Learning Function-Valued Functions in RKHSs with Integral Losses

PhD Defense, Alex Lambert

July 2021

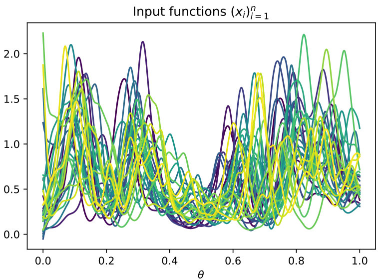
Telecom Paris

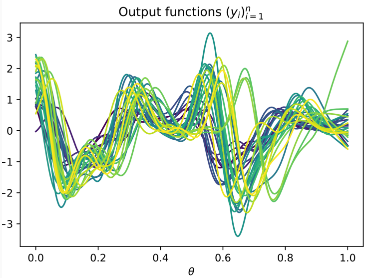
Table of contents

General Motivation

Predict Lip Acceleration from EMG Signals

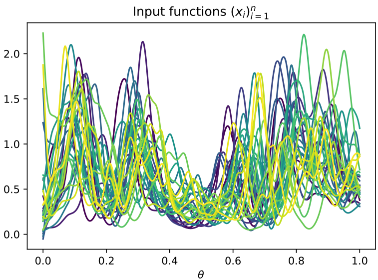
Better understand the diction mechanism

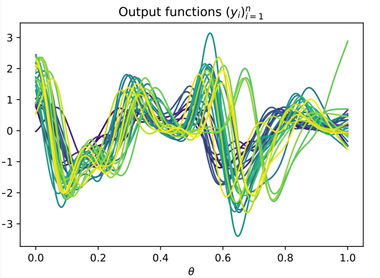




Predict Lip Acceleration from EMG Signals

Better understand the diction mechanism

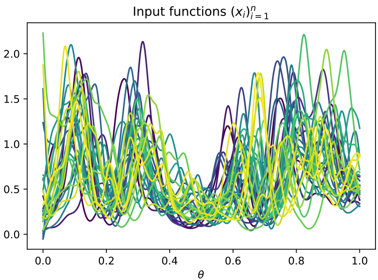


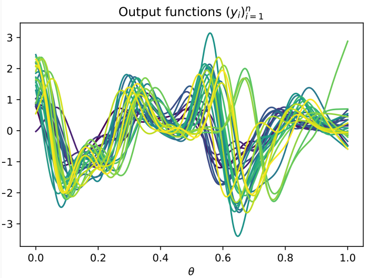


Underlying random variables (X, Y) function-valued.

Predict Lip Acceleration from EMG Signals

Better understand the diction mechanism



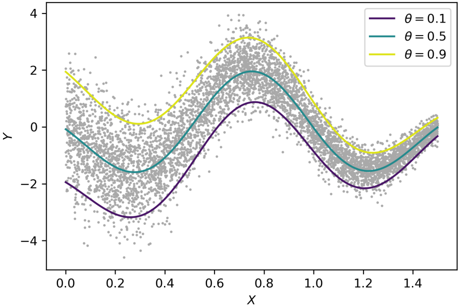


Underlying random variables (X, Y) function-valued.

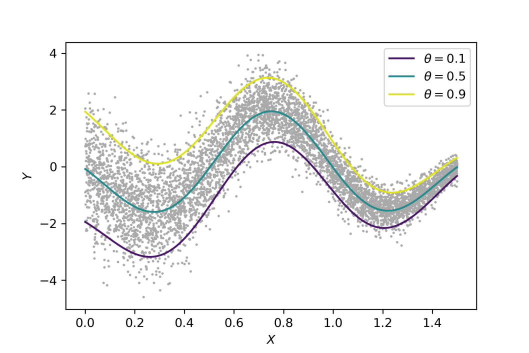
Goal

Learn a model h such that $h(X) \approx Y$

Assess Statistical Risk

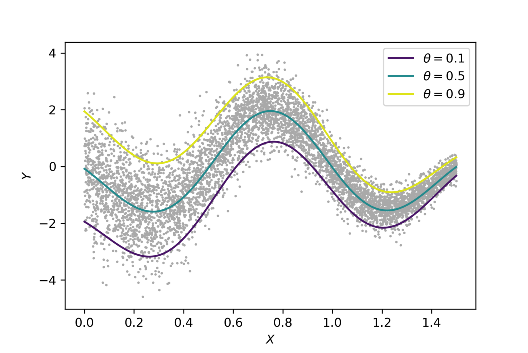


Assess Statistical Risk



Underlying random variables (X, Y) in $\mathbb{R}^d \times \mathbb{R}$

Assess Statistical Risk



Underlying random variables (X, Y) in $\mathbb{R}^d \times \mathbb{R}$

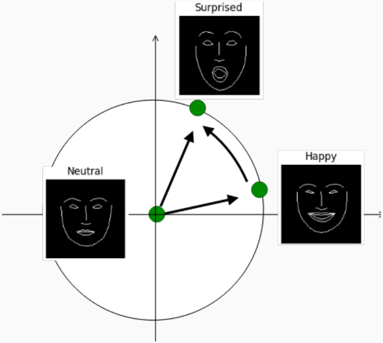
Goal

Learn a model h such that $h(x)(\theta)$ estimates the conditional θ-quantile of Y given $X = x$

Emotion Transfer for Faces

Transfering a target emotion to an input facial representation

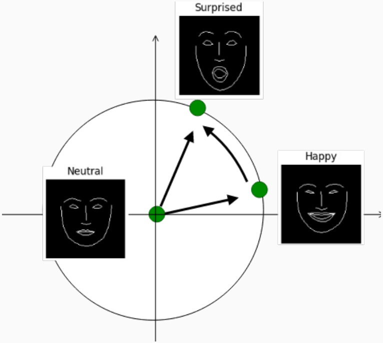
Move continuously from one emotion to another



Emotion Transfer for Faces

Transfering a target emotion to an input facial representation

Move continuously from one emotion to another



Goal

Learn a model $h\colon \mathcal{X} \to (\Theta \to \mathcal{X})$ such that $h(x)(\theta)$ transfers emotion θ to the input x

Common Ground Between These Problems

Target functions $h^*\colon \mathcal{X} \to (\Theta \to \mathbb{R}^p)$ function-valued

$$h^* \in \underset{h \text{ measurable}}{\arg\min} \underbrace{\mathbb{E}_{(\mathsf{X},\mathsf{Y})}\left[\int_\Theta \ell(\theta, h(\mathsf{X})(\theta), \mathsf{Y}(\theta))\mathrm{d}\mu(\theta)\right]}_{\mathcal{R}(h)}$$

Common Ground Between These Problems

Target functions $h^*: \mathcal{X} \to (\Theta \to \mathbb{R}^p)$ function-valued

$$h^* \in \underset{h \text{ measurable}}{\arg\min} \underbrace{\mathbb{E}_{(X,Y)}\left[\int_\Theta \ell(\theta, h(X)(\theta), Y(\theta))\mathrm{d}\mu(\theta)\right]}_{\mathcal{R}(h)}$$

- Lip acceleration prediction: θ is time, $\Theta = [0, 1]$
- Risk assessment: θ is quantile level, $\Theta = (0, 1)$
- Emotion transfer: θ encodes emotion, $\Theta = \mathcal{B}_1 \subset \mathbb{R}^2$ [Rus80] or $\subset \mathbb{R}^s$ [VA19]

Goal of this thesis

Learn function-valued functions, *i.e.* mappings

$$h : \mathcal{X} \to (\Theta \to \mathbb{R}^p)$$

Benefits:

- Regression with functional data [RS97]
- New angle to multi-task learning [EP04]
- Imposing functional constraints

Goal of this thesis

Learn function-valued functions, *i.e.* mappings

$$h\colon \mathcal{X} \to (\Theta \to \mathbb{R}^p)$$

Benefits:

- Regression with functional data [RS97]
- New angle to multi-task learning [EP04]
- Imposing functional constraints

Challenges: **Representation** and **Computability**

Vector-valued RKHSs chosen as hypothesis space [Ped57]

Modeling Function-Valued Functions

Scalar kernels and RKHSs

Scalar kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [Aro50]

- Symmetric: $k(x, x') = k(x', x)$
- Positive definite function: $\sum_{i,j \in [n]} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$

Associated RKHS $\mathcal{H}_k = \overline{\mathrm{Span}}\{k(\cdot, x) \ : \ x \in \mathcal{X}\}$
Reproducing property: $h(x) = \langle h, k(\cdot, x) \rangle_{\mathcal{H}_k}$

Scalar kernels and RKHSs

Scalar kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [Aro50]

- Symmetric: $k(x, x') = k(x', x)$
- Positive definite function: $\sum_{i,j\in[n]} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$

Associated RKHS $\mathcal{H}_k = \overline{\mathrm{Span}}\{k(\cdot, x) \ : \ x \in \mathcal{X}\}$
Reproducing property: $h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k}$

Machine learning problem with $(x_i, y_i)_{i\in[n]} \in (\mathcal{X} \times \mathbb{R})^n$:

$$\hat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}_k} \underbrace{\frac{1}{n}\sum_{i\in[n]} \ell(h(x_i), y_i)}_{\text{data fitting}} + \underbrace{\frac{\lambda}{2}\|h\|^2_{\mathcal{H}_k}}_{\text{regularization}}$$

Representer theorem [SC08]

$$\exists(\hat{\alpha}_i)_{i\in[n]} \in \mathbb{R}^n \ \ s.t. \quad \hat{h}(x) = \sum_{i\in[n]} \hat{\alpha}_i k(x, x_i)$$

Integral Operators

- Represent functions in RKHSs to handle $\langle \cdot, \cdot \rangle_{L^2[\Theta,\mu]}$

Let Θ compact, μ probability measure, k continuous

$$T_k \colon \begin{pmatrix} L^2[\Theta,\mu] & \to & L^2[\Theta,\mu] \\ f & \mapsto & \left(\theta \mapsto \int_\Theta f(\theta')k(\theta,\theta')\mathrm{d}\mu(\theta')\right) \end{pmatrix}$$

Spectral decomposition:

$$\forall f \in L^2[\Theta,\mu], \quad T_k f = \sum_{j=1}^{\infty} \lambda_j \langle f, \psi_j \rangle \psi_j$$

- $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0$ nonnegative eigenvalues
- $(\psi_j)_{j=1}^{\infty}$ orthonormal system in $L^2[\Theta,\mu]$

Truncated basis with first m eigenvectors

$$h \approx \sum_{j=1}^{m} \beta_j \psi_j$$

Operator-Valued Kernels and vv-RKHSs

VV-RKHS framework [CDT06]:

- Hilbert space of functions with values in a Hilbert space $\mathcal{Y}$
- Associated to an operator-valued kernel acting on $\mathcal{Y}$
- Drives regularization

VV-RKHS framework [CDT06]:

- Hilbert space of functions with values in a Hilbert space $\mathcal{Y}$
- Associated to an operator-valued kernel acting on $\mathcal{Y}$
- Drives regularization

Scalar-valued kernel	Operator-valued kernel
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	$K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$
$k(x, x') = k(x', x)$	$K(x, x') = K(x', x)^{\#}$
$\sum_{i,j \in [n]} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$	$\sum_{i,j \in [n]} \langle K(x_i, x_j) y_i, y_j \rangle_{\mathcal{Y}} \geqslant 0$
	$K_x : y \in \mathcal{Y} \mapsto (x' \mapsto K(x', x) y)$
$\mathcal{H}_k = \overline{\mathrm{Span}}\{k(\cdot, x) \ : \ x \in \mathcal{X}\}$	$\mathcal{H}_K = \overline{\mathrm{Span}}\{K_x y \ : \ (x, y) \in \mathcal{X} \times \mathcal{Y}\}$
$h(x) = \langle h, k(\cdot, x) \rangle_{\mathcal{H}_k} \in \mathbb{R}$	$h(x) = K_x^{\#} h \in \mathcal{Y}$

Regularized Empirical Risk Minimization (RERM) in vv-RKHSs

Family of problems:

- Data $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ i.i.d.
- Convex loss $L \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

$$\hat{h} := \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(h(x_i), y_i) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

Regularized Empirical Risk Minimization (RERM) in vv-RKHSs

Family of problems:

- Data $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ i.i.d.
- Convex loss $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

$$\hat{h} := \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(h(x_i), y_i) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

Representer theorem [MP05]

$$\exists (\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n \;\; s.t. \quad \hat{h}(x) = \sum_{i \in [n]} K(x, x_i)\hat{\alpha}_i$$

Challenge: $\mathcal{Y}$ infinite-dimensional

Parametric Duality for RERM in vv-RKHSs

Parametric duality for convex optimization [Roc70]

Fenchel-Legendre conjugate of a function $f \colon \mathcal{Y} \to \mathbb{R}$:

$$f^\star(y) := \sup_{y' \in \mathcal{Y}} \langle y, y' \rangle_{\mathcal{Y}} - f(y')$$

Notation: $L_i : y \mapsto L(y, y_i)$

Dual optimization problem [BSD16]

It holds that $\hat{h}(x) = \frac{1}{\lambda n} \sum_{i \in [n]} K(x, x_i) \hat{\alpha}_i$, where

$$(\hat{\alpha}_i)_{i=1}^n = \underset{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n}{\arg\min} \sum_{i \in [n]} L_i^\star(-\alpha_i) + \frac{1}{2\lambda n} \sum_{i,j \in [n]} \left\langle \alpha_i, K(x_i, x_j) \alpha_j \right\rangle_{\mathcal{Y}}$$

Challenge: $\mathcal{Y}$ infinite-dimensional

Modeling Function-Valued Functions in vv-RKHSs

Our use case: $\mathcal{Y}$ space of functions $\Theta \to \mathbb{R}^p$

Simplest case $p = 1$, combine two scalar kernels

$$k_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad\qquad k_{\Theta} \colon \Theta \times \Theta \to \mathbb{R}$$

Modeling Function-Valued Functions in vv-RKHSs

Our use case: $\mathcal{Y}$ space of functions $\Theta \to \mathbb{R}^p$

Simplest case $p = 1$, combine two scalar kernels

$$k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad\qquad k_{\Theta} : \Theta \times \Theta \to \mathbb{R}$$

View 1: $\mathcal{Y} = \mathcal{H}_{k_\Theta}$

$$K(x, x') = k_{\mathcal{X}}(x, x') \mathbf{Id}_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$$

Modeling Function-Valued Functions in vv-RKHSs

Our use case: $\mathcal{Y}$ space of functions $\Theta \to \mathbb{R}^p$

Simplest case $p = 1$, combine two scalar kernels

$$k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad\qquad k_\Theta : \Theta \times \Theta \to \mathbb{R}$$

View 1: $\mathcal{Y} = \mathcal{H}_{k_\Theta}$

$$K(x, x') = k_{\mathcal{X}}(x, x')\mathbf{Id}_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$$

View 2: $\mathcal{Y} = L^2[\Theta, \mu]$

$$K(x, x') = k_{\mathcal{X}}(x, x')T_{k_\Theta} \in \mathcal{L}(L^2[\Theta, \mu])$$

Modeling Function-Valued Functions in vv-RKHSs

Our use case: $\mathcal{Y}$ space of functions $\Theta \to \mathbb{R}^p$

Simplest case $p = 1$, combine two scalar kernels

$$k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad k_{\Theta} : \Theta \times \Theta \to \mathbb{R}$$

View 1: $\mathcal{Y} = \mathcal{H}_{k_\Theta}$

$$K(x, x') = k_{\mathcal{X}}(x, x') \mathbf{Id}_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$$

View 2: $\mathcal{Y} = L^2[\Theta, \mu]$

$$K(x, x') = k_{\mathcal{X}}(x, x') T_{k_\Theta} \in \mathcal{L}(L^2[\Theta, \mu])$$

Structure: $\mathcal{H}_K \simeq \mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_\Theta} \simeq \mathcal{H}_{k_{\mathcal{X}} \otimes k_\Theta}$

Same space of functions up to a reparameterization

Optimization Challenges

Goal

Find the coefficients $\hat{\alpha}_i$

Two main challenges:

Representation

- $\hat{\alpha}_i$ function of θ
- $\mathcal{Y}$ infinite-dimensional, either $\mathcal{H}_{k_\Theta}$ or $L^2[\Theta, \mu]$

Computability

- $L(f, g) = \int_\Theta \ell(\theta, f(\theta), g(\theta)) \mathrm{d}\mu(\theta)$
- $L_i^\star(-\alpha_i)$ involves $\mathcal{Y}$
- $K(x_i, x_j)\alpha_i = k_\mathcal{X}(x_i, x_j) T_{k_\Theta} \alpha_i$

Proposed Solutions

In this thesis

Global study of primal and dual methods

type	$\mathcal{Y}$	parameterization	loss	algorithm
closed form	$L^2[\Theta, \mu]$	eigenbasis of T_{k_Θ}	square loss	analytic
• closed form	$\mathcal{H}_{k_\Theta}$	double representer	square loss	analytic
• primal	$\mathcal{H}_{k_\Theta}$	double representer	sampled	GD
primal	$\mathcal{H}_{k_\Theta}$	ORFF	any	SGD
• dual	$L^2[\Theta, \mu]$	eigenbasis of T_{k_Θ}	compatibility loss/T_{k_Θ}	GD
dual	$L^2[\Theta, \mu]$	linear splines	prox computable	PGD

Proposed Solutions

In this thesis

Global study of primal and dual methods

type	$\mathcal{Y}$	parameterization	loss	algorithm
closed form	$L^2[\Theta, \mu]$	eigenbasis of T_{k_Θ}	square loss	analytic
• closed form	$\mathcal{H}_{k_\Theta}$	double representer	square loss	analytic
• primal	$\mathcal{H}_{k_\Theta}$	double representer	sampled	GD
primal	$\mathcal{H}_{k_\Theta}$	ORFF	any	SGD
• dual	$L^2[\Theta, \mu]$	eigenbasis of T_{k_Θ}	compatibility loss/T_{k_Θ}	GD
dual	$L^2[\Theta, \mu]$	linear splines	prox computable	PGD

Today:

- Primal with view 1 for infinite task learning
- Dual with view 2 for robust functional output regression
- Closed form with view 1 for emotion transfer

Infinite Task Learning

Jointly Learning Many Tasks

Extending multi-task learning [EP04]

Jointly Learning Many Tasks

Extending multi-task learning [EP04]

Learning tasks with a free parameter θ:

- Quantile regression [KB78] (θ quantile level)
- Cost-sensitive classification [ZE01] (θ imbalanced coefficient)
- One-class SVM [Sch+01] (θ proportion of outliers)

Jointly Learning Many Tasks

Extending multi-task learning [EP04]

Learning tasks with a free parameter θ:

- Quantile regression [KB78] (θ quantile level)
- Cost-sensitive classification [ZE01] (θ imbalanced coefficient)
- One-class SVM [Sch+01] (θ proportion of outliers)

Goal

Jointly learn these tasks for a continuum of θ [Tak+13]

Jointly Learning Many Tasks

Extending multi-task learning [EP04]

Learning tasks with a free parameter θ:

- Quantile regression [KB78] (θ quantile level)
- Cost-sensitive classification [ZE01] (θ imbalanced coefficient)
- One-class SVM [Sch+01] (θ proportion of outliers)

Goal

Jointly learn these tasks for a continuum of θ [Tak+13]

Multi-task learning	Infinite-task learning
Finite number of $(\theta_j)_{j=1}^{p}$	Infinite number of θ
$\mathbb{R}^p$-valued model	function-valued model
sum of loss functions	$\int$ of loss functions

Conditional Quantiles

Conditional quantile:
Take (X, Y) random variables in $\mathbb{R}^d \times \mathbb{R}$, $(X, Y) \sim \mathbb{P}_{(X,Y)}$

$$q(x)(\theta) := \inf\{t \in \mathbb{R} \,|\, \mathbb{P}(Y \leqslant t \,|\, X = x) \geqslant \theta\}, \quad \theta \in (0, 1)$$

Shape:
$q(x)$ increasing function of θ

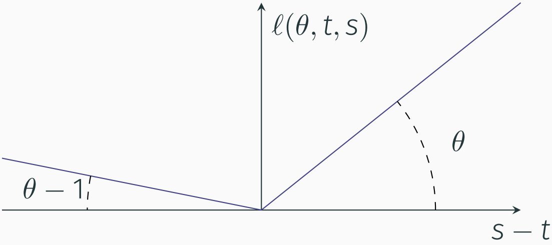
Pinball Loss

Variational formula:

$$q(x)(\theta) \in \arg\min_{t\in\mathbb{R}} \mathbb{E}\left[\ell(\theta,t,Y)|X=x\right]$$

where $\ell(\theta,\cdot,\cdot)$ is the pinball loss [KB78]:

$$\ell(\theta,t,s) = \max\left(\theta(s-t),(\theta-1)(s-t)\right)$$



Pinball loss for $\theta = 0.8$

Problem Formulation

Task at level $\theta \in \Theta$

$$\min_{h \text{ measurable}} \mathbb{E}_{(X,Y)} \left[\ell(\theta, h(X), Y)\right]$$

described by $\ell \colon \Theta \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Given $\mathcal{S} := (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathbb{R})^n$ i.i.d. following $\mathbb{P}_{X,Y}$

Optimization problem

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(h(x_i), y_i) + \frac{\lambda}{2} \|h\|_{\mathcal{H}_K}^2$$

- $L(h(x), y) = \int_\Theta \ell(\theta, h(x)(\theta), y) \mathrm{d}\mu(\theta)$
- μ encodes importance of tasks
- $\mathcal{R}_\mathcal{S}(h) := \frac{1}{n} \sum_{i \in [n]} L(h(x_i), y_i)$ empirical risk
- View 1: $K = k_\mathcal{X} \, \mathbf{Id}_{\mathcal{H}_{k_\Theta}}$

Sampled Empirical Risk

Representer theorem [MP05]:

$$\exists (\hat{\alpha}_i)_{i=1}^n \in \mathcal{H}_{k_\Theta}^n \quad , \quad \hat{h} = \sum_{i \in [n]} K(\cdot, x_i) \hat{\alpha}_i$$

Not enough: representation, computability

Sampled Empirical Risk

Representer theorem [MP05]:

$$\exists (\hat{\alpha}_i)_{i=1}^n \in \mathcal{H}_{k_\Theta}^n \quad , \quad \hat{h} = \sum_{i \in [n]} K(\cdot, x_i)\hat{\alpha}_i$$

Not enough: representation, computability

Sampled empirical risk

$$\widetilde{\mathcal{R}}_{\mathcal{S}}(h) := \frac{1}{n} \sum_{i \in [n]} \sum_{j \in [m]} \eta_{ij} \ell(\theta_{ij}, h(x_i)(\theta_{ij}), y_i)$$

- Monte-Carlo: $\eta_{ij} = \frac{1}{m}$, $(\theta_{ij})_{j=1}^m \overset{\text{i.i.d.}}{\sim} \mu$
- Quasi Monte-Carlo: $\eta_{ij} = \frac{1}{m}$, θ_{ij} low discrepancy (Sobol)
- Kernel quadrature rules, ...

Double Representer Theorem

Approximated problem

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \widetilde{\mathcal{R}}_{\mathcal{S}}(h) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

Double representer theorem (chapter 3)

$$\hat{h}(x)(\theta) = \sum_{i \in [n]} \sum_{j \in [m]} \hat{\alpha}_{ij} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_{ij}), \quad \hat{\alpha}_{ij} \in \mathbb{R}$$

Idea: reproducing property in both $\mathcal{H}_{k_{\mathcal{X}}}$ and $\mathcal{H}_{k_{\Theta}}$

- Finite parameterization of the solution $\in \mathbb{R}^{n \times m}$
- Computable loss
- Plug-in prefered solver depending on ℓ

Goal: bound with high probability

$$\mathcal{R}(\hat{h}) - \mathcal{R}_{\mathcal{S}}(\hat{h})$$

Generalization Bounds for Quantile Regression

Goal: bound with high probability

$$\mathcal{R}(\hat{h}) - \mathcal{R}_{\mathcal{S}}(\hat{h})$$

Framework of uniform stability [BE02]

- Suitable to vv-RKHSs [Kad+16]
- Trade $\mathcal{R}_{\mathcal{S}}(\hat{h})$ against $\widetilde{\mathcal{R}}_{\mathcal{S}}(\hat{h})$

Generalization Bounds for Quantile Regression

Goal: bound with high probability

$$\mathcal{R}(\hat{h}) - \mathcal{R}_{\mathcal{S}}(\hat{h})$$

Framework of uniform stability [BE02]

- Suitable to vv-RKHSs [Kad+16]
- Trade $\mathcal{R}_{\mathcal{S}}(\hat{h})$ against $\widetilde{\mathcal{R}}_{\mathcal{S}}(\hat{h})$

Generalization bound for QR with QMC approximation (chapter 5)

$$\mathcal{R}(\hat{h}) \leqslant \widetilde{\mathcal{R}}_{\mathcal{S}}(\hat{h}) + \mathcal{O}_{\mathbb{P}_{X,Y}}\left(\frac{1}{\lambda\sqrt{n}}\right) + \mathcal{O}\left(\frac{\log m}{\sqrt{\lambda} m}\right)$$

- Requires bounded $k_{\mathcal{X}}, k_{\Theta}, Y$
- Choosing $m \approx \sqrt{\lambda n}$

Example of functional constraint: $\forall x, q(x)$ is nondecreasing
Add soft constraint to encourage non crossing quantiles:

$$\Omega_{\mathrm{nc}}(h) := \lambda_{\mathrm{nc}} \int_{\mathcal{X}} \int_{\Theta} \left|-(\partial_{\Theta} h)(x)(\theta)\right|_{+} \mathrm{d}\mathbb{P}_{X}(x) \mathrm{d}\mu(\theta)$$

Approximated as

$$\widetilde{\Omega}_{\mathrm{nc}}(h) := \lambda_{\mathrm{nc}} \frac{1}{n_{\mathrm{nc}} m_{\mathrm{nc}}} \sum_{i \in n_{\mathrm{nc}}} \sum_{j \in m_{\mathrm{nc}}} \left|-(\partial_{\Theta} h)(\tilde{x}_i)(\tilde{\theta}_j)\right|_{+}.$$

- Grid $(\tilde{x}_i)_{i \in n_{\mathrm{nc}}}$, $(\tilde{\theta}_j)_{j \in m_{\mathrm{nc}}}$
- Hyperparameter $\lambda_{\mathrm{nc}} > 0$

Handling Shape Constraints

Problem to solve:

$$\hat{h} = \underset{h \in \mathcal{H}_K}{\arg\min}\ \widetilde{\mathcal{R}}_{\mathcal{S}}(h) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K} + \widetilde{\Omega}_{\mathrm{nc}}(h)$$

Working with RKHSs -> access to derivatives [Zho08]

Double representer theorem with derivatives (chapter 5)

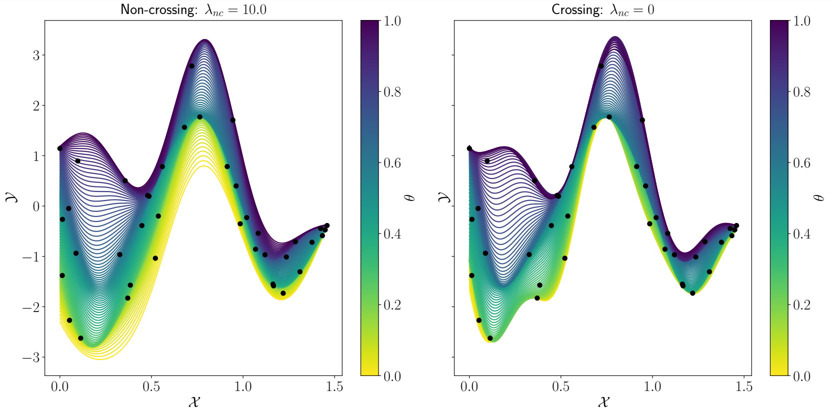
$\exists (\hat{\alpha}_{ij})_{i,j \in [n]\times[m]} \in \mathbb{R}^{nm}$ and $(\hat{\beta}_{ij})_{i,j \in [n_{\mathrm{nc}}]\times[m_{\mathrm{nc}}]} \in \mathbb{R}^{n_{\mathrm{nc}} m_{\mathrm{nc}}}$ s.t.

$$\hat{h}(x)(\theta) = \sum_{i,j \in [n]\times[m]} \hat{\alpha}_{ij} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_j) + \sum_{i,j \in [n_{\mathrm{nc}}]\times[m_{\mathrm{nc}}]} \hat{\beta}_{ij} k_{\mathcal{X}}(x, \tilde{x}_i) \partial_2 k_{\Theta}(\theta, \tilde{\theta}_j)$$

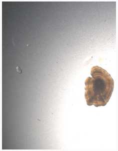
- Finite dimensional representation
- Price to pay: tune λ_{nc}, modify loss

Numerical Illustration

Small data regime prone to crossing quantiles ($n = 40$)

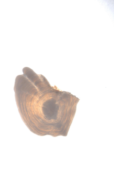


Managing fish resources: estimate age of fishes using otholiths pictures [Ord+20]



Age 3

Age 7

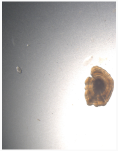


Age 10

Age 13

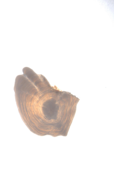
Deep Kernel Models for QR

Managing fish resources: estimate age of fishes using otholiths pictures [Ord+20]



Age 3

Age 7



Age 10

Age 13

X is image, Y is age of the fish

Using a deep kernel [Yan+15; MZS17]

$$k_{\mathcal{X}}(x, x') = k_{\mathcal{V}}(\phi_\omega(x), \phi_\omega(x'))$$

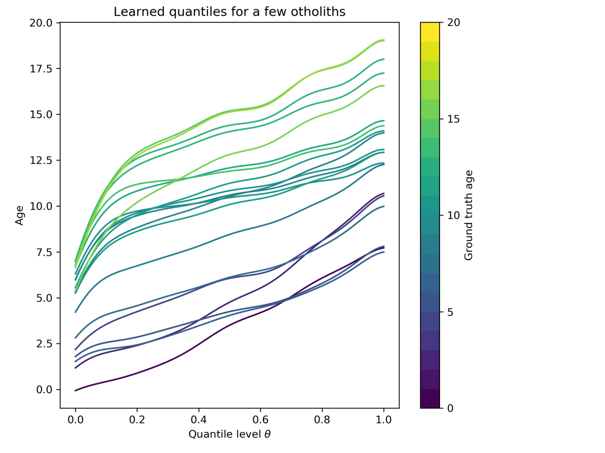
- $\phi_\omega : \mathcal{X} \to \mathcal{V}$ neural architecture (Inception v3, [Sze+16])
- $k_{\mathcal{V}}$ kernel on the feature space

Numerical Illustration

Use of Random Fourier Features [RR07] for $k_{\mathcal{Y}}$ and k_{Θ}

-> Finite dimensional representation by design

Joint optimization on α (kernel) and ω (neural)



Functional Output Regression: Beyond the Square Loss

Functional Output Regression

Data $(x_i, y_i)_{i=1}^n$ i.i.d. realisations of (X, Y). Response variable Y is a function: $y_i \in L^2[\Theta, \mu]$

Regularized empirical risk minimization in vv-RKHS:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(y_i - h(x_i)) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

- $L(f) = \frac{1}{2} \int_\Theta f^2(\theta) \mathrm{d}\theta$ closed-form (ridge regression) with view 1 [Lia07], view 2 [Kad+16]

Functional Output Regression

Data $(x_i, y_i)_{i=1}^n$ i.i.d. realisations of (X, Y). Response variable Y is a function: $y_i \in L^2[\Theta, \mu]$

Regularized empirical risk minimization in vv-RKHS:

$$\hat{h} \in \underset{h \in \mathcal{H}_K}{\arg\min} \; \frac{1}{n} \sum_{i \in [n]} L(y_i - h(x_i)) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

- $L(f) = \frac{1}{2} \int_\Theta f^2(\theta) \mathrm{d}\theta$ closed-form (ridge regression) with view 1 [Lia07], view 2 [Kad+16]

Goal

Enforce robustness or sparsity for $\hat{h}$ through L

Exploit duality: use $\mathcal{Y} = L^2[\Theta, \mu]$, $K = k_{\mathcal{X}} T_{k_\Theta}$

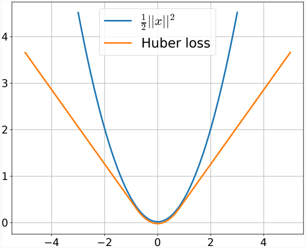
The Huber Loss

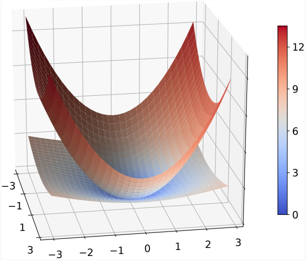
Combine $f, g \colon \mathcal{Y} \to \mathbb{R}$ through infimal convolution [BC+11]

$$f \,\square\, g(y) = \inf_{y' \in \mathcal{Y}} f(y - y') + g(y')$$

Huber loss of parameter $\kappa > 0$:

$$L = \frac{1}{2} \|\cdot\|_{\mathcal{Y}}^2 \,\square\, \kappa \|\cdot\|_{\mathcal{Y}}$$





- Asymptotics as $\kappa \|\cdot\|_{\mathcal{Y}}$ instead of $\|\cdot\|_{\mathcal{Y}}^2$

A Dual Approach

Dual problem:

$$(\hat{\alpha}_i)_{i=1}^n = \underset{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n}{\arg\min} \sum_{i\in[n]} L^\star(-\alpha_i) - \langle \alpha_i, y_i \rangle_{\mathcal{Y}} + \frac{1}{2\lambda n} \sum_{i,j\in[n]} \left\langle \alpha_i, K(x_i, x_j)\alpha_j \right\rangle_{\mathcal{Y}}$$

Infimal convolution and duality:

$$L^\star = \left(\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, \kappa\|\cdot\|_{\mathcal{Y}}\right)^\star = \underbrace{\left(\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2\right)^\star}_{\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2} + \underbrace{\left(\kappa\|\cdot\|_{\mathcal{Y}}\right)^\star}_{\chi_{\mathcal{B}_\kappa}(\cdot)}$$

where χ indicator function, $\mathcal{B}_\kappa$ ball of radius κ

Learning with Huber Loss

Dual problem, Huber loss (chapter 4)

$$\inf_{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n} \sum_{i \in [n]} \frac{1}{2} \|\alpha_i\|_{\mathcal{Y}}^2 - \langle \alpha_i, y_i \rangle_{\mathcal{Y}} + \frac{1}{2\lambda n} \sum_{i,j \in [n]} k_{\mathcal{X}}(x_i, x_j) \left\langle \alpha_i, T_{k_\Theta} \alpha_j \right\rangle_{\mathcal{Y}}$$
$$s.t. \;\; \forall i \in [n], \;\; \|\alpha_i\|_{\mathcal{Y}} \leqslant \kappa$$

Challenges: compute $\left\langle \alpha_i, T_{k_\Theta} \alpha_j \right\rangle_{\mathcal{Y}}$, handle the constraints

Learning with Huber Loss

Dual problem, Huber loss (chapter 4)

$$\inf_{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n} \sum_{i\in[n]} \frac{1}{2} \|\alpha_i\|_{\mathcal{Y}}^2 - \langle \alpha_i, y_i \rangle_{\mathcal{Y}} + \frac{1}{2\lambda n} \sum_{i,j\in[n]} k_{\mathcal{X}}(x_i, x_j) \left\langle \alpha_i, T_{k_\Theta} \alpha_j \right\rangle_{\mathcal{Y}}$$
$$s.t.\ \ \forall i \in [n],\ \|\alpha_i\|_{\mathcal{Y}} \leqslant \kappa$$

Challenges: compute $\left\langle \alpha_i, T_{k_\Theta} \alpha_j \right\rangle_{\mathcal{Y}}$, handle the constraints

Represent the $(\alpha_i)_{i\in[n]}$ using eigenbasis $(\psi_j)_{j\in[m]}$ of T_{k_Θ}

$$\underbrace{\alpha_i}_{\in \mathcal{Y}} = \sum_{j\in[m]} \underbrace{\alpha_{ij}}_{\in \mathbb{R}} \underbrace{\psi_j}_{\in \mathcal{Y}}$$

- Finite dimensional parameterization by $\boldsymbol{\alpha} \in \mathbb{R}^{nm}$

Learning with Huber Loss

Notation

- Gram matrix $\mathbf{K}_{\mathcal{X}} = [k_{\mathcal{X}}(x_i, x_j)]_{i,j \in [n] \times [m]} \in \mathbb{R}^{n \times n}$
- Eigenvalues matrix $\Lambda = \operatorname{diag}\{(\lambda_j)_{j \in [m]}\} \in \mathbb{R}^{m \times m}$
- Data-fitting term $\mathbf{R} = [\langle y_i, \psi_j \rangle_{\mathcal{Y}}]_{i,j \in [n] \times [m]} \in \mathbb{R}^{n \times m}$
- $\|\cdot\|_{2,\infty}$: maximum of row-wise $\|\cdot\|_2$

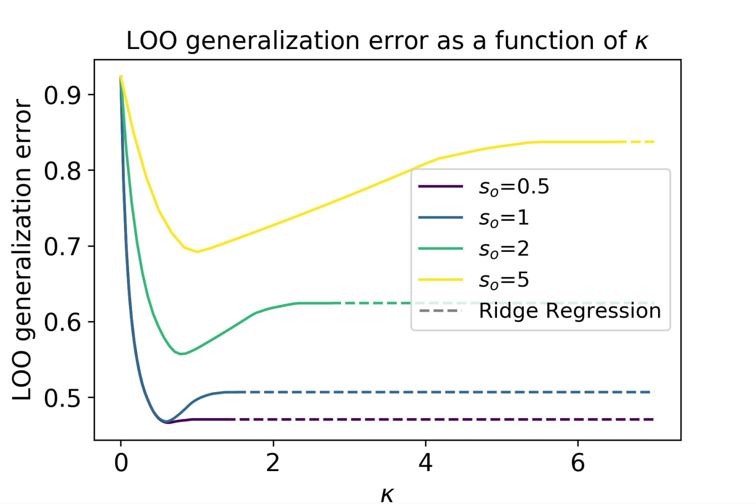
Practical optimization problem (chapter 4)

$$\inf_{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}} \quad \operatorname{Tr}\left(\frac{1}{2}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \boldsymbol{\alpha}\mathbf{R}^\top + \frac{1}{2\lambda n}\mathbf{K}_{\mathcal{X}}\boldsymbol{\alpha}\Lambda\boldsymbol{\alpha}^\top\right)$$
$$s.t. \quad \|\boldsymbol{\alpha}\|_{2,\infty} \leqslant \kappa.$$

- Solvable using projected gradient descent
- When κ is large, recover ridge regression [Kad+16]

Lip dataset, augmented with outliers.

s_o: scale of the outlier



Emotion Transfer for Facial Landmarks

Problem Formulation

- Facial representation space $\mathcal{X} = \mathbb{R}^d$
- Emotion embedding space $\Theta \subset \mathbb{R}^s$

Goal

Learn a model $h \colon \mathcal{X} \to (\Theta \to \mathcal{X})$ such that $h(x)(\theta)$ transfers emotion θ to the input x

Given $(\underbrace{x_i}_{\in \mathcal{X}}, \underbrace{(y_{ij})_{j\in[m]}}_{\in \mathcal{X}^m})_{i\in[n]}$ observed at emotions $(\theta_{ij})_{i,j\in[n]\times[m]}$

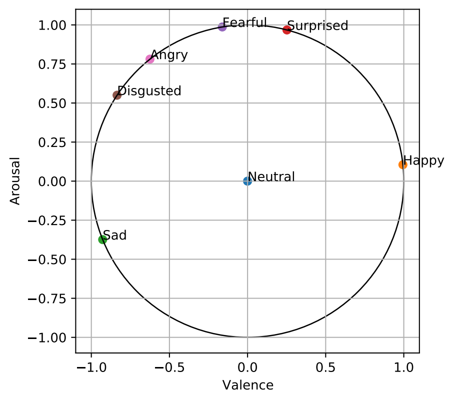
Empirical risk

$$\mathcal{R}_{\mathcal{S}}(h) = \frac{1}{nm} \sum_{i\in[n]} \sum_{j\in[m]} \left\| h(x_i)(\theta_{ij}) - y_{ij} \right\|^2_{\mathbb{R}^d}$$

Emotion Encoding

Pre-defined ℓ_2 normalized embedding in valence-arousal space [Rus80]

Centroids from AffectNet database [Kol+19]



Modeling with view 1

$$h : \mathcal{X} \mapsto \underbrace{\underbrace{(\Theta \mapsto \mathcal{X})}_{\in \mathcal{H}_G}}_{\in \mathcal{H}_K}$$

- Scalar kernel k_Θ
- Scalar kernel $k_\mathcal{X}$
- Positive self-adjoint matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ encoding output similarities

$$G(\theta, \theta') = k_\Theta(\theta, \theta')\mathbf{A}, \qquad K(x, x') = k_\mathcal{X}(x, x')\mathbf{Id}_{\mathcal{H}_G}$$

Optimization problem

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \mathcal{R}_{\mathcal{S}}(h) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K}$$

Representer theorem (chapter 3)

$$\hat{h}(x)(\theta) = \sum_{i \in [n]} \sum_{j \in [m]} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_{ij}) \mathbf{A} \hat{\alpha}_{ij}, \quad \hat{\alpha}_{ij} \in \mathbb{R}^d$$

Optimization Problem

Optimization problem

$$\hat{h} = \underset{h \in \mathcal{H}_K}{\arg\min}\ \mathcal{R}_{\mathcal{S}}(h) + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_K}$$

Representer theorem (chapter 3)

$$\hat{h}(x)(\theta) = \sum_{i \in [n]} \sum_{j \in [m]} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_{ij}) \mathsf{A} \hat{\alpha}_{ij}, \quad \hat{\alpha}_{ij} \in \mathbb{R}^d$$

In matrix form (Sylvester equation)

$$\mathsf{K}\hat{\boldsymbol{\alpha}}\mathsf{A} + nm^2\lambda\hat{\boldsymbol{\alpha}} = \mathsf{Y}$$

- $\boldsymbol{\alpha} \in \mathbb{R}^{nm \times d}$, $\mathsf{K} \in \mathbb{R}^{nm \times nm}$, $\mathsf{Y} \in \mathbb{R}^{nm \times d}$
- If $\mathsf{A} = \mathsf{Id}_{\mathbb{R}^d}$

$$\hat{\boldsymbol{\alpha}} = (\mathsf{K} + \lambda nm\, \mathsf{Id}_{\mathbb{R}^{nm}})^{-1}\mathsf{Y}$$

Experimental Results

GAN-based baseline [Cho+18]

Mean square error

Methods	KDEF frontal	RaFD frontal
Ours	0.011 ± 0.001	0.007 ± 0.001
StarGAN	0.029 ± 0.003	0.024 ± 0.007

Classification accuracy

Methods	KDEF frontal	RaFD frontal
Ours	74.81 ± 3.10	77.11 ± 3.97
StarGAN	70.69 ± 8.46	65.88 ± 8.92

Conclusion and Perspectives

Conclusion

Rich framework of integral losses and vv-RKHSs

Tractable optimization problems after approximation

Conclusion

Rich framework of integral losses and vv-RKHSs
Tractable optimization problems after approximation

- New angle to multi-task learning: functional view
- Adapted to functional output regression

Conclusion

Rich framework of integral losses and vv-RKHSs
Tractable optimization problems after approximation

- New angle to multi-task learning: functional view
- Adapted to functional output regression

Take home message for optimization

- Primal: often the simplest
- Dual: convoluted losses
- Manageable computational complexity

Conclusion

Rich framework of integral losses and vv-RKHSs
Tractable optimization problems after approximation

- New angle to multi-task learning: functional view
- Adapted to functional output regression

Take home message for optimization

- Primal: often the simplest
- Dual: convoluted losses
- Manageable computational complexity

Python library torch_itl

Perspectives

More complex Θ

- Quantile regression with $Y \in$ Hilbert

Perspectives

More complex Θ

- Quantile regression with $Y \in$ Hilbert

Bayes risk analysis

- Bound $\mathcal{R}(\hat{h}) - \inf_h \mathcal{R}(h)$

Perspectives

More complex Θ

- Quantile regression with $Y \in$ Hilbert

Bayes risk analysis

- Bound $\mathcal{R}(\hat{h}) - \inf_h \mathcal{R}(h)$

Varying notions of outliers

- $\frac{1}{2} \|\cdot\|_{\mathcal{Y}}^2 \,\square\, \kappa \|\cdot\|_?$

Perspectives

More complex Θ

- Quantile regression with $Y \in$ Hilbert

Bayes risk analysis

- Bound $\mathcal{R}(\hat{h}) - \inf_h \mathcal{R}(h)$

Varying notions of outliers

- $\frac{1}{2} \|\cdot\|_{\mathcal{Y}}^2 \,\square\, \kappa \|\cdot\|_?$

Richer kernels

- Beyond separable $K = K_{\mathcal{X}} \otimes K_{\Theta}$, deep kernels

Perspectives

More complex Θ

- Quantile regression with $Y \in$ Hilbert

Bayes risk analysis

- Bound $\mathcal{R}(\hat{h}) - \inf_h \mathcal{R}(h)$

Varying notions of outliers

- $\frac{1}{2} \|\cdot\|_{\mathcal{Y}}^2 \square \kappa \|\cdot\|_?$

Richer kernels

- Beyond separable $K = K_{\mathcal{X}} \otimes K_{\Theta}$, deep kernels

References

N. Aronszajn. “Theory of reproducing kernels.” In: *Transactions of the American Mathematical Society* (1950), pp. 337–404 (cit. on pp. 17, 18).

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011 (cit. on p. 55).

Olivier Bousquet and André Elisseeff. “Stability and generalization.” In: *Journal of Machine Learning Research* 2 (2002), pp. 499–526 (cit. on pp. 43–45).

References ii

Céline Brouard, Marie Szafranski, and Florence D'Alché-Buc. "Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels." In: *Journal of Machine Learning Research* 17.1 (2016), pp. 6105–6152 (cit. on p. 24).

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem." In: *Analysis and Applications* 4.04 (2006), pp. 377–408 (cit. on pp. 20, 21).

Yunjey Choi et al. "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8789–8797 (cit. on p. 67).

References iii

Theodoros Evgeniou and Massimiliano Pontil. "Regularized multi-task learning." In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2004, pp. 109–117 (cit. on pp. 14, 15, 33–36).

Hachem Kadri et al. "Operator-valued Kernels for Learning from Functional Response Data." In: *Journal of Machine Learning Research* 17.20 (2016), pp. 1–54 (cit. on pp. 43–45, 53, 54, 59).

Roger Koenker and Gilbert Bassett Jr. "Regression quantiles." In: *Econometrica: Journal of the Econometric Society* (1978), pp. 33–50 (cit. on pp. 33–36, 38).

Dimitrios Kollias et al. "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond." In: *International Journal of Computer Vision* 127.6-7 (2019), pp. 907–929 (cit. on p. 63).

Heng Lian. "Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces." In: *Canadian Journal of Statistics* 35.4 (2007), pp. 597–606 (cit. on pp. 53, 54).

Charles Micchelli and Massimiliano Pontil. "On Learning Vector-Valued Functions." In: *Neural Computation* 17 (2005), pp. 177–204 (cit. on pp. 22, 23, 40, 41).

Siamak Mehrkanoon, Andreas Zell, and Johan AK Suykens. "Scalable Hybrid Deep Neural Kernel Networks.." In: *ESANN*. 2017 (cit. on pp. 49, 50).

Alba Ordonez et al. "Explaining decisions of deep neural networks used for fish age prediction." In: *PloS one* 15.6 (2020), e0235013 (cit. on pp. 49, 50).

G. Pedrick. *Theory of reproducing kernels for Hilbert spaces of vector-valued functions*. Tech. rep. University of Kansas, Department of Mathematics, 1957 (cit. on pp. 14, 15).

R Tyrrell Rockafellar. *Convex analysis*. Vol. 36. Princeton university press, 1970 (cit. on p. 24).

A. Rahimi and B. Recht. “Random Features for Large-Scale Kernel Machines.” In: *Advances in Neural Information Processing Systems (NIPS)*. 2007, pp. 1177–1184 (cit. on p. 51).

J. Ramsay and B. Silverman. *Functional Data Analysis*. 1997 (cit. on pp. 14, 15).

James A Russell. “A circumplex model of affect.” In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178 (cit. on pp. 12, 13, 63).

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008 (cit. on pp. 17, 18).

Bernhard Schölkopf et al. "Estimating the support of a high-dimensional distribution." In: *Neural computation* 13.7 (2001), pp. 1443–1471 (cit. on pp. 33–36).

Christian Szegedy et al. "Rethinking the inception architecture for computer vision." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. on pp. 49, 50).

Ichiro Takeuchi et al. "Parametric task learning." In: *Advances in Neural Information Processing Systems (NIPS)*. 2013, pp. 1358–1366 (cit. on pp. 33–36).

Raviteja Vemulapalli and Aseem Agarwala. “A compact embedding for facial expression similarity.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5683–5692 (cit. on pp. 12, 13).

Zichao Yang et al. “Deep fried convnets.” In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1476–1483 (cit. on pp. 49, 50).

Bianca Zadrozny and Charles Elkan. “Learning and making decisions when costs and probabilities are both unknown.” In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2001, pp. 204–213 (cit. on pp. 33–36).

Ding-Xuan Zhou. “Derivative reproducing properties for kernel methods in learning theory.” In: *Journal of Computational and Applied Mathematics* 220 (2008), pp. 456–463 (cit. on p. 47).

Appendix

Random Fourier Features

Valid for shift-invariant kernels: $k(x, x') = k_0(x - x')$
Bochner's theorem: $\exists$ unique finite measure ρ_k s.t.

$$k(x, z) = \int_{\mathbb{R}^d} \cos\left(\langle \omega, x - z \rangle\right) \mathrm{d}\rho_k(\omega).$$

Given some integer m and $(\omega_j)_{j=1}^m$ i.i.d. sampled from ρ_k, define

$$\forall (x, x') \in \mathcal{X}^2, \quad \tilde{k}(x, x') = \frac{1}{m} \sum_{j=1}^{m} \cos\left(\langle \omega_j, x - x' \rangle\right)$$

Feature map

$$\tilde{\phi}(x) = \frac{1}{\sqrt{m}} (\cos\left(\omega_1^\top x\right), \ldots, \cos\left(\omega_m^\top x\right), \sin\left(\omega_1^\top x\right), \ldots, \sin\left(\omega_m^\top x\right))^\top.$$

Integral Operator Eigendecomposition

Problem: find (λ, ψ)

$$T_k\psi = \lambda\psi$$

Hard in general, few closed form (Laplace kernel & μ Lebesgue)
Reduces to SVD with RFF

$$\Psi_{i,j} = \int_\Theta \cos(\omega_i^\top \theta)\cos(\omega_j^\top \theta)\mathrm{d}\mu(\theta) \quad \Psi_{i+m,j+m} = \int_\Theta \sin(\omega_i^\top \theta)\sin(\omega_j^\top \theta)\mathrm{d}\mu(\theta)$$
$$\Psi_{i+m,j} = \int_\Theta \sin(\omega_i^\top \theta)\cos(\omega_j^\top \theta)\mathrm{d}\mu(\theta) \quad \Psi_{i,j+m} = \int_\Theta \cos(\omega_i^\top \theta)\sin(\omega_j^\top \theta)\mathrm{d}\mu(\theta)$$

Eigendecomposition of Ψ gives coefficients/eigenvalues

Experimental Setup Quantile Regression

- $k_{\mathcal{X}}$, k_{Θ} Gaussian
- Smoothed pinball ”a la Huber”
- LBFGS on $\boldsymbol{\alpha}$

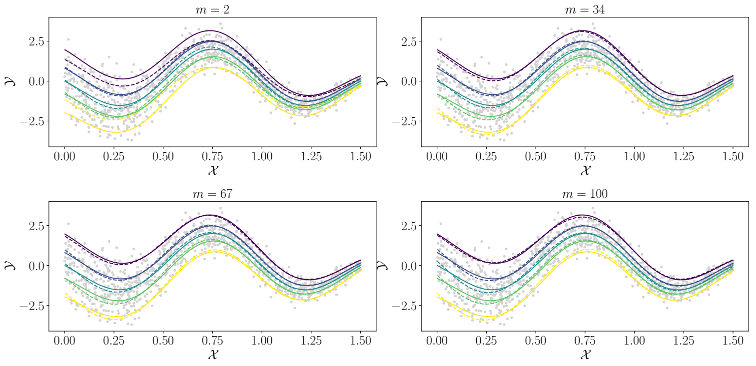
Other possibility: duality
Non smooth in primal -> Smooth in dual + linear constraints

Quantitative Results Quantile Regression

DATASET	JQR				IND-QR				IQR	
	(PINBALL	PVAL)	(CROSS	PVAL)	(PINBALL	PVAL)	(CROSS	PVAL)	PINBALL	CROSS
CobarOre	159 ± 24	$9 \cdot 10^{-01}$	0.1 ± 0.4	$6 \cdot 10^{-01}$	150 ± 21	$2 \cdot 10^{-01}$	0.3 ± 0.8	$7 \cdot 10^{-01}$	165 ± 36	2.0 ± 6.0
engel	175 ± 555	$6 \cdot 10^{-01}$	0.0 ± 0.2	$1 \cdot 10^{+00}$	63 ± 53	$8 \cdot 10^{-01}$	4.0 ± 12.8	$8 \cdot 10^{-01}$	47 ± 6	0.0 ± 0.1
BostonHousing	49 ± 4	$8 \cdot 10^{-01}$	0.7 ± 0.7	$2 \cdot 10^{-01}$	49 ± 4	$8 \cdot 10^{-01}$	$\mathbf{1.3 \pm 1.2}$	$1 \cdot 10^{-05}$	49 ± 4	0.3 ± 0.5
caution	88 ± 17	$6 \cdot 10^{-01}$	0.1 ± 0.2	$6 \cdot 10^{-01}$	89 ± 19	$4 \cdot 10^{-01}$	$\mathbf{0.3 \pm 0.4}$	$2 \cdot 10^{-04}$	85 ± 16	0.0 ± 0.1
ftcollinssnow	154 ± 16	$8 \cdot 10^{-01}$	0.0 ± 0.0	$6 \cdot 10^{-01}$	155 ± 13	$9 \cdot 10^{-01}$	0.2 ± 0.9	$8 \cdot 10^{-01}$	156 ± 17	0.1 ± 0.6
highway	103 ± 19	$4 \cdot 10^{-01}$	0.8 ± 1.4	$2 \cdot 10^{-02}$	99 ± 20	$9 \cdot 10^{-01}$	$\mathbf{6.2 \pm 4.1}$	$1 \cdot 10^{-07}$	105 ± 36	0.1 ± 0.4
heights	127 ± 3	$1 \cdot 10^{+00}$	0.0 ± 0.0	$1 \cdot 10^{+00}$	127 ± 3	$9 \cdot 10^{-01}$	0.0 ± 0.0	$1 \cdot 10^{+00}$	127 ± 3	0.0 ± 0.0
sniffer	43 ± 6	$8 \cdot 10^{-01}$	0.1 ± 0.3	$2 \cdot 10^{-01}$	44 ± 5	$7 \cdot 10^{-01}$	$\mathbf{1.4 \pm 1.2}$	$6 \cdot 10^{-07}$	44 ± 7	0.1 ± 0.1
snowgeese	55 ± 20	$7 \cdot 10^{-01}$	0.3 ± 0.8	$3 \cdot 10^{-01}$	53 ± 18	$6 \cdot 10^{-01}$	0.4 ± 1.0	$5 \cdot 10^{-02}$	57 ± 20	0.2 ± 0.6
ufc	81 ± 5	$6 \cdot 10^{-01}$	$\mathbf{0.0 \pm 0.0}$	$4 \cdot 10^{-04}$	82 ± 5	$7 \cdot 10^{-01}$	$\mathbf{1.0 \pm 1.4}$	$2 \cdot 10^{-04}$	82 ± 4	0.1 ± 0.3
BigMac2003	80 ± 21	$7 \cdot 10^{-01}$	$\mathbf{1.4 \pm 2.1}$	$4 \cdot 10^{-04}$	74 ± 24	$9 \cdot 10^{-02}$	$\mathbf{0.9 \pm 1.1}$	$7 \cdot 10^{-05}$	84 ± 24	0.2 ± 0.4
UN3	98 ± 9	$8 \cdot 10^{-01}$	0.0 ± 0.0	$1 \cdot 10^{-01}$	99 ± 9	$1 \cdot 10^{+00}$	$\mathbf{1.2 \pm 1.0}$	$1 \cdot 10^{-05}$	99 ± 10	0.1 ± 0.4
birthwt	141 ± 13	$1 \cdot 10^{+00}$	0.0 ± 0.0	$6 \cdot 10^{-01}$	140 ± 12	$9 \cdot 10^{-01}$	0.1 ± 0.2	$7 \cdot 10^{-02}$	141 ± 12	0.0 ± 0.0
crabs	$\mathbf{11 \pm 1}$	$4 \cdot 10^{-05}$	0.0 ± 0.0	$8 \cdot 10^{-01}$	$\mathbf{11 \pm 1}$	$2 \cdot 10^{-04}$	$\mathbf{0.0 \pm 0.0}$	$2 \cdot 10^{-05}$	13 ± 3	0.0 ± 0.0
GAGurine	61 ± 7	$4 \cdot 10^{-01}$	0.0 ± 0.1	$3 \cdot 10^{-03}$	62 ± 7	$5 \cdot 10^{-01}$	$\mathbf{0.1 \pm 0.2}$	$4 \cdot 10^{-04}$	62 ± 7	0.0 ± 0.0
geyser	105 ± 7	$9 \cdot 10^{-01}$	0.1 ± 0.3	$9 \cdot 10^{-01}$	105 ± 6	$9 \cdot 10^{-01}$	0.2 ± 0.3	$6 \cdot 10^{-01}$	104 ± 6	0.1 ± 0.2
gilgais	51 ± 6	$5 \cdot 10^{-01}$	0.1 ± 0.1	$1 \cdot 10^{-01}$	49 ± 6	$6 \cdot 10^{-01}$	$\mathbf{1.1 \pm 0.7}$	$2 \cdot 10^{-05}$	49 ± 7	0.3 ± 0.3
topo	69 ± 18	$1 \cdot 10^{+00}$	0.1 ± 0.5	$1 \cdot 10^{+00}$	71 ± 20	$1 \cdot 10^{+00}$	$\mathbf{1.7 \pm 1.4}$	$3 \cdot 10^{-07}$	70 ± 17	0.0 ± 0.0
mcycle	66 ± 9	$9 \cdot 10^{-01}$	0.2 ± 0.3	$7 \cdot 10^{-03}$	66 ± 8	$9 \cdot 10^{-01}$	$\mathbf{0.3 \pm 0.3}$	$7 \cdot 10^{-06}$	65 ± 9	0.0 ± 0.1
cpus	$\mathbf{7 \pm 4}$	$2 \cdot 10^{-04}$	$\mathbf{0.7 \pm 1.0}$	$5 \cdot 10^{-04}$	$\mathbf{7 \pm 5}$	$3 \cdot 10^{-04}$	$\mathbf{1.2 \pm 0.8}$	$6 \cdot 10^{-08}$	16 ± 10	0.0 ± 0.0

Impact of m in Quantile Regression

Number of sampled locations $(\theta_j)_{j=1}^m$



Deep Kernel Learning with Random Fourier Features

Parameterized model:

$$h(x)(\theta) = \tilde{\phi}(x)^\top \boldsymbol{\alpha} \tilde{\phi}(\theta)$$

Optimization problem:

$$\min_{\mathsf{v} \in \mathcal{V}^n} \mathbb{E}_{\theta\sim\mu} \underbrace{\left[\frac{1}{n} \sum_{i=1}^{n} \ell \left(\theta, \left[\left(\mathsf{K}_{\mathcal{X}} \otimes \widetilde{\Phi}(\theta)^\sharp \right) \mathsf{v} \right]_i, y_i(\theta) \right) + \frac{\lambda}{2} \operatorname{Tr}\left(\mathsf{K}_{\mathcal{X}} \mathsf{v}\mathsf{v}^\top \right) \right]}_{:= \mathcal{J}(\theta, \mathsf{v})}.$$

-> Stochastic gradient descent, compatible with NN

Projected GD for Huber loss

$$\mathcal{J}(\boldsymbol{\alpha}) := \mathrm{Tr}\left(\frac{1}{2}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \boldsymbol{\alpha}\mathsf{R}^\top + \frac{1}{2\lambda n}\mathsf{K}_{\mathcal{X}}\boldsymbol{\alpha}\Lambda\boldsymbol{\alpha}^\top\right)$$

Gradient step:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \gamma\left(\boldsymbol{\alpha}^{(t)} + \frac{1}{\lambda n}\mathsf{K}_{\mathcal{X}}\boldsymbol{\alpha}^{(t)}\Lambda - \mathsf{R}\right)$$

Projection step:

$$\boldsymbol{\alpha}_{i:}^{(t+1)} = \min\left(\frac{\kappa}{\left\|\boldsymbol{\alpha}_{i:}^{(t+1)}\right\|_2}, 1\right)\boldsymbol{\alpha}_{i:}^{(t+1)}$$

Stepsize $\gamma = \frac{1}{C}$:

$$\nabla\mathcal{J}(\boldsymbol{\alpha}) = \boldsymbol{\alpha} + \frac{1}{\lambda n}\mathsf{K}_{\mathcal{X}}\boldsymbol{\alpha}\Lambda - \mathsf{R}, \quad C = 1 + \frac{1}{\lambda n}\left\|\mathsf{K}_{\mathcal{X}}\right\|_{\mathrm{op}}\lambda_1$$

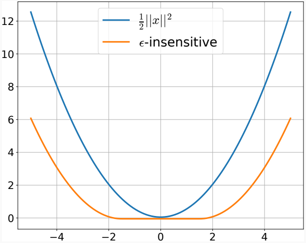
ϵ-insensitive Losses

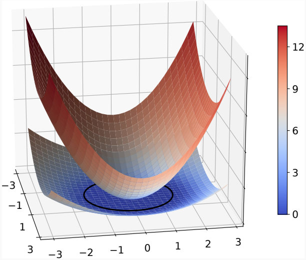
$$\forall y \in \mathcal{Y}, \quad L_\epsilon(y) = \begin{cases} 0 & \text{if } \|y\|_{\mathcal{Y}} \leqslant \epsilon \\ \inf\limits_{\|d\|_{\mathcal{Y}} \leqslant 1} L(y - \epsilon d) & \text{otherwise} \end{cases}$$

Using convolutions:

$$L_\epsilon = L \,\square\, \chi_{\mathcal{B}_\epsilon}(\cdot)$$

$$L_\epsilon^\star = L^\star + \epsilon \|\cdot\|$$

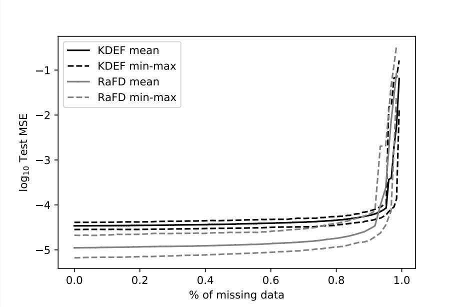




Experimental Results: Partial Observations

What if we do not observe all emotions for all subjects ?

- Random mask $(\eta_{i,j})_{i\in[n],j\in[m]} \in \{0,1\}^{n\times m}$
- Use $z_i(\theta_{i,j})$ only if $\eta_{i,j} = 1$
- Percentage of missing data $p := \frac{1}{nm}\sum_{i,j\in[n]\times[m]} \eta_{i,j}$



Logarithm of the test MSE (min-mean-max) as a function of the percentage of missing data.

Qualitative Results

Radial sampling in the emotion direction