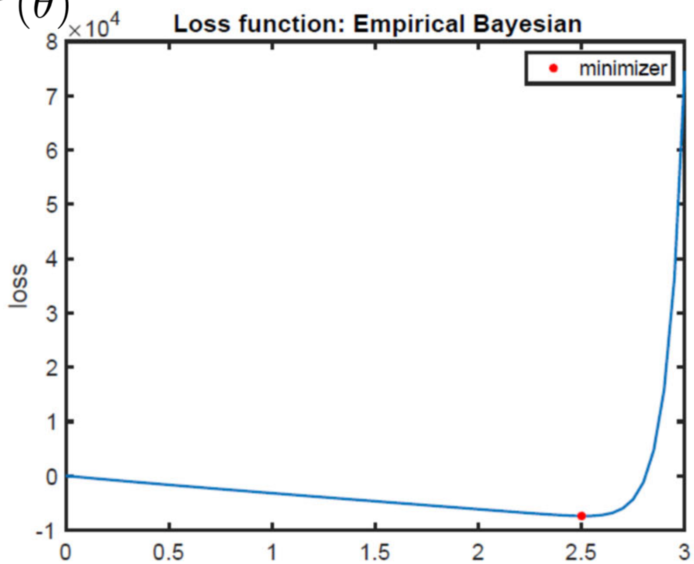
$\theta^{EB} \to s$ and $\theta^{KF} \to \frac{s - \frac{d}{2}}{2}$ in probability

Experiment

$d = 1,\ s = 2.5,\ \#$ of data $N = 2^9$

EB loss

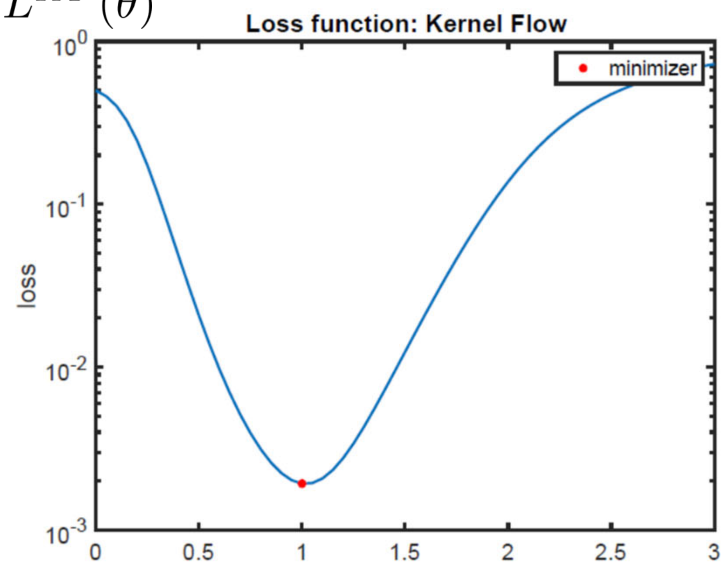
$L^{EB}(\theta)$



θ

KF loss

$L^{KF}(\theta)$



θ

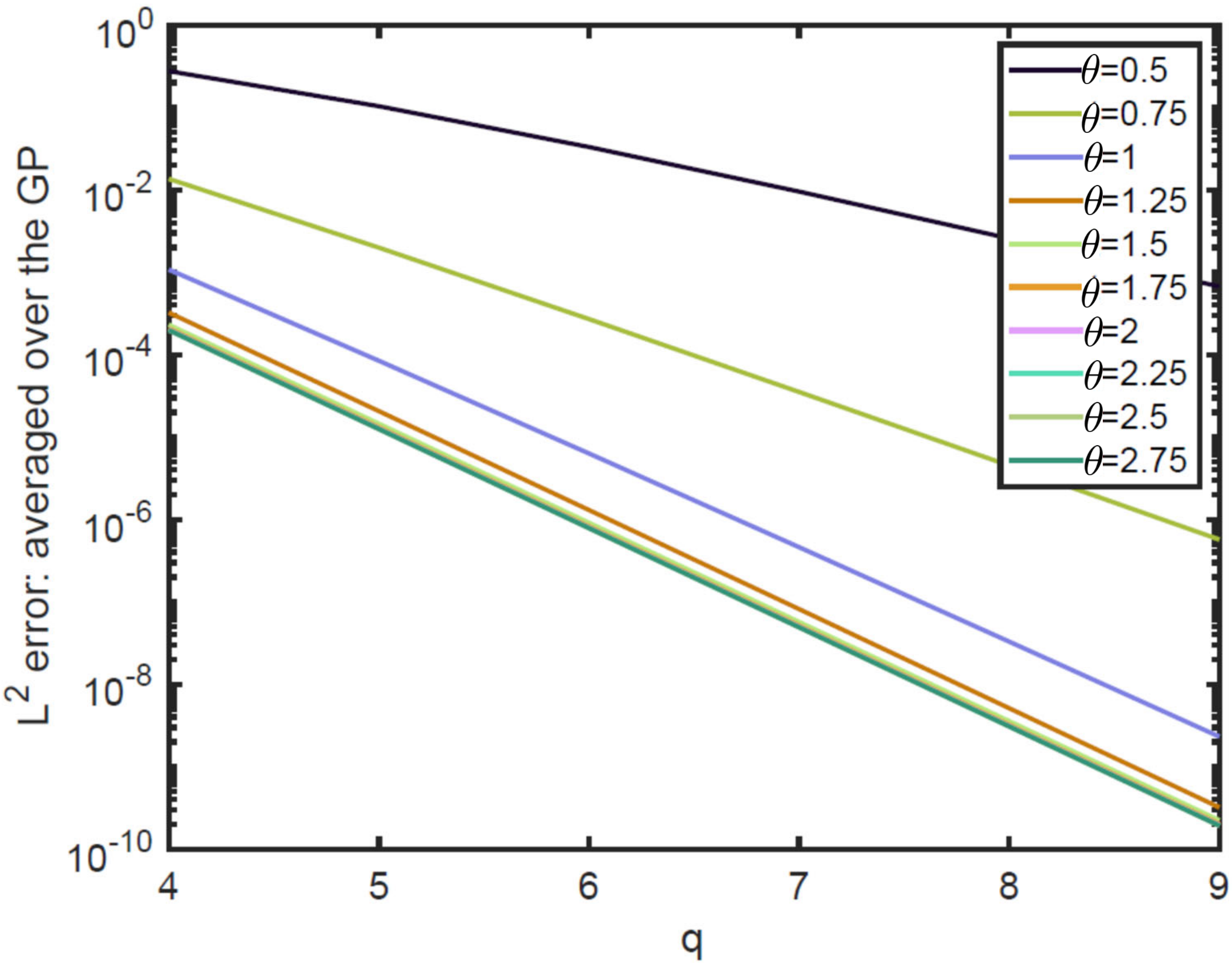
Question?

How are the limits s $(= 2.5)$ and $\frac{s - \frac{d}{2}}{2}$ $(= 1)$ special?

What is the implicit bias in the EB and KF algorithms?

Experiment 1

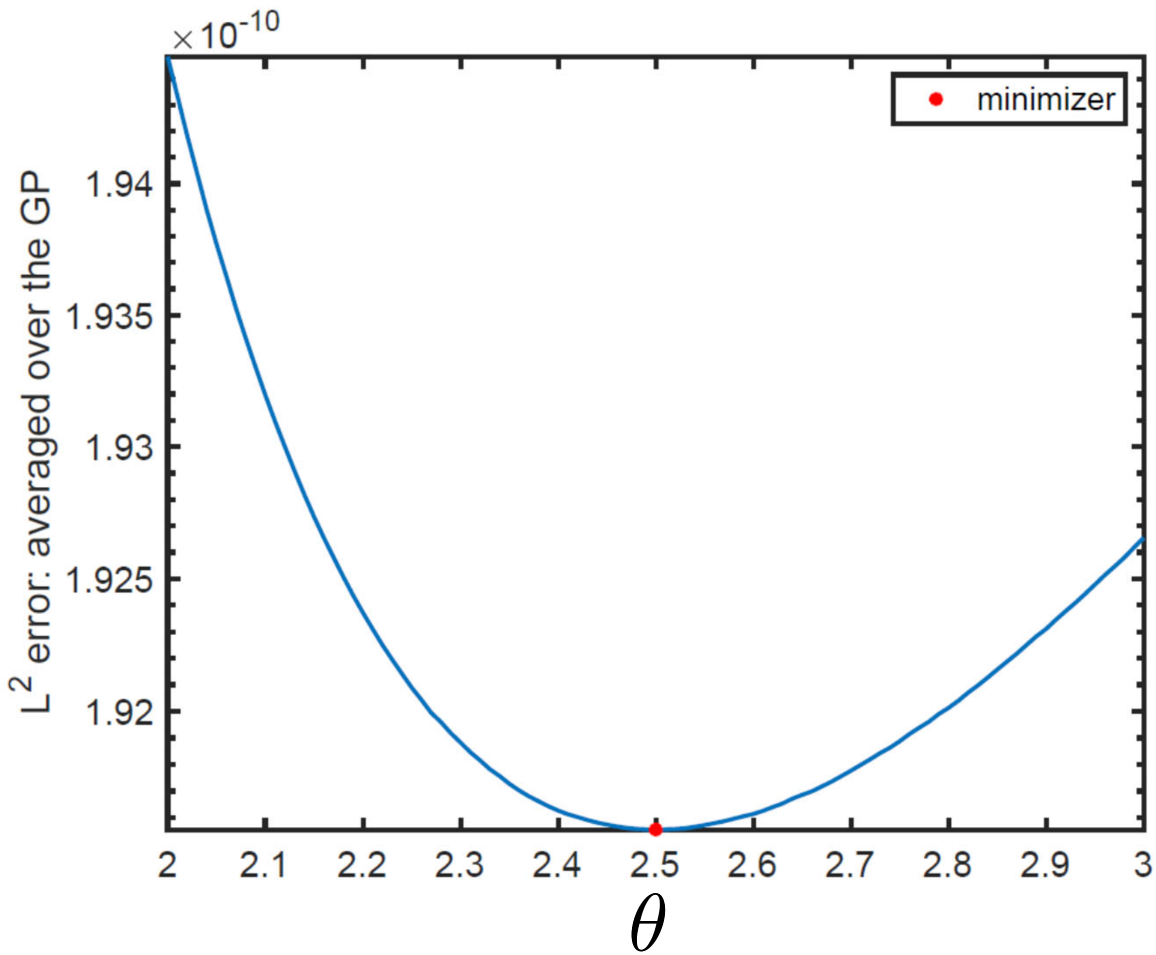
- Compute $\mathbb{E}_{f^\dagger}\left\|f^\dagger(\cdot) - f(\cdot,\theta,X_q)\right\|_{L^2}^2$ vs q



- $\frac{s-\frac{d}{2}}{2}$ $(=1)$ is the smallest θ that suffices to achieve fastest rate in L^2

Experiment 2

- $q = 9$. Compute $\mathbb{E}_{f^\dagger} \left\| f^\dagger(\cdot) - f(\cdot, \theta, X_q) \right\|_{L^2}^2$ vs θ



- s $(= 2.5)$ is the θ that minimizes the mean squared error

Takeaway message

- EB selects the θ that minimizes the mean squared error.
- KF selects the smallest θ that suffices for the fastest rate of convergence in mean squared error.

More comparisons

- EB may be brittle (not robust) to model misspecification
- KF has some degree of robustness to model misspecification

G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross validation. 1980.

M. L. Stein. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. 1990.

F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. 2013.

Chen, O., Stuart. Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation. 2020.

Extrapolation problem

Given time series $z_1, \dots, z_N$

predict $z_{N+1}, z_{N+2}, z_{N+3}, \dots$

Assumption

$$\boxed{z_{k+1} = f^\dagger(z_k, \dots, z_{k-\tau^\dagger+1})}$$

$f^\dagger$, $\tau^\dagger$ unknown

Fundamental problem

[Box, Jenkins, 1976]: Time Series Analysis

Mezíc, Klus, Budišić, R. Mohr,...: Koopman operator

[Alexander, Giannakis, 2020]: Operator theoretic framework

[Bittracher et al, 2019]: kernel embeddings of transition manifolds

[Brunton, Proctor, Kutz, 2016]: SINDy

Brian, Hunt, Ott, Pathak, Lu, Hunt, Girvan, Ott,...: Reservoir computing

Ralaivola, Chattopadhyay,...: LSTM

Simplest solution

Approximate $f^\dagger$ with Kernel interpolant f

$$\boxed{f(z_k, \ldots, z_{k-\tau^\dagger+1}) = z_{k+1}} \qquad k = \tau^\dagger, \tau^\dagger + 1, \ldots, N-1$$

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

$$X_k = (z_k, \ldots, z_{k-\tau^\dagger+1})$$

$$Y_k = z_{k+1} = f^\dagger(X_k)$$

Predict future values of the time series by simulating the dynamical system

$$\boxed{s_{k+1} = f(s_k, \ldots, s_{k-\tau^\dagger+1})}$$

Learning dynamical systems from data: a simple cross-validation perspective. B. Hamzi and H. Owhadi. 2020. arXiv:2007.05074

Boumediene Hamzi

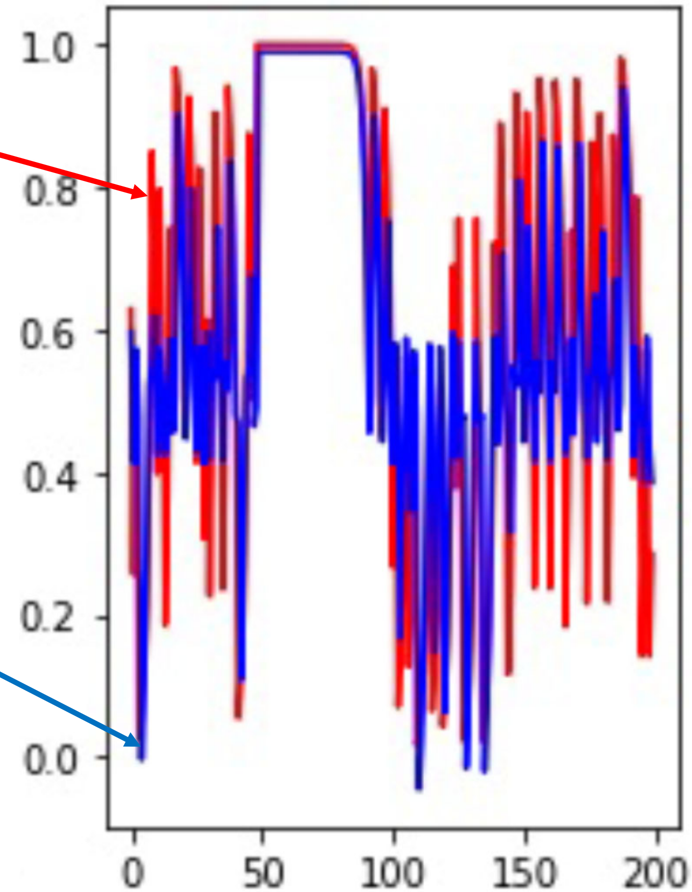
Example: Bernoulli map

$$z_{k+1} = 2z_k \mod 1$$

$$K(x,x') = e^{-\|x-x'\|^2}$$

True dynamic

Predicted dynamic



Example: Bernoulli map

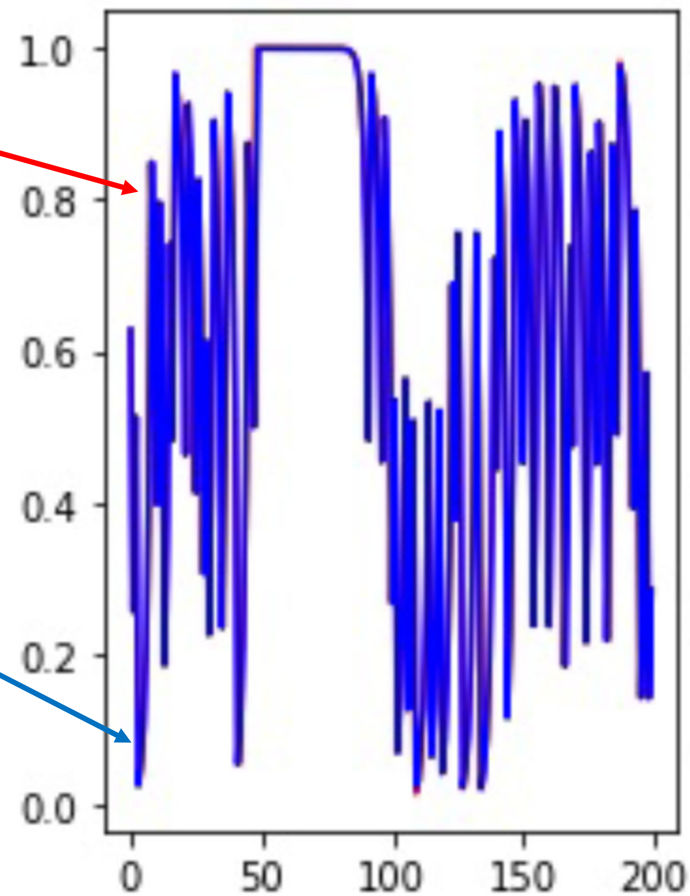
$$\boxed{z_{k+1} = 2z_k \mod 1}$$

$$\boxed{K(x,x') = \alpha_0 \max\{0, 1 - \frac{\|x-x'\|^2}{\sigma_0}\} + \alpha_1 e^{-\frac{\|x-x'\|^2}{\sigma_1^2}}}$$

True dynamic

$\alpha_0, \sigma_0, \alpha_1, \sigma_1^2$:
Learned parameters
(using Kernel Flows)

Predicted dynamic



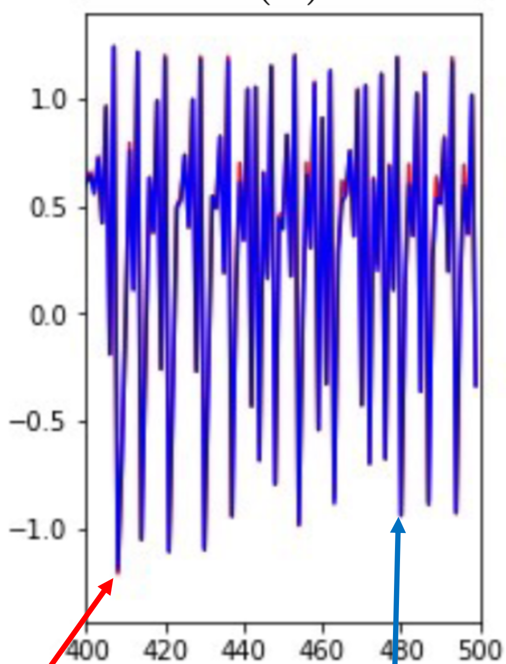
Example: Hénon map

$$
\begin{aligned}
x(k+1) &= 1 - ax(k)^2 + y(k) \\
y(k+1) &= bx(k)
\end{aligned}
$$

$$K(x,x') = \begin{pmatrix} k_1(x,x') & 0 \\ 0 & k_2(x,x') \end{pmatrix}$$

$$k_i(x,y) = \alpha_i + (\beta_i + ||x-y||_2^{\kappa_i})^{\sigma_i} + \delta_i e^{-||x-y||_2^2/\mu_i^2}$$

$x(k)$



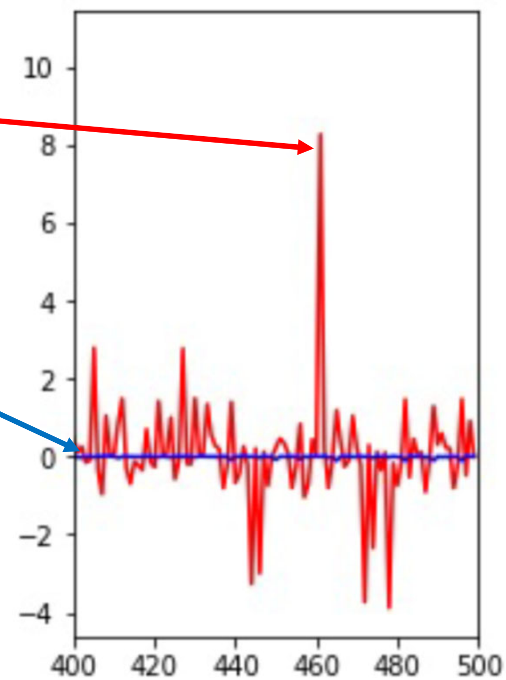
True dynamic Predicted dynamic

Learned kernel

Prediction error

Initial kernel

Learned kernel

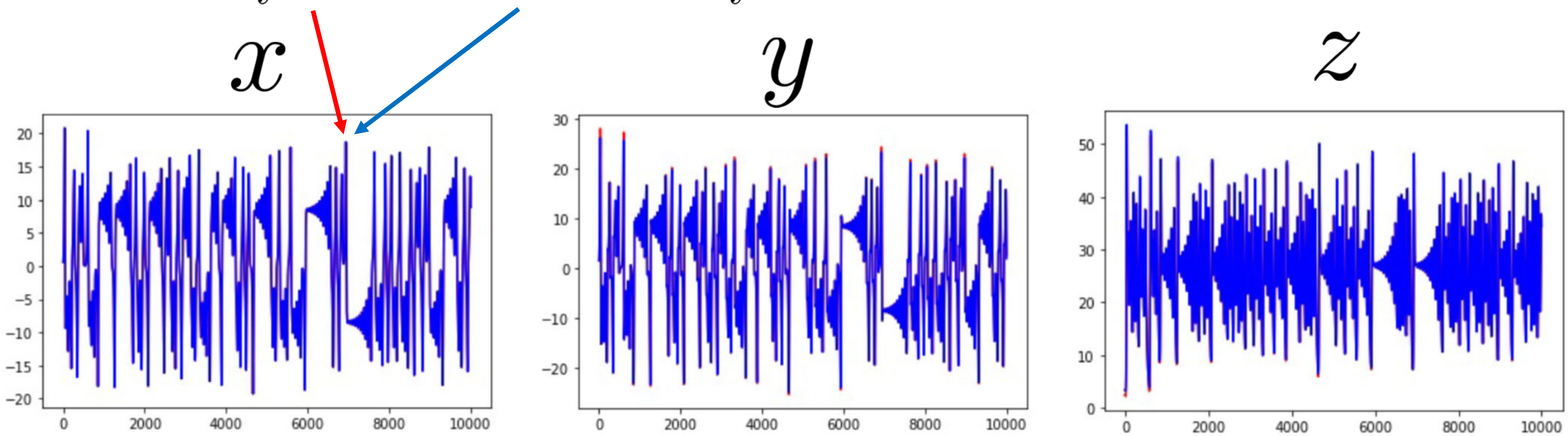


Example: Lorenz system

$$\begin{aligned}
\frac{dx}{dt} &= s(y-x) \\
\frac{dy}{dt} &= rx - y - xz \\
\frac{dz}{dt} &= xy - bz
\end{aligned}$$

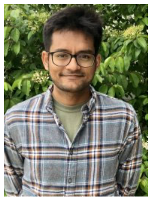
$$k_i(x,y) = \alpha_i + (\beta_i + ||x-y||_2^{\kappa_i})^{\sigma_i} + \delta_i e^{-||x-y||_2^2/\mu_i^2}$$

True dynamic   Predicted dynamic with learned kernel

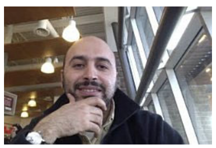


Not limited to toy problems

Also works for extrapolating climate/weather time series



Romit
Maulik
(ANL)



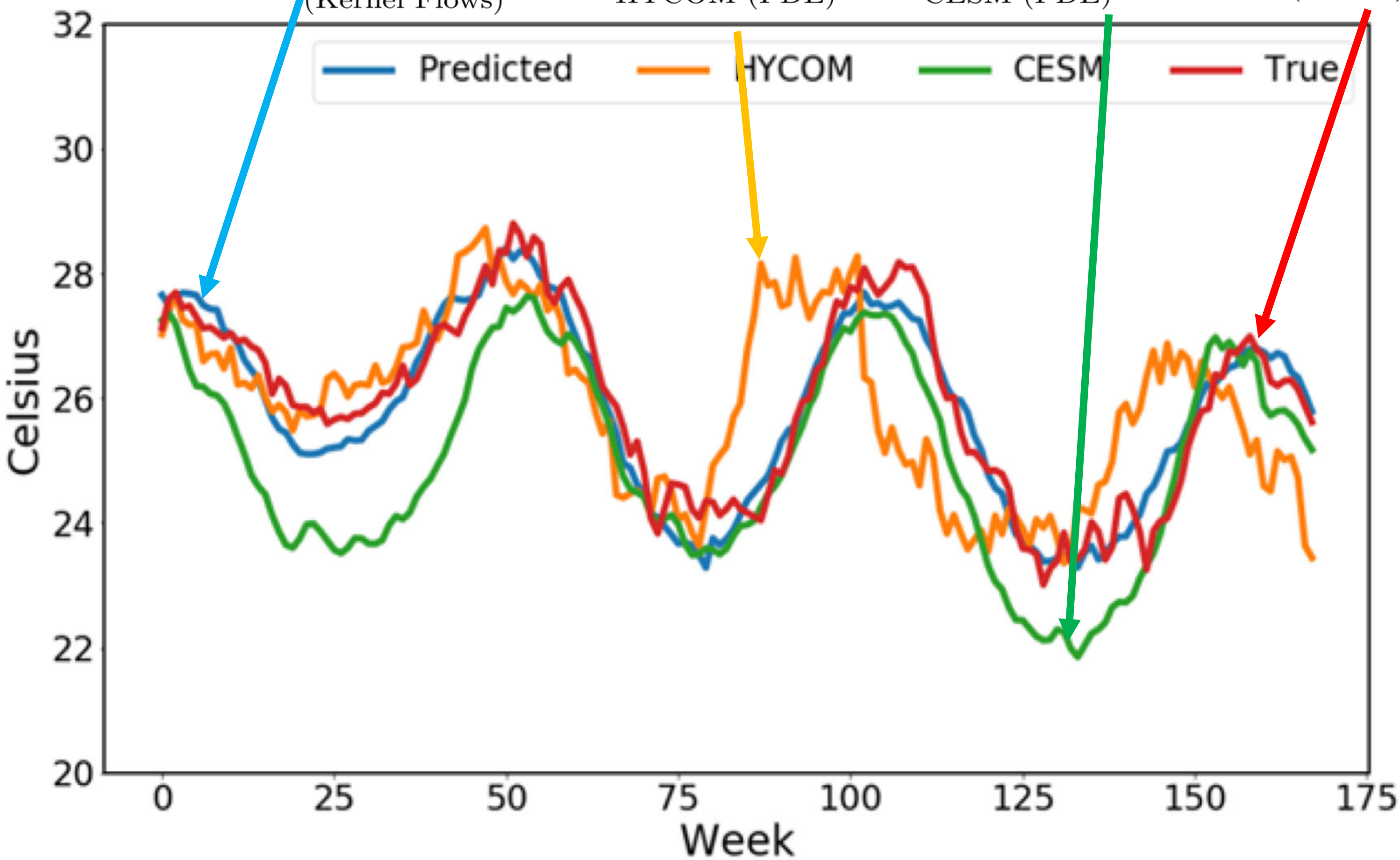
Boumediene
Hamzi

Predicted
(Kernel Flows)

HYCOM (PDE)

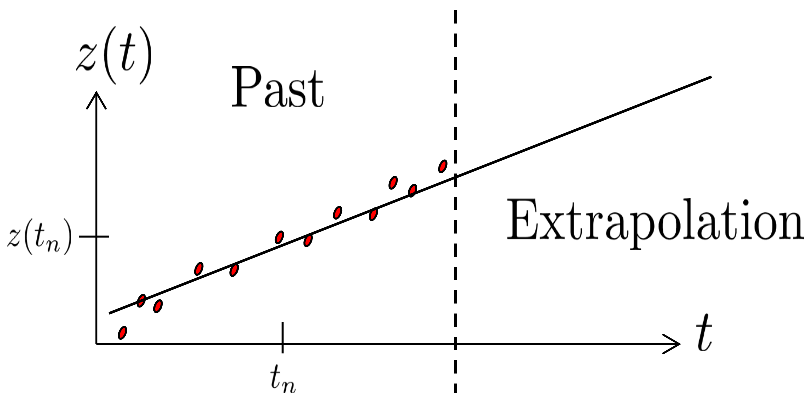
CESM (PDE)

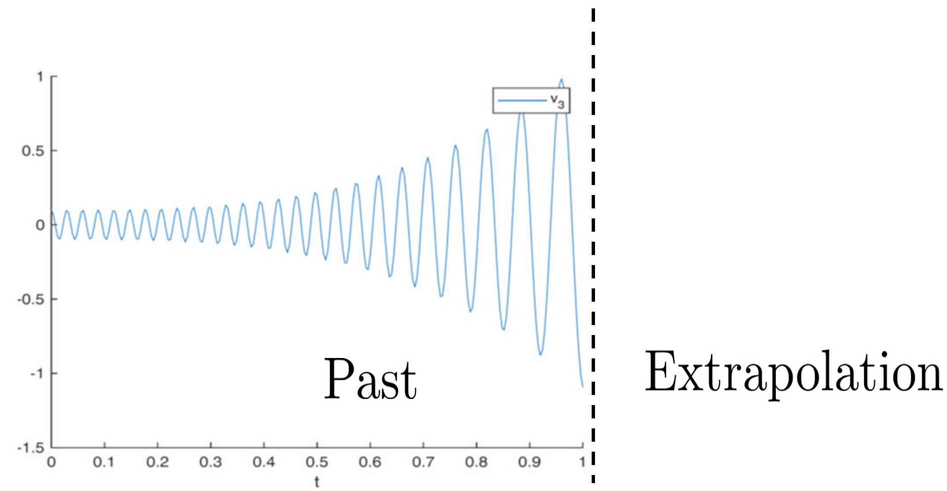
True (NOAA)



Takeaway message

Kernel methods can perform well on extrapolation problems if the kernel is also learned from data





Learning dynamical systems from data: a simple cross-validation perspective. B. Hamzi and H. Owhadi. 2020. arXiv:2007.05074

Kernel Mode Decomposition and programmable/interpretable regression networks, O., Scovel, Yoo, 2019 arXiv:1907.08592

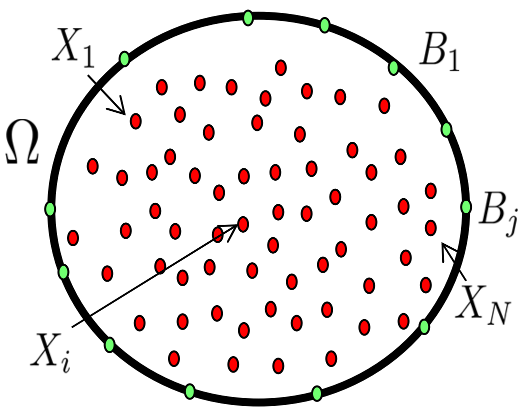
Which kernel do we pick?

- Deep learning approach

Physics informed neural networks [Raissi, Perdikaris, Karniadakis, 2017]

$$\boxed{\begin{cases} \mathcal{L}f^\dagger = g, & x \in \Omega, \\ f^\dagger = 0, & x \in \partial\Omega, \end{cases}}$$

$$g \in C(\Omega)$$



$f(x, \theta)$: Neural network with parameters θ

$$\boxed{\min_\theta \sum_i \left|\mathcal{L}f(X_i, \theta) - g(X_i)\right|^2 + \sum_j \left|f(B_j, \theta)\right|^2}$$

Can be remarkably efficient on complex problems with partial information:

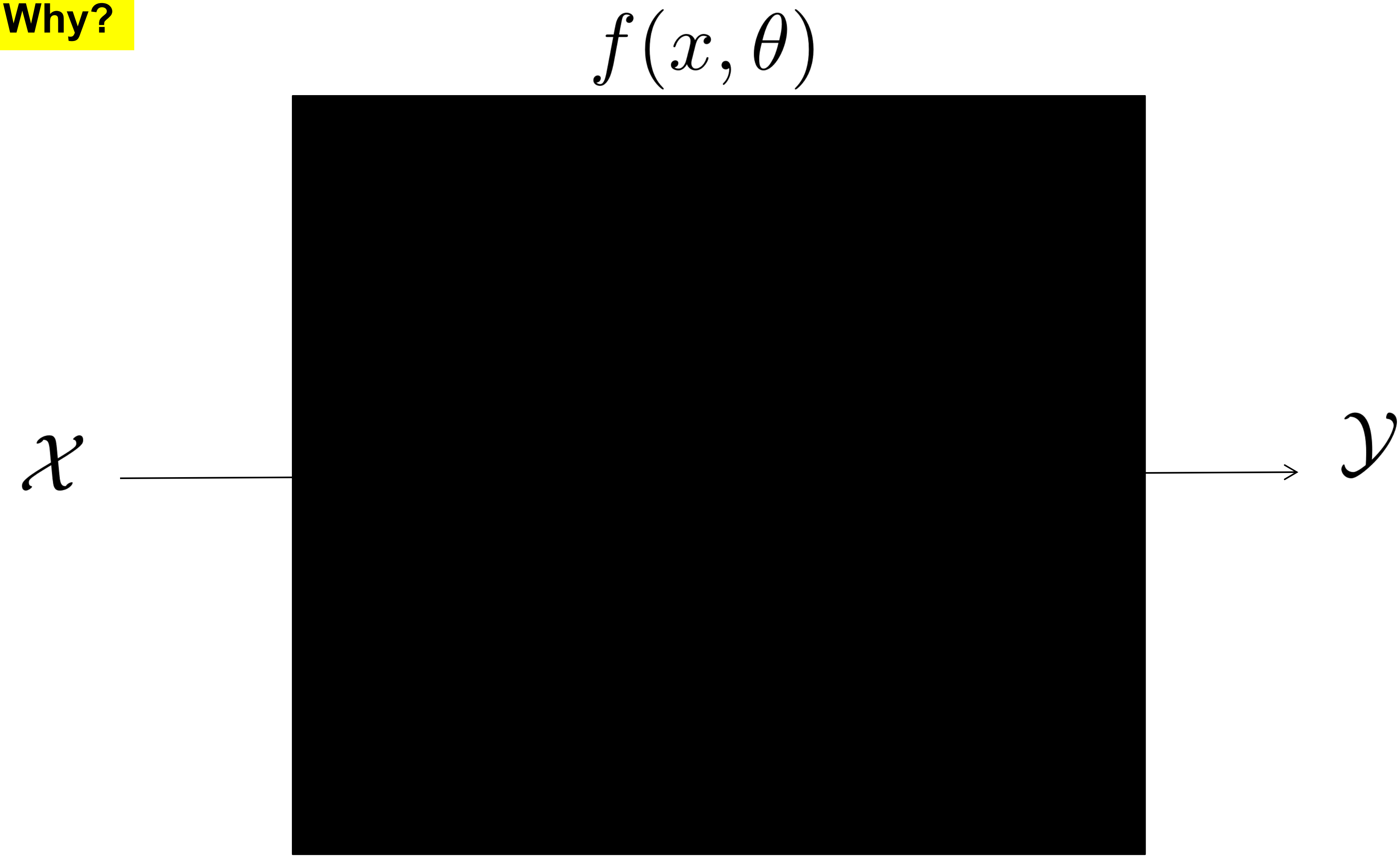
[Raissi, Yazdani, Karniadakis, Science 2020]: Hidden fluid mechanics, Learning velocity and pressure fields from flow visualizations

But can fail on simple problems

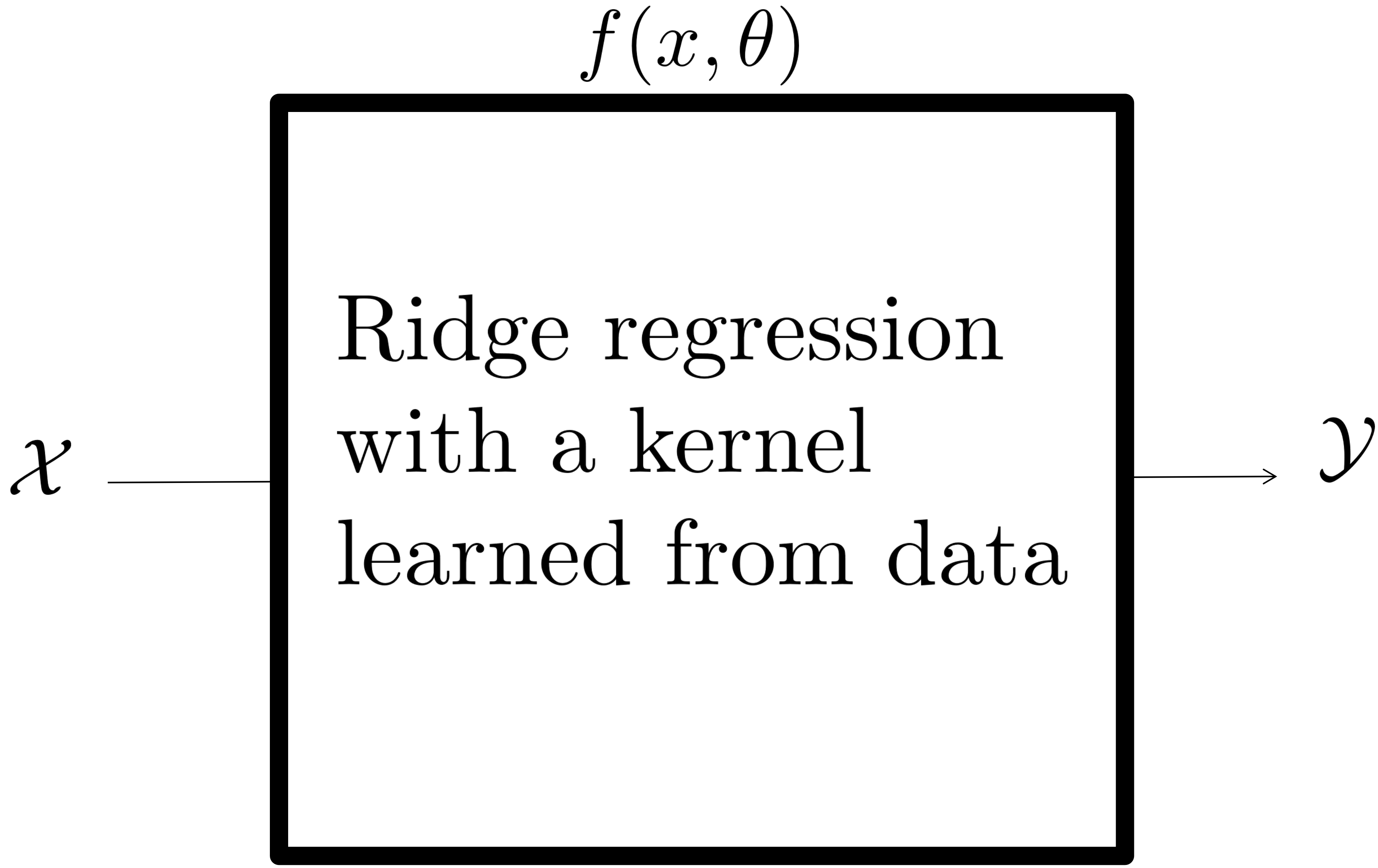
[Wang, Yu, Perdikaris, 2020]: When and why PINNs fail to train.

[van der Meer, Oosterlee, Borovykh, 2020]: Can fail on simple problems

Why?



We need to open the neural network back box



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

Problem

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

$f^{\dagger}$: Unknown

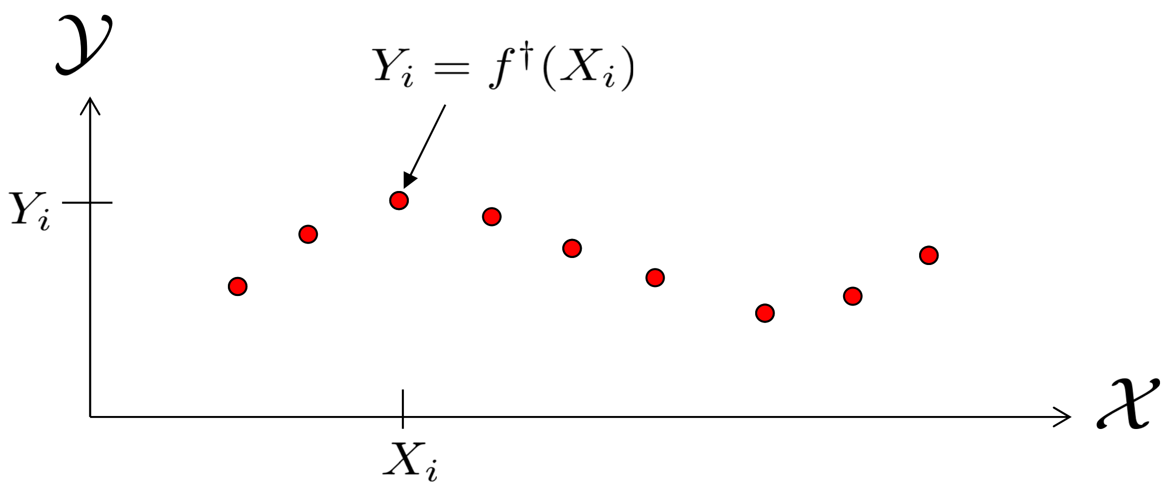
Given $f^{\dagger}(X) = Y$ with $(X, Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^{\dagger}$

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

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

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

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



Problem

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

$f^\dagger$: Unknown

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

Ridge regression solution

Approximate $f^\dagger$ with minimizer of

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

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

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

$\mathcal{L}(\mathcal{Y})$: Set of bounded linear operators on $\mathcal{Y}$.

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

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

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

Artificial neural network solution

Approximate $f^\dagger$ with

$$\boxed{f = f_D \circ \cdots \circ f_1}$$

$$\mathcal{X} = \mathcal{X}_1 \xrightarrow{f_1} \mathcal{X}_2 \quad \cdots \quad \mathcal{X}_k \xrightarrow{f_k} \mathcal{X}_{k+1} \quad \cdots \quad \mathcal{X}_{D+1} = \mathcal{Y}$$

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

$\mathbf{a}$: 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

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

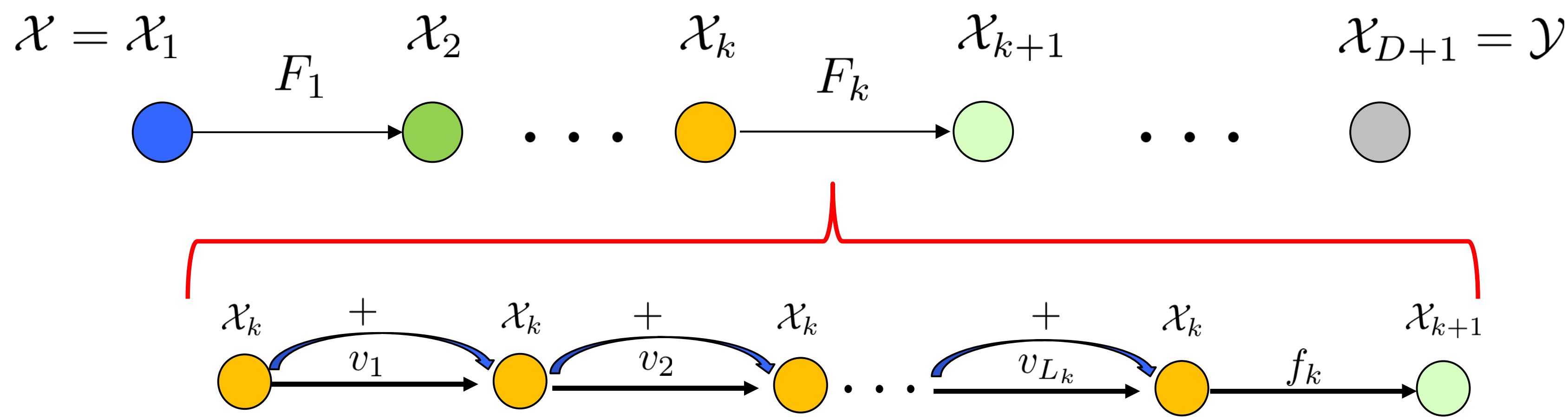
$$\|Y\|^2_{\mathcal{Y}^N} := \sum_{i=1}^N \|Y_i\|^2_{\mathcal{Y}}$$

Residual neural network solution

Approximate $f^\dagger$ with

[He et Al, 2016]

$$\boxed{f = F_D \circ \cdots \circ F_1}$$



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

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

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

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

ODE/Dynamical system interpretation of ResNets

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

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

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

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

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

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

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

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

Mechanical regression

Approximate $f^\dagger$ with

$$\boxed{f^\ddagger = f \circ \phi_L}$$

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

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

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

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

Particular case: ResNet block with L2 regularization on weights and biases!

Particular case

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

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

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

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

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

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

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

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

Mechanical regression

Approximate $f^\dagger$ with

$$\boxed{f^\ddagger = f \circ \phi_L}$$

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

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

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

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

Theorem

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

$$\boxed{f \circ \phi^v(x)}$$

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

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

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

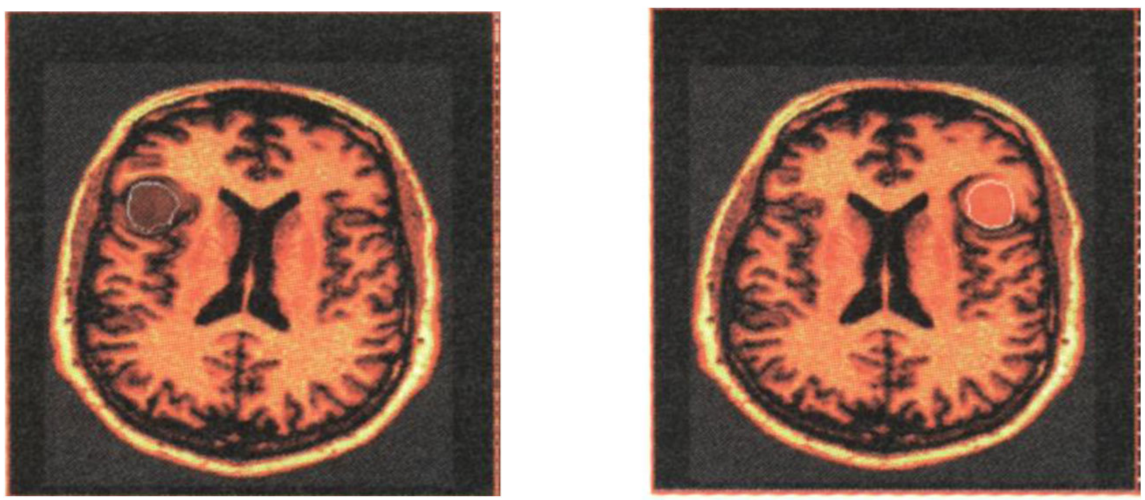
What kind of optimization problem is this?

Looks like an image registration/computational anatomy variational problem

Image registration

How to best align image I and image I'?

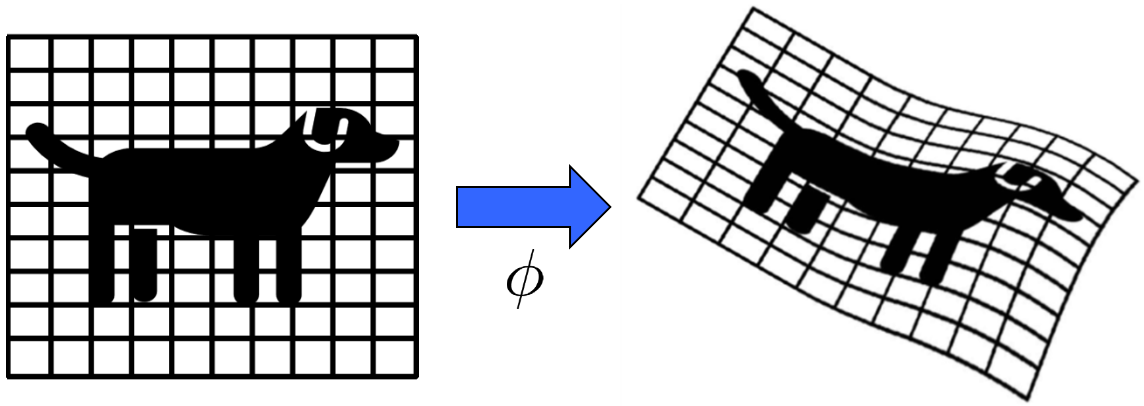
I $\qquad$ I'



[Grenander, Miller, 1998]: Computational anatomy

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

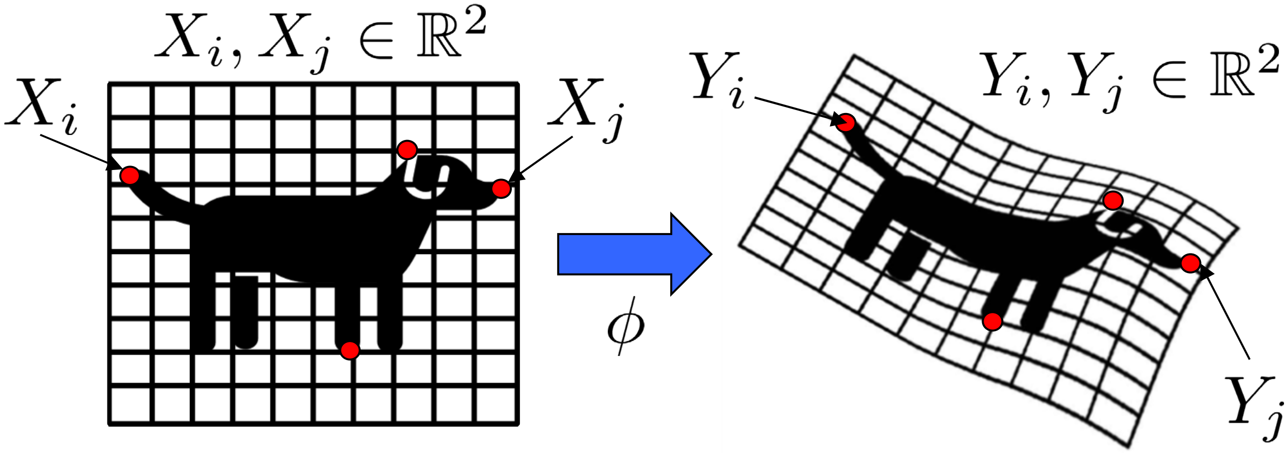
Image registration



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

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

Image registration with landmarks

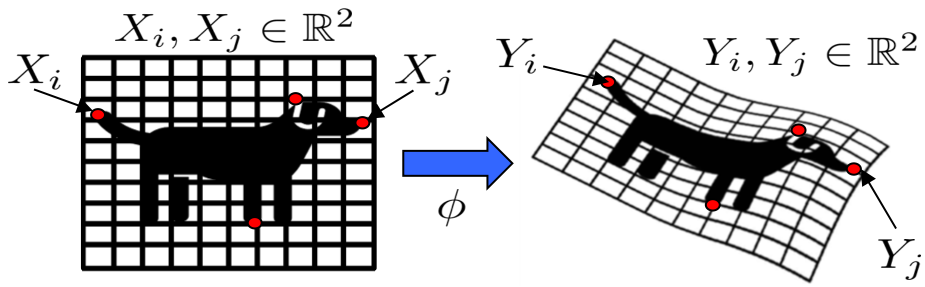


$$\min_v \lambda \int_0^1 \|\Delta v\|^2_{L^2([0,1]^2)}\, dt + \sum_i |\phi^v(X_i,1) - Y_i|^2$$

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

[Joshi, Miller, 2000]: Landmark matching

Image registration with landmark matching

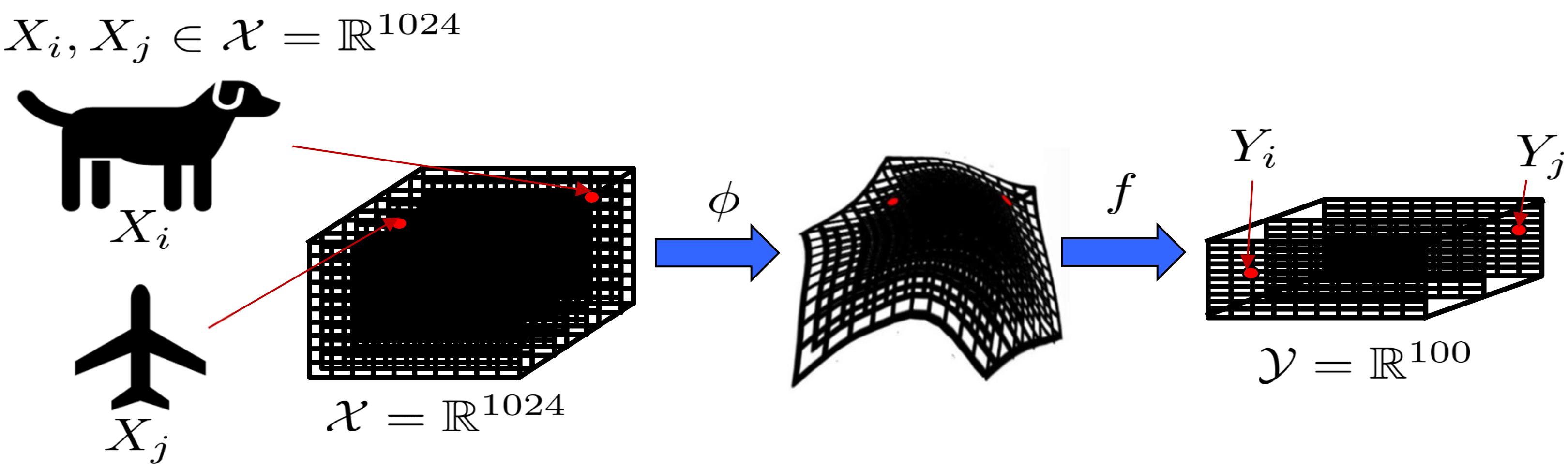


$$\min_v \lambda \int_0^1 \|\Delta v\|^2_{L^2([0,1]^2)}\, dt + \sum_i |\phi^v(X_i,1) - Y_i|^2$$

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

Generalization

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



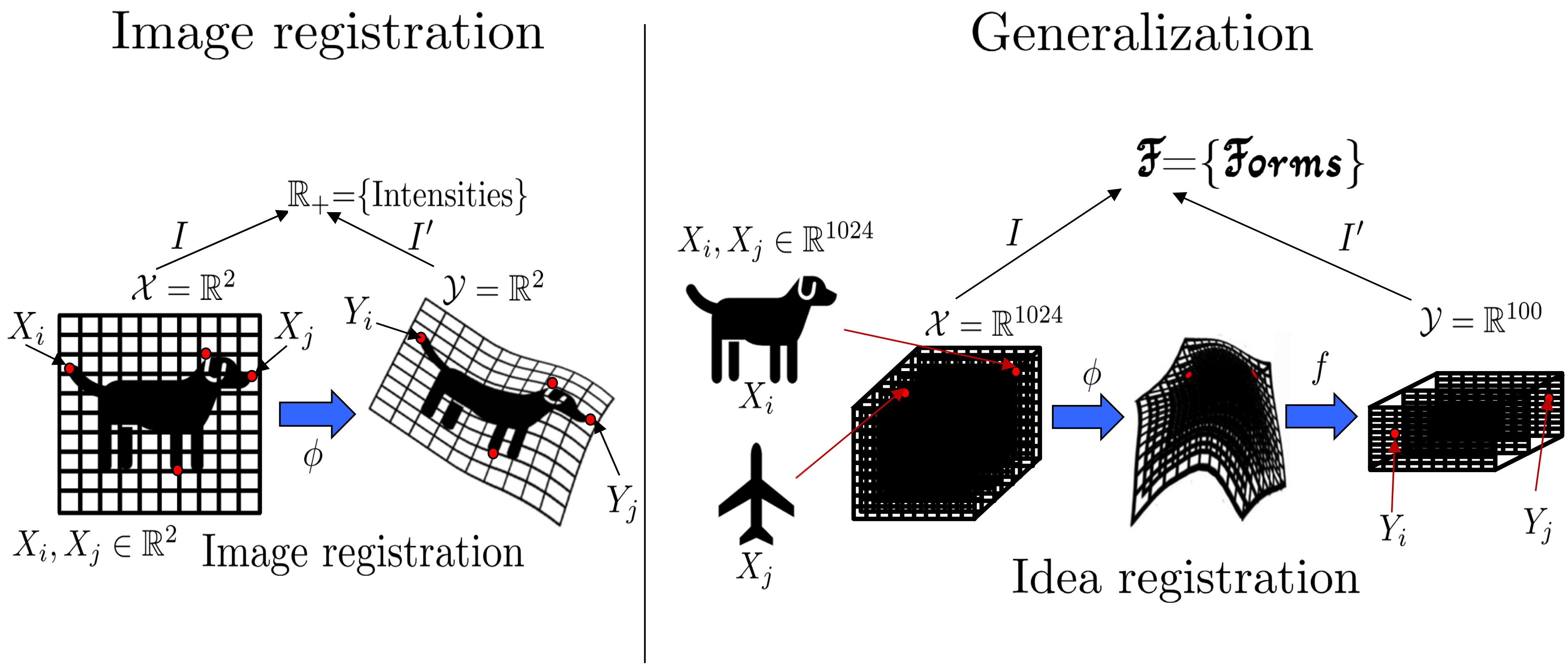


	Image registration	Idea registration
	Image $I : [0,1]^2 \to \mathbb{R}_+$ $I' : [0,1]^2 \to \mathbb{R}_+$	Idea $I : \mathcal{X} \to \mathfrak{F}$ $I' : \mathcal{Y} \to \mathfrak{F}$
X_i, Y_i	Landmark/material points $X_i \in [0,1]^2$, $Y_i \in [0,1]^2$	Data points $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$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

Idea registration is ridge regression with a warped kernel

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

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

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

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

$$\boxed{f^{RR} = f}$$

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

Idea registration is Gaussian Process Regression with a prior learned from data

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

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

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

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

$$\boxed{f^{RR} = f}$$

Theorem

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

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

Idea registration is Gaussian Process Regression with a prior learned from data

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

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

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

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

$$\boxed{f^{RR} = f}$$

Theorem

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

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

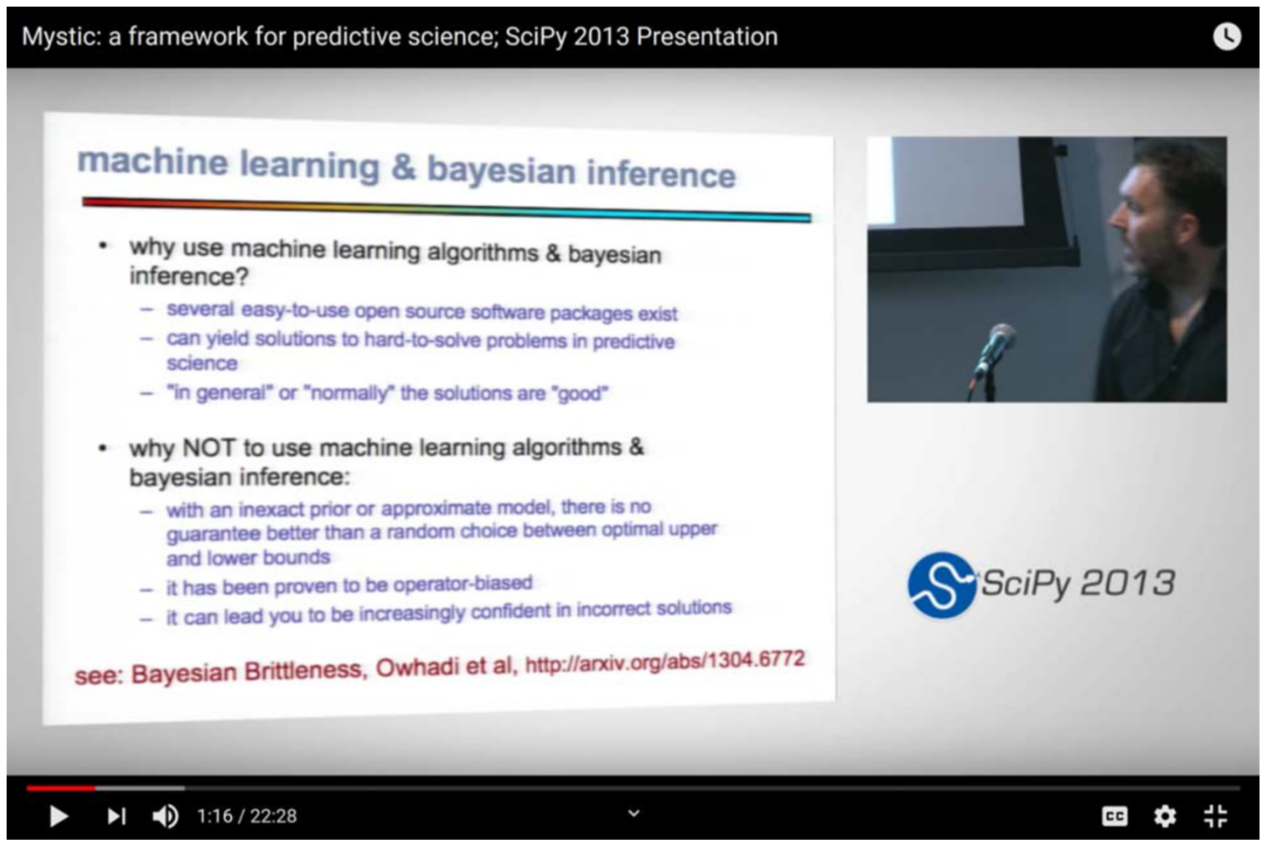
$$f^{\mathrm{IR}}(x) = \mathbb{E}_{\substack{\xi \sim \mathcal{N}(0, K^v) \\ Z \sim \mathcal{N}(0, \lambda I)}} \left[\xi(x) \mid \xi(X) = Y + Z \right]$$

[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

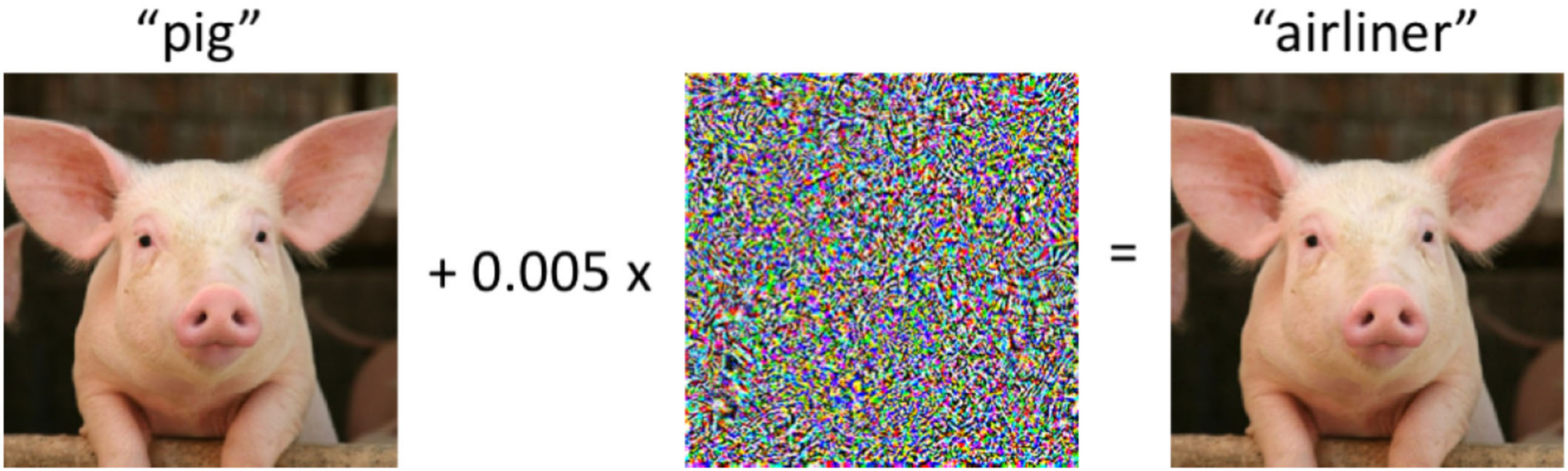
https://youtu.be/o-nwSnLC6DU?t=74

Brittleness of Bayesian inference implies the brittleness of ANNs



[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]

How do we fix it?

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

Training without regularization

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

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

Training with regularization

$$\Gamma \Longleftrightarrow \Gamma + \underbrace{rI}_{\text{nugget}}$$

$$K \Longleftrightarrow K + \overbrace{\rho I}^{\text{nugget}}$$

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

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

Kernel methods **Idea registration**

Kernel methods	Idea registration
Kernel:	Kernel representation
⇕	⇕
Feature map:	Feature map representation
⇕	⇕
RKHS space:	Idea registration
⇕	⇕
GP:	Bayesian MAP estimation

Idea registration

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

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

Theorem

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

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

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

Idea registration

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

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

Corollary

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

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

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

$\Downarrow$

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

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

In feature space

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

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

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

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

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

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

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

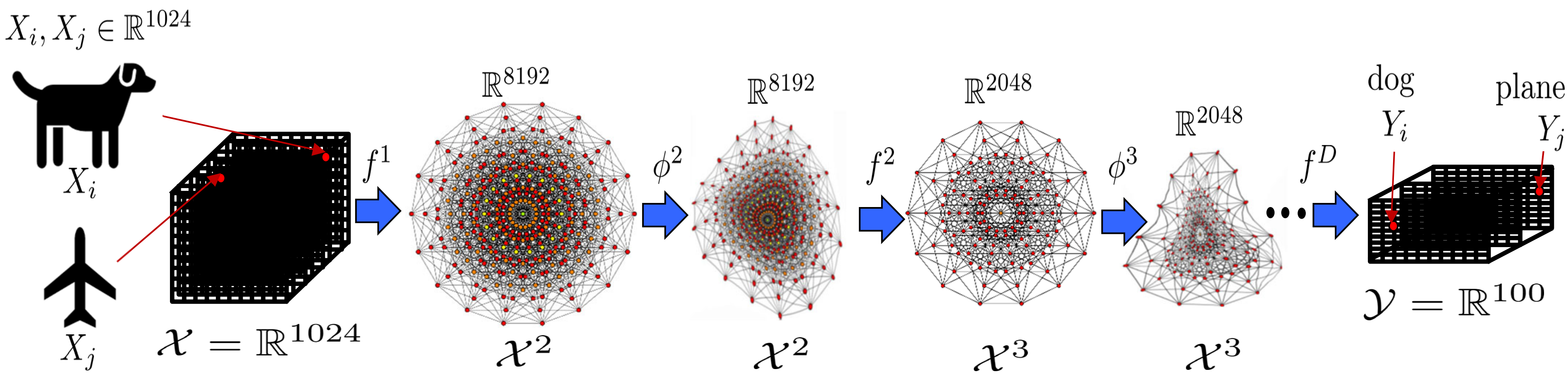
$\phi^{\zeta}(x, t)$: solution of

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

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

Composed idea registration



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

$\boxed{\text{ANNs and ResNets are solvers for discretized idea } \textit{form}\text{ation problems!}}$

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

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]

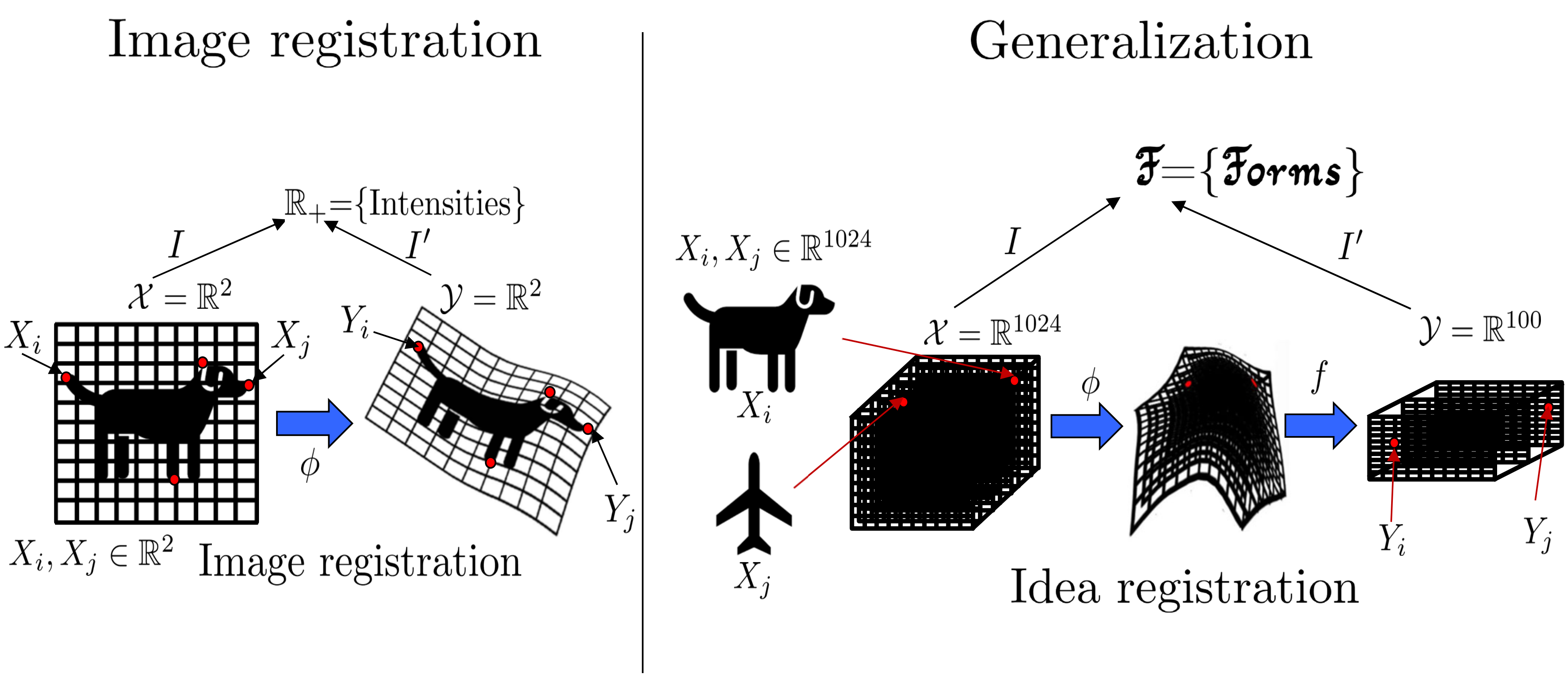
This work

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

Thank you

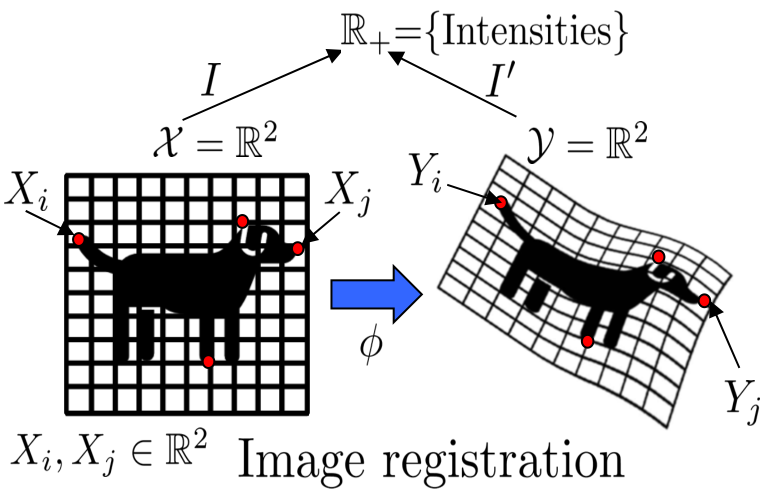
Main message

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

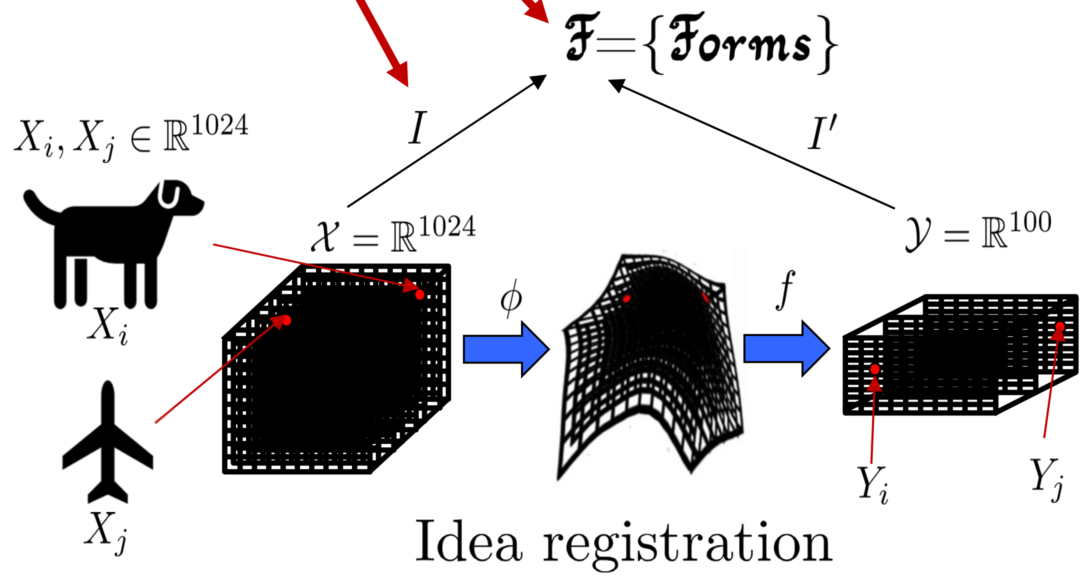


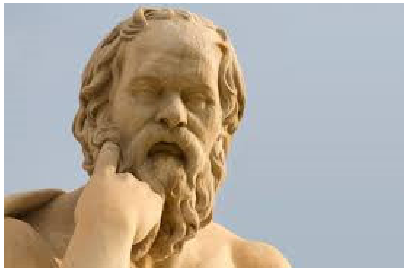
What are these?

Image registration



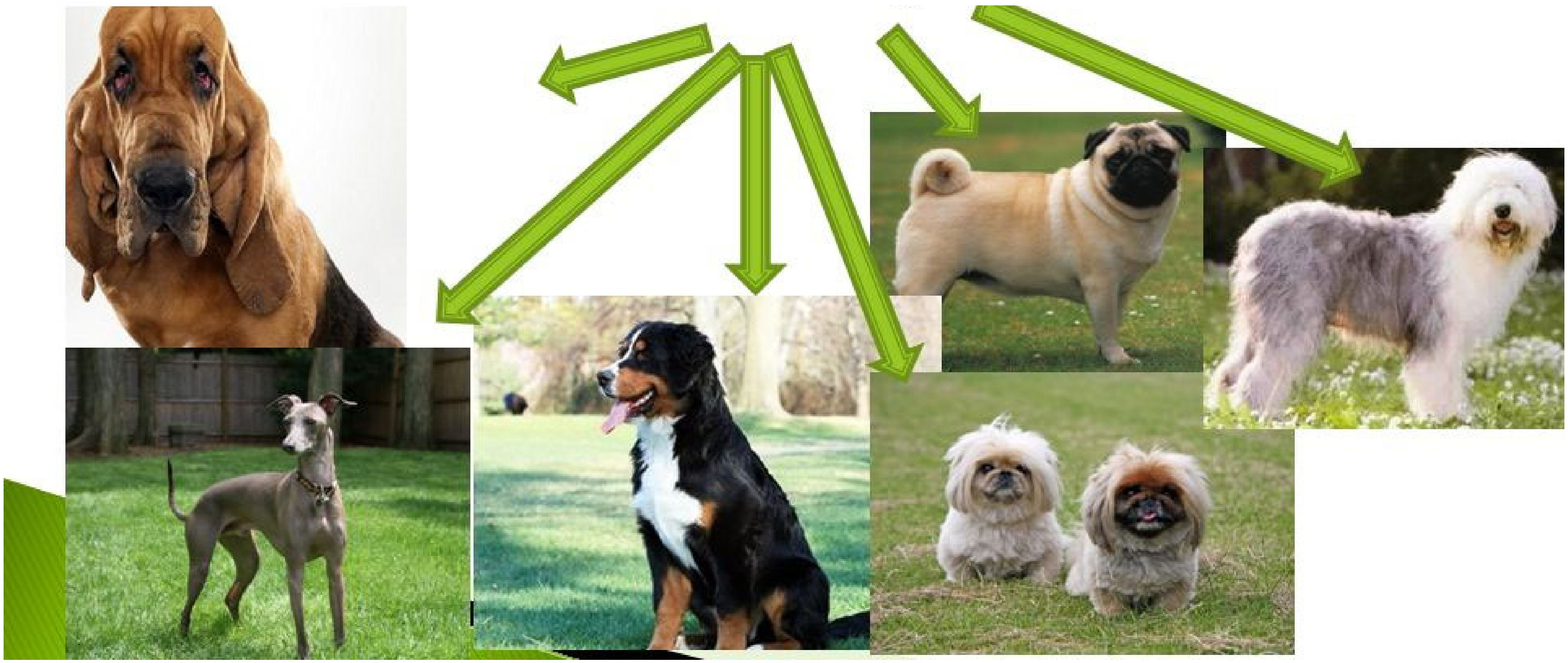
Generalization





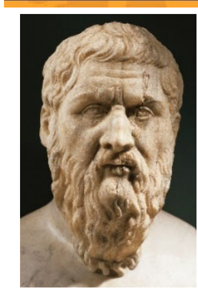
Socrates

How do we know that these are all dogs?



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

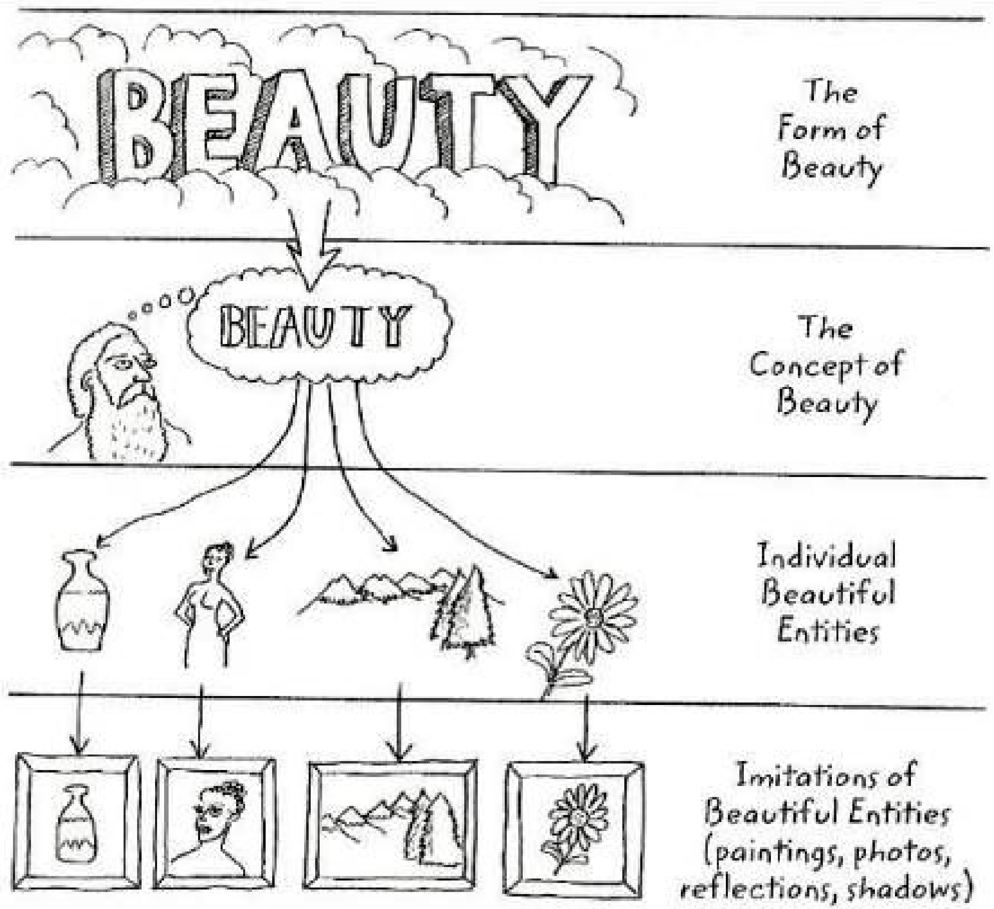
Plato's allegory of the cave



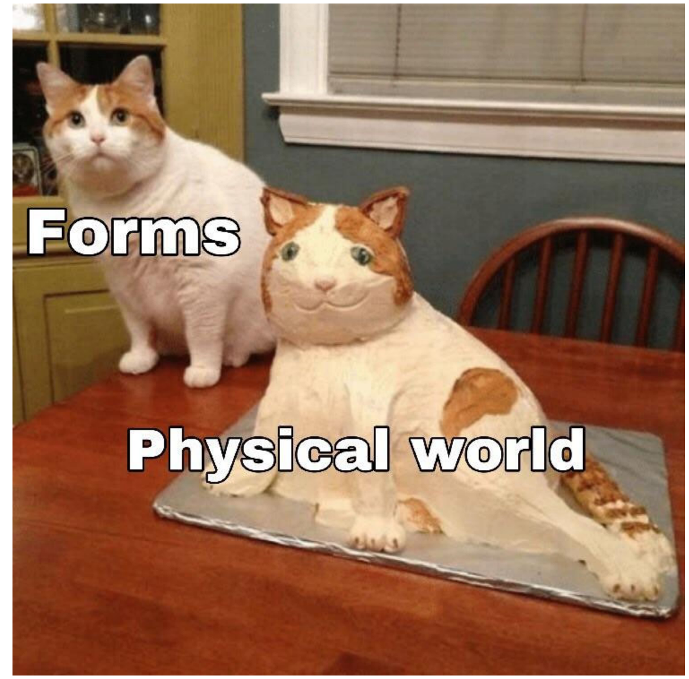
https://www.studiobinder.com/blog/platos-allegory-of-the-cave/

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

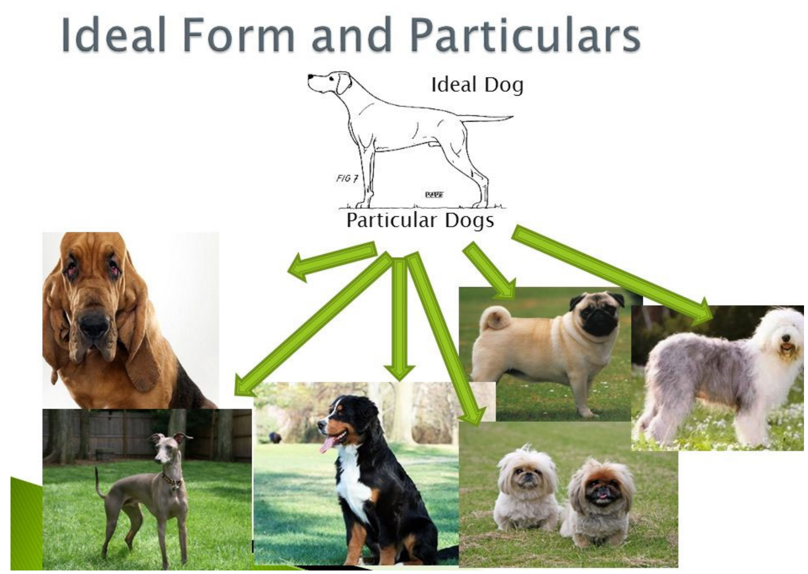
Plato's theory of forms



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`