Machine Learning in Banach Spaces: A Black-box or White-box Method?

Qi Ye
School of Mathematical Sciences
South China Normal University

Workshop of Applied Functional Analysis
August 28 - September 2, 2022
Banff International Research Station

Outline

Plato,
428/424-348 BC

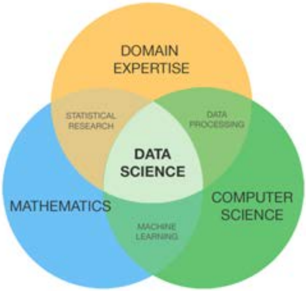
Platonism is a form of realism. Mathematics is abstract, has no space-time or causal properties, and is immutable.

--- *Platonism in the Philosophy of Mathematics*

David Hilbert,
1862-1943

Formalism states that the theory of mathematical logic is the rules for certain string operations.

--- *Formalism in the Philosophy of Mathematics*



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

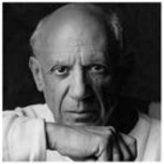
We should dramatically increase investment for research and development in privacy-enhancing technologies, encouraging cross-cutting research that involves not only computer science and mathematics, but also social science, communications, and legal disciplines.

--- *Big Data:Seizing Opportunities, Preserving Values* (Obama White House Archives)

Steve Jobs,
1955-2011

Innovation is to make all sorts of things together.

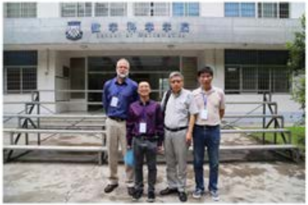
--- *Steve Jobs*



Pablo Picasso,
1881-1973

Good artists copy, great artists steal.

--- *Pablo Picasso*



Master Advisor : Liren Huang

Doctoral Advisor : Gregory E. Fasshauer

Postdoctoral Collaborator : Yuesheng Xu

Optimization

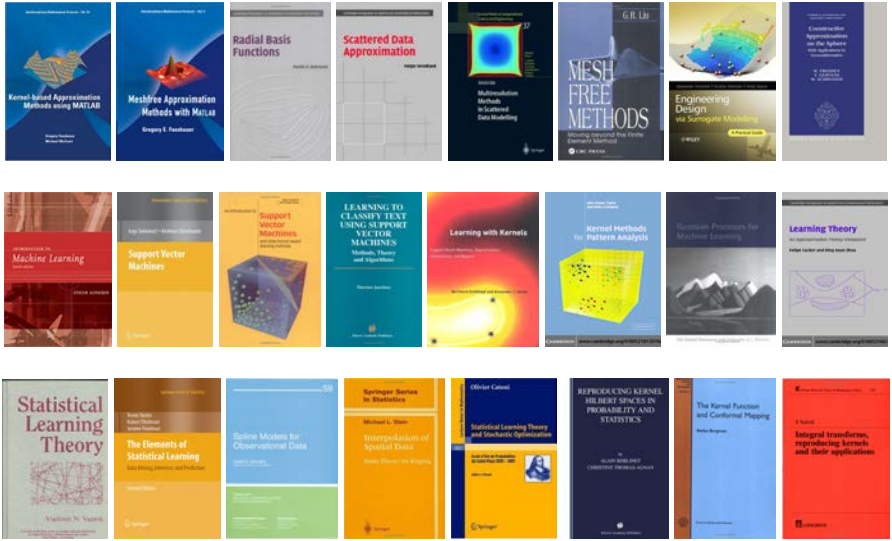
Kung Fu Ng
PhD Advisor of Huang

Approximation Theory

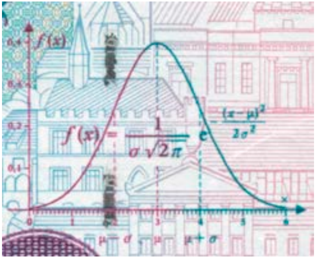
Larry L. Schumaker
PhD Advisor of Fasshauer

Inverse Problem

Charles A. Micchelli
Postdoctoral Collaborator of Xu


Kernel-based Approximation Methods using MATLAB
Meshfree Approximation Methods with MATLAB
Gregory E. Fasshauer
Radial Basis Functions
Scattered Data Approximation
Multiresolution Methods in Scattered Data Modelling
G.R. Liu
MESH FREE METHODS
Moving beyond the Finite Element Method
Engineering Design via Surrogate Modelling
Machine Learning
Support Vector Machines
Support Vector Machines
LEARNING TO CLASSIFY TEXT USING SUPPORT VECTOR MACHINES
Methods, Theory and Algorithms
Learning with Kernels
Kernel Methods for Pattern Analysis
Learning Theory
Statistical Learning Theory
Vladimir N. Vapnik
The Elements of Statistical Learning
Spline Models for Observational Data
Springer Series in Statistics
Michael L. Stein
Interpolation of Spatial Data
Olivier Catoni
Statistical Learning Theory and Stochastic Optimization
REPRODUCING KERNEL HILBERT SPACES IN PROBABILITY AND STATISTICS
The Kernel Function and Conformal Mapping
Integral transforms, reproducing kernels and their applications

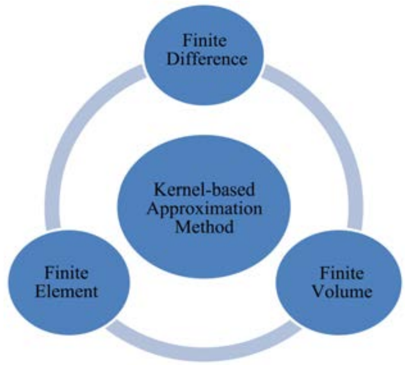
Carl Friedrich Gauß, 1777-1855



Gaussian Kernels

Gauss mentioned Gaussian kernels that now so often carry his name in 1809 in his second book [12].

By the techniques of numerical analysis, stochastic analysis, and regression analysis, we use the kernel-based approximation methods to combine the algorithms of finite difference, finite elment and finite volume in a system.

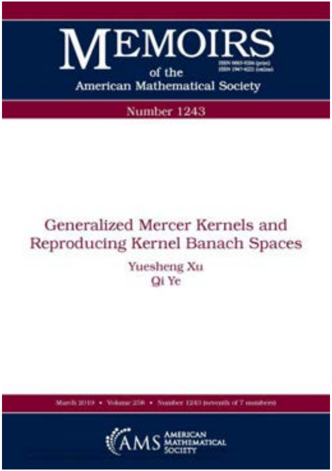


The data-driven methods are used to analyze black-box models, and the model-driven methods are used to analyze white-box models.

Remark. With the same thought of the Tai Chi diagram, we use the discrete local information of the black-box and white-box models to construct the global approximate solutions by the algorithms of regularized learning.

Outline

- The details of reproducing kernel Banach spaces (RKBS) are mentioned in my paper [29] for 122 pages.
- It studies the constructions and theorems of RKBS and verifies the learning theory in RKBS, specially, the sparse learning methods in 1-norm RKBS.



- L. Huang, C. Liu, L. Tan, and Q. Ye, Generalized representer theorems in Banach spaces, Analysis and Applications. **19** (2021), 125–146 [14].
- Qi Ye. Analysis of regularized learning for generalized data in Banach spaces. arXiv:2109. 03159 (2021), 1-36 [30].

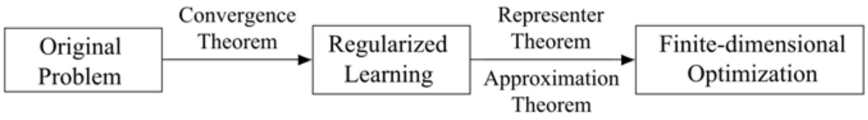
The work of regularized learning in Banach spaces provides another road to study the algorithms of machine learning including:

- the interpretability in approximation theory,
- the nonconvexity and nonsmoothness in optimization theory,
- the generalization and overfitting in regularization theory.

The original problem is to solve an exact solution from

$$\inf_{f \in \mathcal{B}} R(f),$$

where $\mathcal{B}$ is a Banach space and $R : \mathcal{B} \to [0, \infty)$ is an expected risk function.



Key to Success of Regularized Learning

Remark. We can solve the finite dimensional optimization problems by

- alternating direction multiplier methods (splitting methods),
- composite optimization methods.

- We have generalized data

$$(\xi_1^*, y_1), (\xi_2^*, y_2), \ldots, (\xi_{N_n}^*, y_{N_n}) \in \mathcal{B}_* \times Y, \quad \text{for } n \in \mathbb{N},$$

which can be written as

$$(\boldsymbol{\xi}_n^*, \boldsymbol{y}_n) \in \mathcal{B}_*^{N_n} \times Y^{N_n}, \quad \text{for } n \in \mathbb{N},$$

where $\mathcal{B}_*$ is a predual space of $\mathcal{B}$ and Y is an output range.

- For each nth approximate step, we have a multi-loss function as

$$\mathbb{L}_n : \mathcal{B}_*^{N_n} \times Y^{N_n} \times \mathbb{R}^{N_n} \to [0, \infty),$$

for example,

$$\mathbb{L}_n(\boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{t}) = \frac{1}{N_n} \sum_{k=1}^{N_n} L(\xi_k^*, y_k, t_k),$$

where L is a classical loss function.

- The regularized learning is to solve an approximate solution from

$$\inf_{f\in\mathcal{B}} R_n(f)+\lambda\Phi(\|f\|),$$

where R_n is an empirical risk function, that is,

$$R_n(f):=\mathbb{L}_n(\boldsymbol{\xi}_n^*,\boldsymbol{y}_n,\langle f,\boldsymbol{\xi}_n^*\rangle),\quad \text{for } f\in\mathcal{B},$$

and $\lambda>0$, $\Phi:[0,\infty)\to[0,\infty)$ is an increasing function.

- The approximate solution can be equivalently or approximately solved by a finite-dimensional optimization, that is,

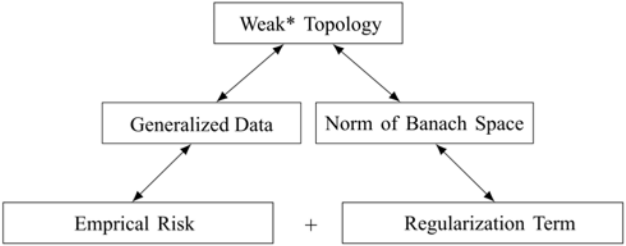
$$\inf_{f\in\mathcal{U}_m} R_n(f)+\lambda\Phi(\|f\|),$$

where $\mathcal{U}_m$ is a subset of $\mathcal{B}$ such that there exists an surjection Γ_m from Ω_m onto $\mathcal{U}_m$, where $\Omega_m\subset\mathbb{R}^m$ and $m\in\mathbb{N}$.

The representer theorems, approximation theorems, and convergence theorems of regularized learning are guaranteed by the assumptions:

1. The R_n converges pointwise to R when $n \to \infty$.
2. The collection of all generalized input data is relatively compact in $\mathcal{B}_*$ and the collection of all multi-loss functions is uniformly local Lipschitz continuous.

$\Rightarrow$ $\{R_n : n \in \mathbb{N}\}$ is weakly* equicontinuous on any closed ball of origin.



Regularized Learning for Generalized Data

- Representer Theorems in [30]
  - Theorem 4.2 shows that the approximate solutions are equivalently solved by the finite-dimensional optimization.
- Approximation Theorems in [30]
  - Theorem 4.5 shows that the approximate solutions are approximately solved by the finite-dimensional optimization.
- Convergence Theorems in [30]
  - Theorem 4.6 shows the existence of a weakly* convergent subnet of approximate solutions to an exact solution from the orignial problem.
  - Theorem 4.7 shows that any element, to which a subnet of approximate solutions weakly* converges, is an exact solution.
  - Theorem 4.8 shows that the subsequence of approximate solutions for special regularization parameters is weakly* convergent.

- Suppose that an exact solution belonging to $\mathcal{B}^b \cap \mathcal{B}^w$ is solved by the both models, that is,

Black-box Model

$$\inf_{f\in\mathcal{B}^b} R^b(f),$$
$$\inf_{f\in\mathcal{B}^b} R_n^b(f) + \lambda \|f\|_b .$$

White-box Model

$$\inf_{f\in\mathcal{B}^w} R^w(f),$$
$$\inf_{f\in\mathcal{B}^w} R_n^w(f) + \lambda \|f\|_w .$$

- Let $\mathcal{B} := \mathcal{B}^b \cap \mathcal{B}^w$ with $\|f\| := \|f\|_b + \|f\|_w$. Thus, the composite model can be written as

$$\inf_{f\in\mathcal{B}} R(f), \quad \inf_{f\in\mathcal{B}} R_n(f) + \lambda \|f\| ,$$
where $R := R^b + R^w,\ R_n := R_n^b + R_n^w .$

- Let $T^b_{n,\lambda} := R^b_n + \lambda\|\cdot\|_b$, $T^w_{n,\lambda} := R^w_n + \lambda\|\cdot\|_w$, "prox" represent the proximal operator, and $\theta > 0$ be a stepsize. Thus, two iterative algorithms can be used to solve the two models above, respectively,

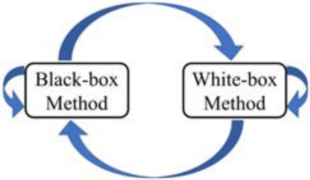
Black-box Method

$$g_{k+1} \in \mathrm{prox}_{\theta T^b_{n,\lambda}}(g_k).$$

White-box Method

$$g_{k+1} \in \mathrm{prox}_{\theta T^w_{n,\lambda}}(g_k).$$

- By the Douglas-Rochford splitting methods, the composite algorithms can be written as

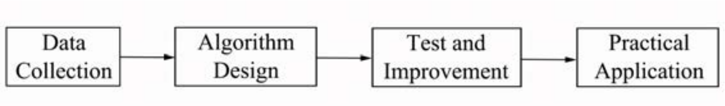


$$\begin{aligned} h^b_k &\in \mathrm{prox}_{\theta T^b_{n,\lambda}}(g_k), \\ h^w_k &\in \mathrm{prox}_{\theta T^w_{n,\lambda}}(2h^b_k - g_k), \\ g_{k+1} &:= g_k + \sigma(h^w_k - h^b_k). \end{aligned}$$

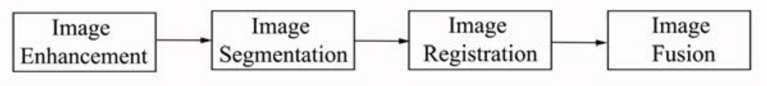
Outline

By the theory of regulizaed learning, we analyze the big data in education and medicine.

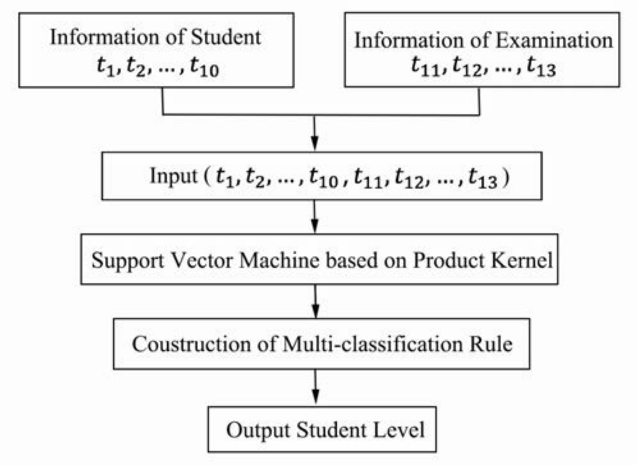
- The big data analysis of mathematics education of middle school represents:



- The big data analysis of digital images of pancreatic cancer represents:

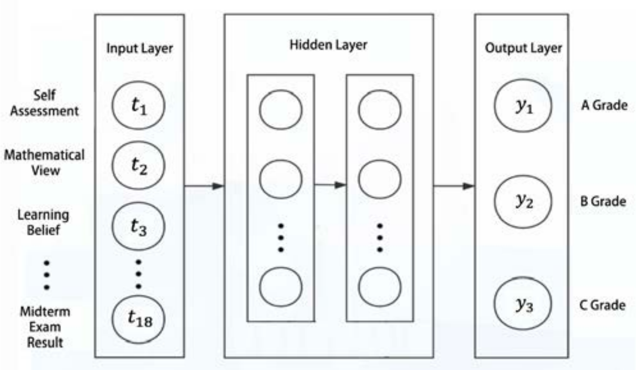


We use the black-box methods of support vector machines to analyze the data of mathematics education for hierarchical teaching.



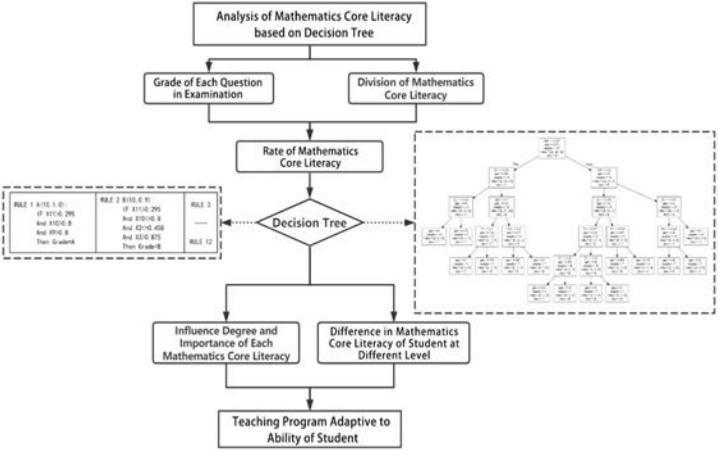
Software Copyright Registration Number: 2021SR0767033

We use the black-box methods of artificial neural networks to analyze the data of mathematics education for hierarchical teaching.



Software Copyright Registration Number: 2021SR0621919

We use the white-box methods of decision trees to analyze the data of mathematics education for hierarchical teaching.



Software Copyright Registration Number: 2022SR0568823

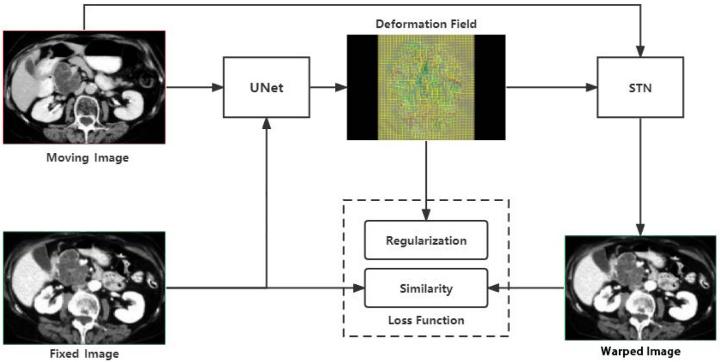
We use the model-driven methods to solve the image registration of pancreatic cancer.

- Diffeomorphic Demons Model (Tom Vercauteren, 2009 [26])

$$L(F, M, s, u) = \|F - M \circ s\|^2 + \lambda\|s - u\| + \mu\|\nabla u\|^2.$$

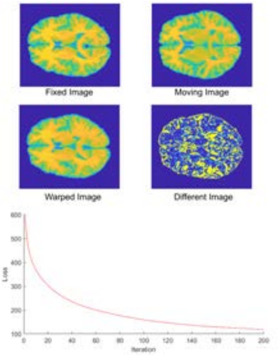
Remark. The F is a fixed image, M is a moving image, s is the approximate transformation, u is the real transformation, λ is the parameter of the spatial uncertainties term, and μ is the parameter of the regularization term.

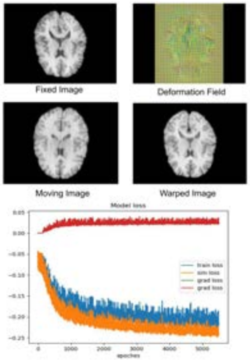
We use the data-driven methods to solve the image registration of pancreatic cancer.



Deep Learning based on Registration: VoxelMorph

Based on the LPBA40 data set, we use the model-driven and data-driven methods to register the brain CT images as follows:



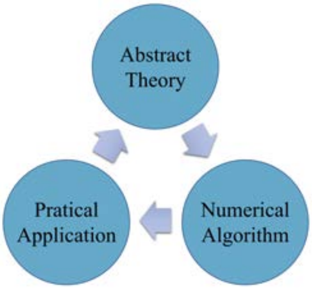


Numerical Results of Diffeomorphic Demons Model and VoxelMorph Model

Remark. The computational time, accuracy, and process of the model-driven method are longer, higher, and smoother than the data-driven method.

By the theory of regularized learning, we will combine the black-box and white-box methods to construct the composite algorithms to solve the problems of the big data analysis in education and medicine. Our original ideas are inspired by the eastern philosophy such as golden mean and Tai Chi diagram.

Plato and Aristotle



I hope someday you'll join us. And the world will live as one.

--- John Winston Lennon

John Winston Lennon,
1940-1980

Thank You!

国家自然科学基金委员会
NSFC
National Natural Science Foundation of China

广东省科学技术厅
Department of Science and Technology of Guangdong Province

广东省教育厅
DEPARTMENT OF EDUCATION OF GUANGDONG PROVINCE

1933
华南师范大学
SOUTH CHINA NORMAL UNIVERSITY

http://mlopt.scnu.edu.cn/en/

Outline

[1] Ethem Alpaydin.
Introduction to Machine Learning.
MIT Press, Cambridge, 2020.

[2] Stefan Bergman.
The Kernel Function and Conformal Mapping.
American Mathematical Society, Providence, 1970.

[3] Alain Berlinet and Christine Thomas-Agnan.
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Springer Science & Business Media, Berlin, 2011.

[4] Martin D. Buhmann.

Radial Basis Functions: Theory and Implementations.

Cambridge University Press, Cambridge, 2003.

[5] Olivier Catoni.

Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001.

Springer Science & Business Media, Berlin, 2004.

[6] Noppadol Chumchob.

Vectorial total variation-based regularization for variational image registration.

IEEE Transactions on Image Processing, 22(11):4551–4559, 2013.

[7] Nello Cristianini and John Shawe-Taylor.

An Introduction to Support Vector Machines and other Kernel-based Learning Methods.

Cambridge University Press, Cambridge, 2000.

[8] Felipe Cucker and Ding Xuan Zhou.

Learning Theory: An Approximation Theory Viewpoint.

Cambridge University Press, Cambridge, 2007.

[9] Gregory E. Fasshauer.

Meshfree Approximation Methods with MATLAB.

World Scientific, Singapore, 2007.

[10] Gregory E. Fasshauer and Michael J. McCourt.

Kernel-based Approximation Methods Using MATLAB.

World Scientific, Singapore, 2015.

[11] Willi Freeden, Theo Gervens, and Michael Schreiner.

Constructive Approximation on the Sphere with Applications to Geomathematics.

Oxford University Press on Demand, Oxford, 1998.

[12] Carl F. Gauß.

Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections.

Hamburg: Friedrich Perthes and I. H. Besser, 1809.

[13] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman.

The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Springer, New York, 2009.

[14] Liren Huang, Chunguang Liu, Lulin Tan, and Qi Ye.

Generalized representer theorems in Banach spaces.

Analysis and Applications, 19(01):125–146, 2021.

[15] Armin Iske.

Multiresolution Methods in Scattered Data Modelling.

Springer Science & Business Media, Berlin, 2004.

[16] Gui-Rong Liu.
Meshfree Methods: Moving Beyond the Finite Element Method.
CRC press, Boca Raton, 2009.

[17] Carl E. Rasmussen and Christopher K. I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, 2005.

[18] Saburou Saitoh.
Integral Transforms, Reproducing Kernels and Their Applications.
CRC Press, Boca Raton, 1997.

[19] Bernhard Schölkopf and Alexander J. Smola.

Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.

MIT Press, Cambridge, 2018.

[20] John Shawe-Taylor and Nello Cristianini.

Kernel Methods for Pattern Analysis.

Cambridge University Press, Cambridge, 2004.

[21] András Sobester, Alexander Forrester, and Andy Keane.

Engineering Design via Surrogate Modelling: A Practical Guide.

John Wiley & Sons, Hoboken, 2008.

[22] Michael L. Stein.

Interpolation of Spatial Data: Some Theory for Kriging.

Springer Science & Business Media, Berlin, 2012.

[23] Ingo Steinwart and Andreas Christmann.

Support Vector Machines.

Springer Science & Business Media, Berlin, 2008.

[24] Joachims Thorsten.

Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.

Kluwer, Alphen aan den Rijn, 2002.

[25] Vladimir N. Vapnik.

Statistical Learning Theory.

Wiley-Interscience, Hoboken, 1998.

[26] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache.

Diffeomorphic demons: Efficient non-parametric image registration.

NeuroImage, 1(45):61–72, 2009.

[27] Grace Wahba.

Spline Models for Observational Data.

SIAM, Philadelphia, 1990.

[28] Holger Wendland.

Scattered Data Approximation.

Cambridge University Press, Cambridge, 2004.

[29] Yuesheng Xu and Qi Ye.

Generalized mercer kernels and reproducing kernel Banach spaces.

Memoirs of the American Mathematical Society, 258(1243):1–122, 2019.

[30] Qi Ye.

Analysis of regularized learning for generalized data in Banach spaces.

arXiv preprint arXiv: 2109.03159, pages 1–36, 2021.