Vector-valued Reproducing Kernel Hilbert Spaces

with applications to Function Extension and Image Colorization

Minh Ha Quang

minh.ha.quang@staff.hu-berlin.de

Humboldt Universität zu Berlin

Outline of the Talk

- **Brief Review of Scalar-valued RKHS**
- Vector-valued RKHS
- Function Extension: 2 algorithms
- Application: Image Colorization
- Learning Theory Estimates (if time permits)

Positive Definite Kernels

- X any nonempty set
- $K : X \times X \to \mathbb{R}$ is a (real-valued) positive definite kernel if it is symmetric and

$$\sum_{i,j=1}^{N} a_i a_j K(x_i, x_j) \geq 0$$

  for any finite set of points $\{x_i\}_{i=1}^{N} \in X$ and real numbers $\{a_i\}_{i=1}^{N} \in \mathbb{R}$.
- Complex-valued kernels are often encountered in complex analysis.

RKHS

- Abstract theory due to Aronszajn (1950).
- K a positive definite kernel on $X \times X$. For each $x \in X$, there is a function $K_x : X \to \mathbb{R}$, with $K_x(t) = K(x, t)$.

$$\mathcal{H}_K = \overline{\{\sum_{i=1}^{N} a_i K_{x_i} : N \in \mathbb{N}\}}$$

with inner product

$$\langle \sum_i a_i K_{x_i}, \sum_j b_j K_{y_j} \rangle_K = \sum_{i,j} a_i b_j K(x_i, y_j)$$

- $\mathcal{H}_K$ = RKHS associated with K (unique).

RKHS

- **Reproducing Property**: for each $f \in \mathcal{H}_K$, for every $x \in X$

$$f(x) = \langle f, K_x \rangle_K$$

- **Assumption**

$$\kappa = \sup_{x \in X} \sqrt{K(x,x)} < \infty$$

- Then

$$||f||_\infty \leq \kappa ||f||_K$$

Examples: RKHS

For $s > n/2$, the Sobolev space $H^s(\mathbb{R}^n)$, with

$$\|f\|^2_{H^s(\mathbb{R}^n)} = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \left| (1 + |\xi|^2)^{s/2} \widehat{f}(\xi) \right|^2 \, d\xi < \infty,$$

is an RKHS, with kernel

$$K(x,y) = \frac{1}{(2\pi)^n} \widehat{\frac{1}{(1+|\xi|^2)^s}}(x-y)$$

Examples: RKHS

- The Gaussian kernel $K(x,y) = \exp(-\frac{|x-y|^2}{\sigma^2})$ on $\mathbb{R}^n$ induces the space

$$\mathcal{H}_K = \{||f||^2_{\mathcal{H}_K} = \frac{1}{(2\pi)^n(\sigma\sqrt{\pi})^n}\int_{\mathbb{R}^n} e^{\frac{\sigma^2|\xi|^2}{4}}|\widehat{f}(\xi)|^2 d\xi < \infty\}.$$

- The Laplacian kernel $K(x,y) = \exp\left(-a|x-y|\right)$, $a > 0$, on $\mathbb{R}^n$ induces the space

$$\mathcal{H}_K = \{||f||^2_{\mathcal{H}_K} = \frac{1}{(2\pi)^n}\frac{1}{aC(n)}\int_{\mathbb{R}^n}(a^2+|\xi|^2)^{\frac{n+1}{2}}|\hat{f}(\xi)|^2 d\xi < \infty\}$$

with $C(n) = 2^n\pi^{\frac{n-1}{2}}\Gamma(\frac{n+1}{2})$

Examples: RKHS

- The Laplacian kernel has less smoothing effect than the Gaussian kernel (may be useful if we do not want very smooth functions)
- Generalization of the Gaussian kernel: $K(x, y) = \exp(-\frac{|x-y|^p}{\sigma^2})$, where $0 \leq p \leq 2$ (Schoenberg 1938).

Outline of the Talk

- Brief Review of Scalar-valued RKHS
- **Vector-valued RKHS**
- Function Extension: 2 algorithms
- Application: Image Colorization
- Learning Theory Estimates (if time permits)

Vector-valued RKHS

- Laurent Schwartz (1964): very general framework for RKHS of functions with values in locally convex topological spaces
- Some recent works in machine learning related literature: Pontil-Micchelli(2005), Caponnetto-De Vito (2006), Reisert-Burkhardt (2007), Carmeli et al (2006).
- Here we will focus on RKHS of functions with values in a Hilbert space.

Operator-valued kernels

- D a nonempty set, $\mathcal{W}$ a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, $\mathcal{L}(W)$ the Banach space of bounded linear operators on $\mathcal{W}$.
- A function $K : D \times D \to \mathcal{L}(\mathcal{W})$ is said to be an **operator-valued positive definite kernel** if for each pair $(x, y) \in D \times D$, $K(x, y) \in \mathcal{L}(\mathcal{W})$ is a self-adjoint operator and

$$\sum_{i,j=1}^{N} \langle w_i, K(x_i, x_j) w_j \rangle_{\mathcal{W}} \geq 0$$

for every finite set of points $\{x_i\}_{i=1}^{N}$ in D and $\{w_i\}_{i=1}^{N}$ in $\mathcal{W}$, where $N \in \mathbb{N}$.

Vector-valued RKHS

- $\mathcal{W}^D$ = vector space of all functions $f : D \to \mathcal{W}$.
- For each $x \in D$ and $w \in \mathcal{W}$, we form a function $K_x w = K(.,x)w \in \mathcal{W}^D$ defined by

$$(K_x w)(y) = K(y,x)w \quad \text{for all} \quad y \in D.$$

- Consider the set $\mathcal{H}_0 = \mathrm{span}\{\mathrm{K_x w} \mid \mathrm{x} \in \mathrm{D},\ \mathrm{w} \in \mathcal{W}\} \subset \mathcal{W}^{\mathrm{D}}$. For $f = \sum_{i=1}^N K_{x_i} w_i$, $g = \sum_{i=1}^N K_{y_i} z_i \in \mathcal{H}_0$, we define

$$\langle\ f, g\ \rangle_{\mathcal{H}_K} = \sum_{i,j=1}^N \langle w_i, K(x_i, y_j) z_j \rangle_{\mathcal{W}}.$$

Vector-valued RKHS

- Taking the closure of $\mathcal{H}_0$ gives the Hilbert space $\mathcal{H}_K$.
- The **reproducing property** is

$$\langle f(x), w\rangle_{\mathcal{W}} = \langle f, K_x w\rangle_{\mathcal{H}_K} \quad \text{for all} \quad f \in \mathcal{H}_K.$$

- For each $x \in D$ and $f \in \mathcal{H}_K$:

$$||f(x)||_{\mathcal{W}} \leq \sqrt{||K(x,x)||}\ ||f||_{\mathcal{H}_K}.$$

Vector-valued RKHS

- Simple example: let $k(x, y)$ be a real-valued positive definite kernel and B a positive definite matrix. Then

$$K(x, y) = Bk(x, y)$$

is a matrix-valued kernel, which induces a vector-valued RKHS

Outline of the Talk

- Brief Review of Scalar-valued RKHS
- Vector-valued RKHS
- **Function Extension: 2 algorithms**
- Application: Image Colorization
- Learning Theory Estimates (if time permits)

Function Extension

- $D \subset \Omega$ are closed sets in a complete separable metric space
- $f : D \to \mathcal{W}$,
- Goal: extend $f : D \to \mathcal{W}$ to $F : \Omega \to \mathcal{W}$, such that F is close to f on the smaller set D, and reasonably well-behaved on the larger set Ω.

Extension Operator

- Assume we have a kernel $K : \Omega \times \Omega \to \mathcal{W}$.
- Assume that $K(x, x)$ is compact for each x, and that $\sup_{x \in \Omega} ||K(x, x)|| < \infty$.
- For $f : D \to \mathcal{W}$, define $L_K : L^2_\mu(D; \mathcal{W}) \to \mathcal{H}_K(\Omega)$, with

$$L_K f(x) = \int_D K(x, y) f(y) d\mu(y),$$

for every $x \in \Omega$. This defines an extension operator. The adjoint operator $L^*_K : \mathcal{H}_K(\Omega) \to L^2_\mu(D; \mathcal{W})$ is the restriction operator: $L^*_K F = F|_D$

Function Extension

- Find the extension function $F : \Omega \to \mathcal{W}$ by solving the minimization problem

$$\inf_{F \in \mathcal{H}_K(\Omega)} ||f - L_K^* F||^2_{L^2_\mu(D;\mathcal{W})} + \gamma ||F||^2_{\mathcal{H}_K(\Omega)},$$

- This problem has a unique solution

$$F_\gamma = (L_K L_K^* + \gamma I)^{-1} L_K f$$

Function Extension: Spectral Algorithm

- Scalar version: Coifman-Lafon (2005)
- Considered as an operator $L^2_\mu(D;\mathcal{W}) \to L^2_\mu(D;\mathcal{W})$, L_K is compact, positive, with orthonormal spectrum $(\lambda_k, \phi_k)_{k=1}^\infty$.
- Eigenfunction extension: for $\lambda_k > 0$, we extend $\phi_k : D \to \mathcal{W}$ to $\Phi_k : \Omega \to \mathcal{W}$ by

$$\Phi_k(x) = \frac{1}{\lambda_k}\int_D K(x,y)\phi_k(y)d\mu(y), \quad \text{for } x \in \Omega.$$

- To be numerically reliable, one may want to consider only $\lambda_k > \delta$, for some given $\delta > 0$.

Function Extension: Spectral Algorithm

- Compute the eigenvalues and eigenfunctions $\{(\lambda_k, \phi_k)\}$ of $L_K : L^2_\mu(D; \mathcal{W}) \to L^2_\mu(D; \mathcal{W})$.
- Compute the expansion coefficients a_k's of f in the basis $\{\phi_k\}$: $f = \sum_k a_k \phi_k$
- Compute $F_\delta = \sum_{k, \lambda_k > \delta} \frac{\lambda_k}{\lambda_k + \gamma} a_k \Phi_k$, for some $\delta > 0$
- Alternatively, to take care of the case $\lambda_k = 0$, compute directly

$$F_\gamma(x) = \sum_{k=1}^{\infty} \frac{a_k}{\lambda_k + \gamma} \int_D K(x, y) \phi_k(y) d\mu(y)$$

.

Function Extension: Least square

- Assume now that $D = \{x_i\}_{i=1}^m$, with $w_i = f(x_i)$.
- An algorithm with real kernel-based flavor:

$$F_\gamma = \arg \min_{F \in \mathcal{H}_K(\Omega)} \frac{1}{m} \sum_{i=1}^m ||F(x_i) - w_i||^2_{\mathcal{W}} + \gamma ||F||^2_{\mathcal{H}_K(\Omega)}.$$

- This has a unique solution $F_\gamma = \sum_{i=1}^m K_{x_i} a_i$, with $F_\gamma(x) = \sum_{i=1}^m K(x, x_i) a_i$, where the vectors $a_i' s \in \mathcal{W}$ satisfy the m linear equations

$$\sum_{j=1}^m K(x_i, x_j) a_j + m\gamma a_i = w_i.$$

Compare two algorithms

- Spectral: theoretically more general (D can be either discrete or continuous)
- If D is discrete and μ is the uniform distribution, then Least square and Spectral are the same analytically.
- Numerically, Least square is easier to implement and should be expected to be more stable (involves solving well-conditioned systems of linear equations, vs finding eigenvalues/eigenfunctions of the Spectral method).
- The basis functions in Least square are exact (based on the given data points)
- Here we will focus on the Least square method for numerical work

Outline of the Talk

- Brief Review of Scalar-valued RKHS
- Vector-valued RKHS
- Function Extension
- **Application: Image Colorization**
- Learning Theory Estimates (if time permits)

Image Colorization

- Joint work with Sung Ha Kang (Georgia Tech) and Triet Le (Yale)
- Ω is the given grayscale image
- $D \subset \Omega$ is the given region with colors (often very small). The initial function here is $f : D \to \mathbb{R}^3$ (red, green, blue)
- Goal: extend the colors to all of Ω.
- Some (among many) other works this on problem: Levin-Lischinski-Weiss(2004), Sapiro(2005), Qiu-Guan(2005), Fornasier (2006), Buades-Coll-Lisani-Sbert(2007), Kang-March(2007), etc

Nonlocal kernel

- Simplest scenario: all the colors are independent.
- $K(x,y) = \mathrm{diag}(k_1(x,y), k_2(x,y), k_3(x,y))$ where each k_i is a scalar-valued kernel.
- Here we will use scalar-valued kernels of the form

$$k(x,y) = \exp(-\frac{|g_r(x) - g_r(y)|^p}{\sigma_1})\exp(-\frac{|x-y|^p}{\sigma_2})$$

where $g_r(x)$ is the patch of radius r centered at x, of size $(2r+1)\times(2r+1)$, with g denoting the gray level.
- Extend the color function using least square RKHS

Chromaticity and Brightness Model

For sharper resulting images, we consider the CB model of color.

- $f(x) = B(x)C(x)$, where $B(x)$ is the brightness, and $C(x) = (r(x), g(x), b(x)) \in S^2$.
- **Assumption**: we are given the brightness $B(x)$ on all of Ω, but $C(x)$ only on D.
- **Need**: to extend $C(x)$ to all of Ω.
- **Problem**: the set of S^2-valued functions is not a vector space

Stereographic Projection

- **Solution** for the S^2-valued Chromaticity function: Stereographic projection
- Since the colors are all nonnegative and for symmetry, we need a symmetric stereographic projection that projects from the first quadrant
- Projection point: $(-\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$
- Projection plane: $X + Y + Z = 0$

Stereographic Projection

- Forward projection from S^2 onto $X+Y+Z=0$:
  $X = \frac{3x-(x+y+z)}{\sqrt{3}(x+y+z+\sqrt{3})}, \quad Y = \frac{3y-(x+y+z)}{\sqrt{3}(x+y+z+\sqrt{3})}, Z = \frac{3z-(x+y+z)}{\sqrt{3}(x+y+z+\sqrt{3})},$
- Inverse projection from $X+Y+Z=0$ onto S^2:
  $x = \frac{2\sqrt{3}X+1-(X^2+Y^2+Z^2)}{\sqrt{3}(1+X^2+Y^2+Z^2)}, \quad y = \frac{2\sqrt{3}Y+1-(X^2+Y^2+Z^2)}{\sqrt{3}(1+X^2+Y^2+Z^2)},$
  $z = \frac{2\sqrt{3}Z+1-(X^2+Y^2+Z^2)}{\sqrt{3}(1+X^2+Y^2+Z^2)}.$

Image Colorization Algorithm

- Given: Brightness $B(x)$ on all of Ω and Chromaticity on small subset $D \subset \Omega$
- Project $C(x) : D \to S^2$ to $C(x) : D \to \mathbb{R}^2$
- Extend $C(x)$ to $\Omega :\to \mathbb{R}^2$ using the least square algorithm in the RKHS induced by the nonlocal kernel above (kernel constructed using $B(x)$)
- Project the results back onto S^2 to get the extended Chromaticity function from $\Omega \to S^2$
- Multiply the resulting Chromaticity with the given Brightness to obtain the final answer.

Colorization Algorithm - Complexity

- Involves solving 2 systems of linear equations, each of size $m \times m$, where $m = |D|$
- Evaluation step involves computing kernel matrix of size $m \times M$, where $M = |\Omega|$
- Main computation time is in computing the kernel
- Explicit and unique solution, no iteration required

Numerical Examples

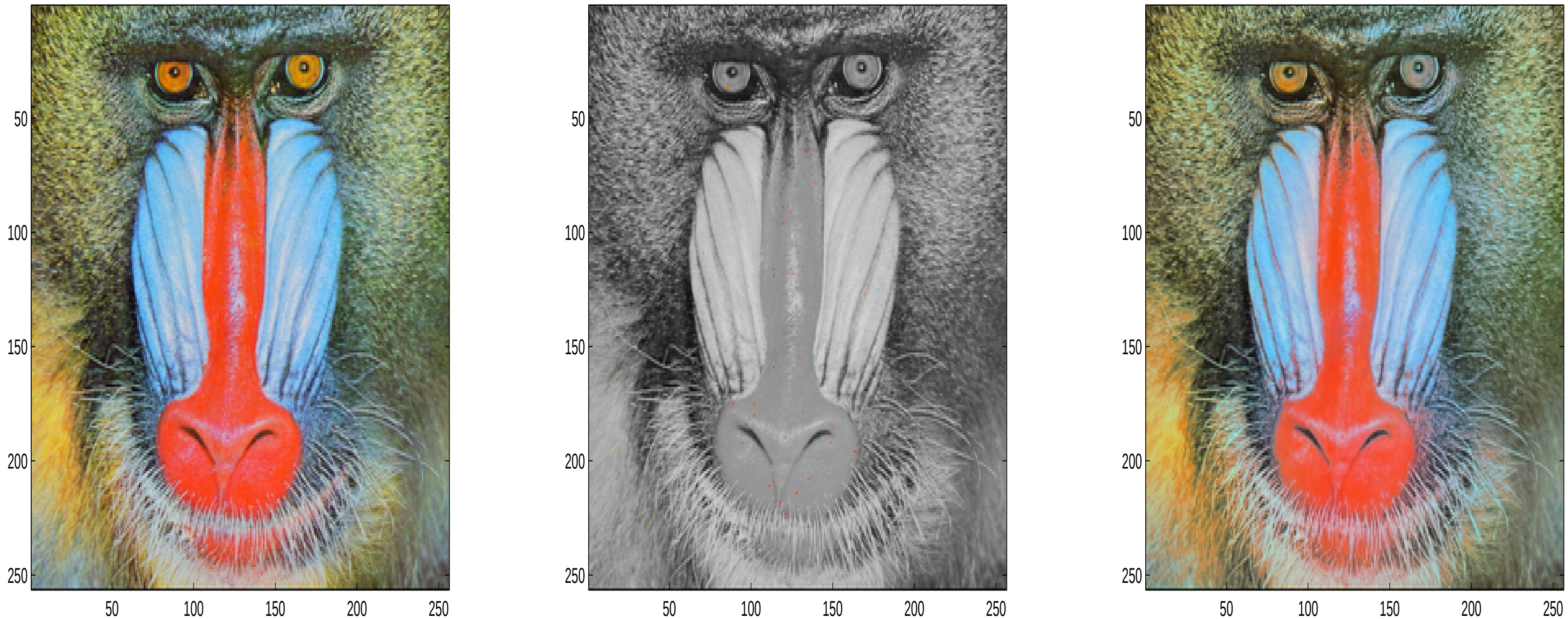


Figure 1: $p = 1$, $r = 1$, $\sigma_1 = 0.5$, $\sigma_2 = 1$. About 0.5% of color is given

Numerical Examples

Figure 2: $p = 1$, $r = 1$, $\sigma_1 = 0.5$, $\sigma_2 = 1$. About 1% of color is given

Numerical Examples

Figure 3: $p = 1$, $r = 1$, $\sigma_1 = 0.5$, $\sigma_2 = 1$. About 1% of color is given

Numerical Examples

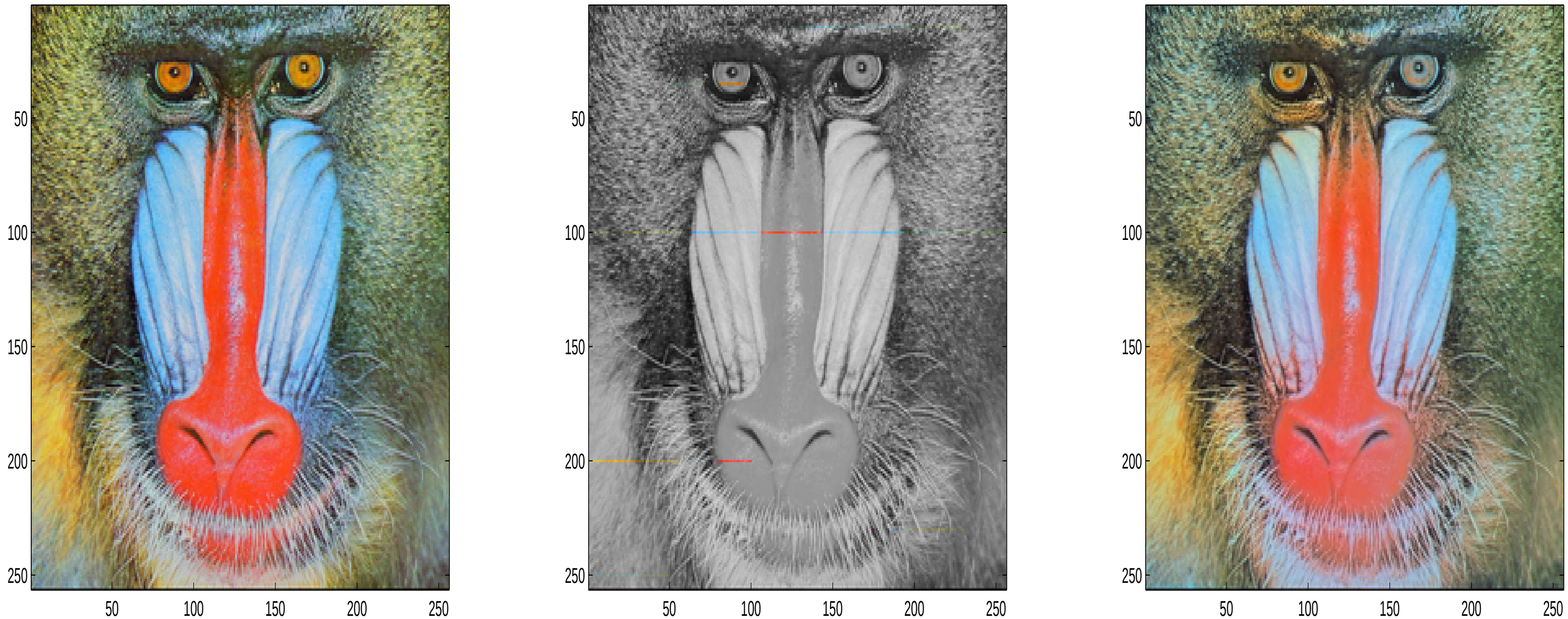


Figure 4: $p = 1$, $r = 2$, $\sigma_1 = 0.5$, $\sigma_2 = 2$. About 0.96% of color is given

Numerical Examples - Cartoon

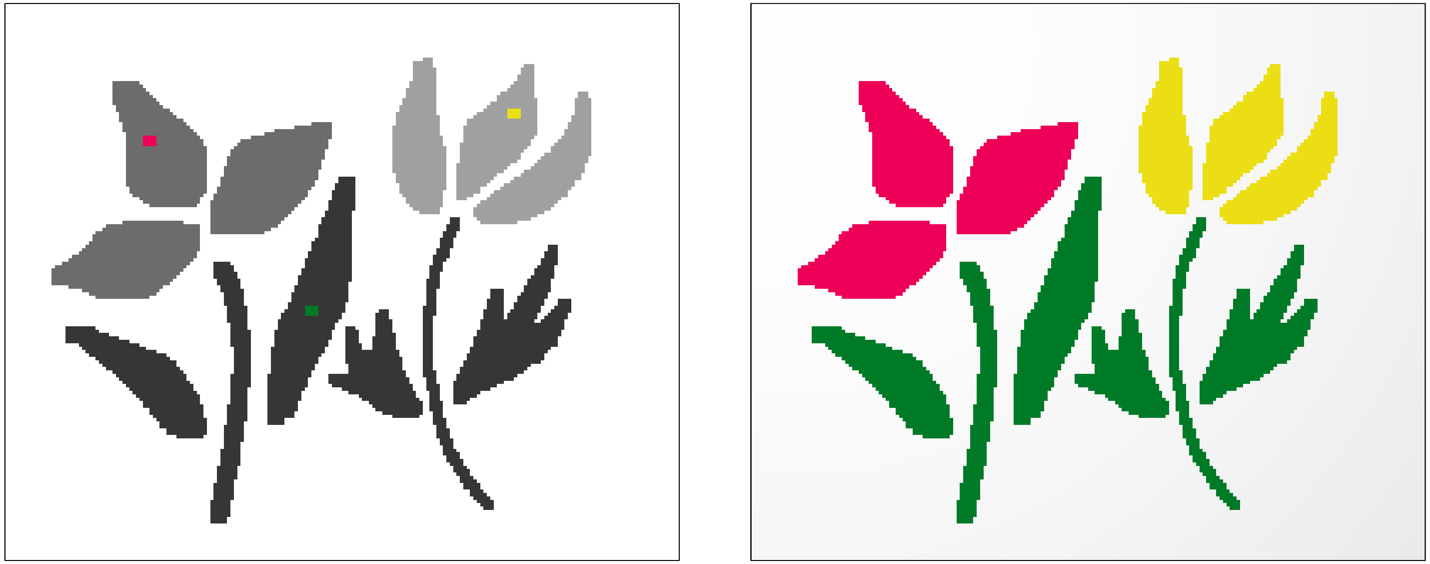


Figure 5: The colorization result with $r = 0$, $p = 2$, $\sigma_1 = 0.001$, and $\sigma_2 = 10$.

Numerical Examples

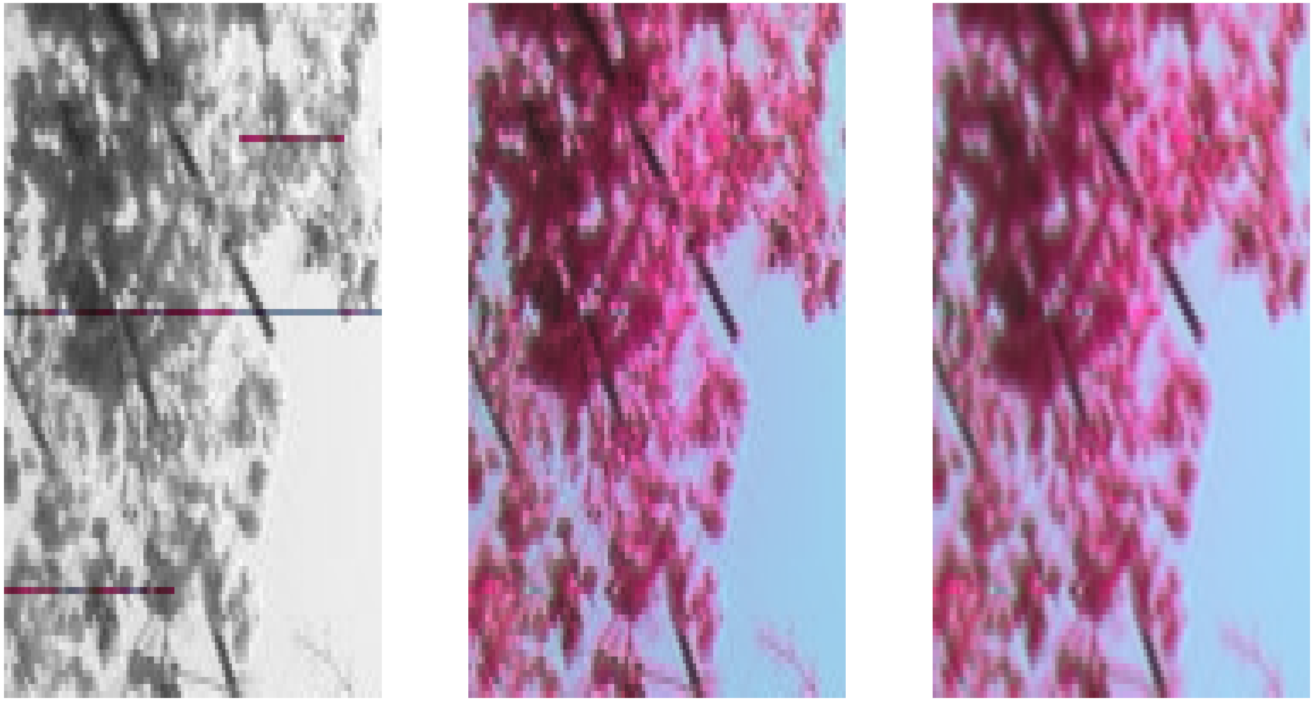


Figure 6: Chromaticity and Brightness model via Stereographic Projection vs. RGB channel: $p = 1$, $r = 2$, $\sigma_1 = 0.5$, and $\sigma_2 = 10$

Numerical Examples

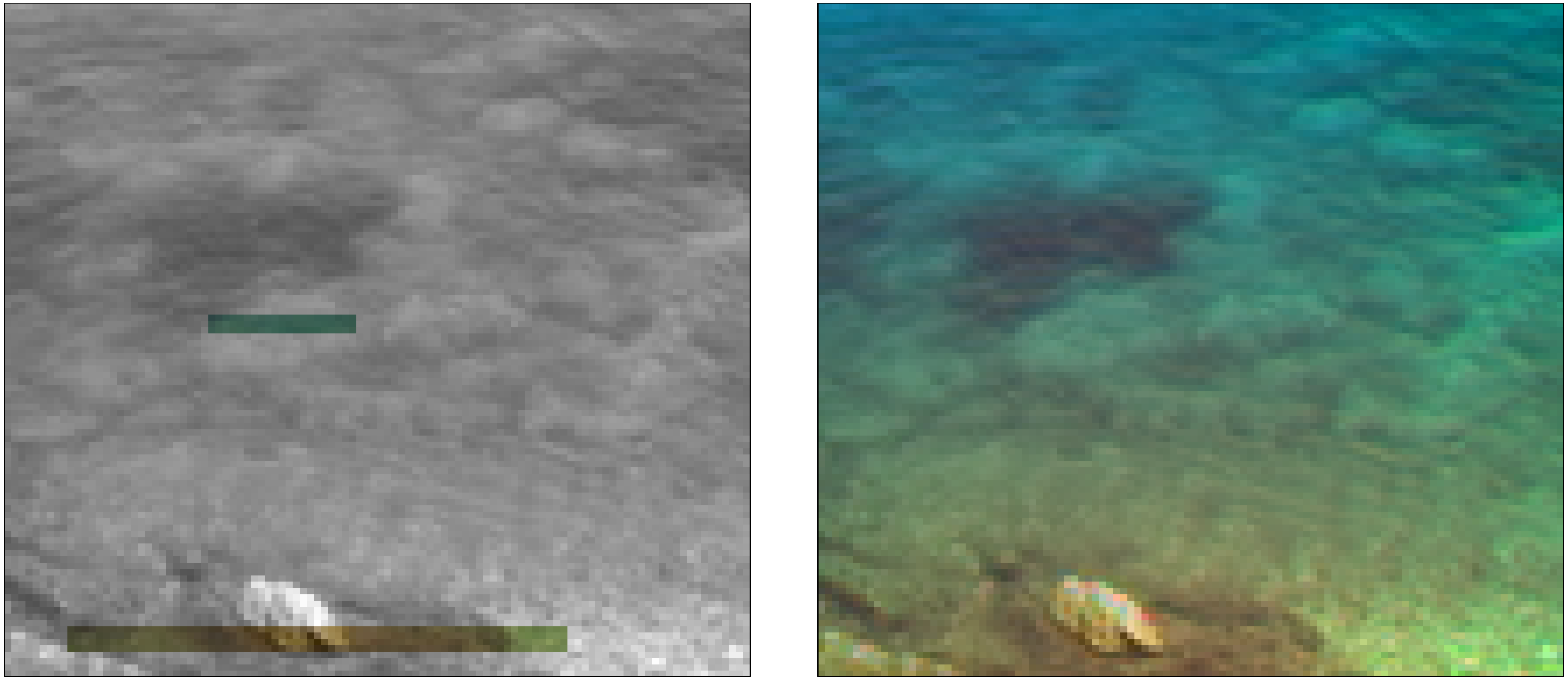


Figure 7: $p = 2$, $r = 2$, $\sigma_1 = 0.1$, and $\sigma_2 = 10$

Numerical Examples

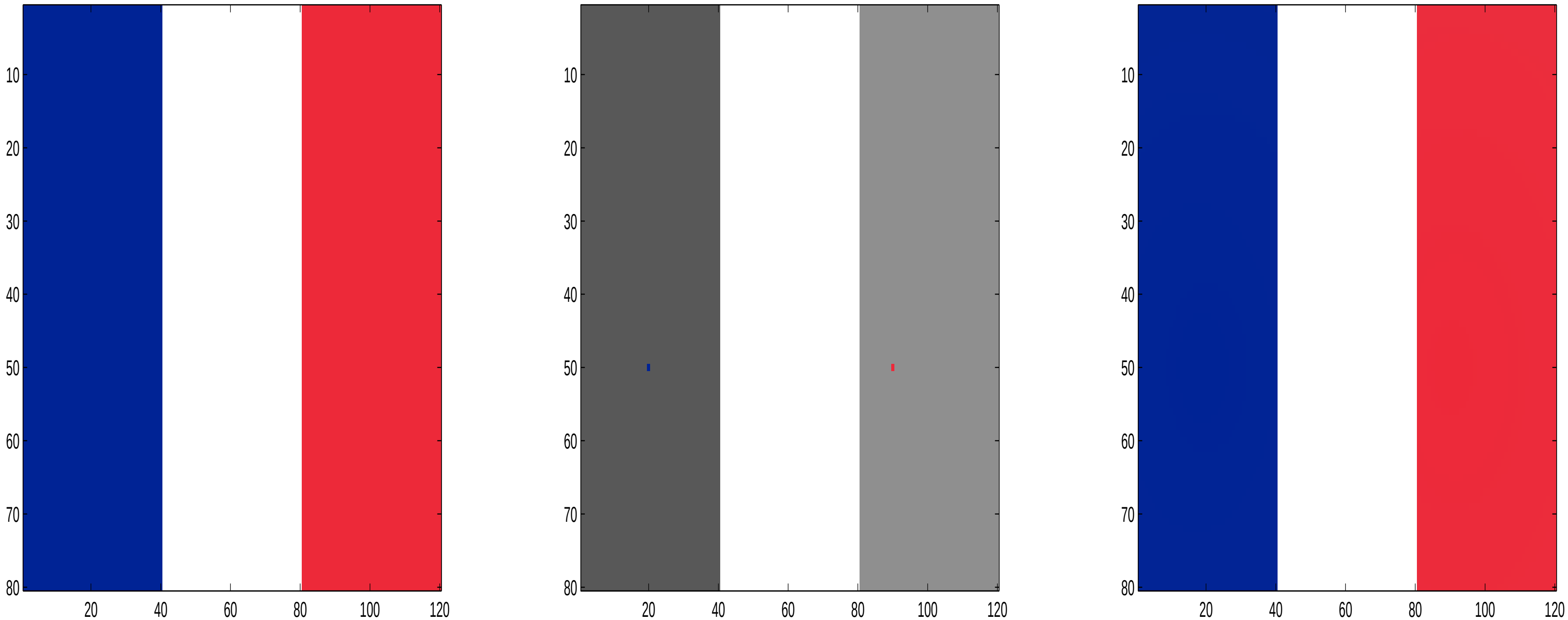


Figure 8: $p = 1$, $r = 0$, $\sigma_1 = 0.05$, $\sigma_2 = 10$.

Numerical Examples

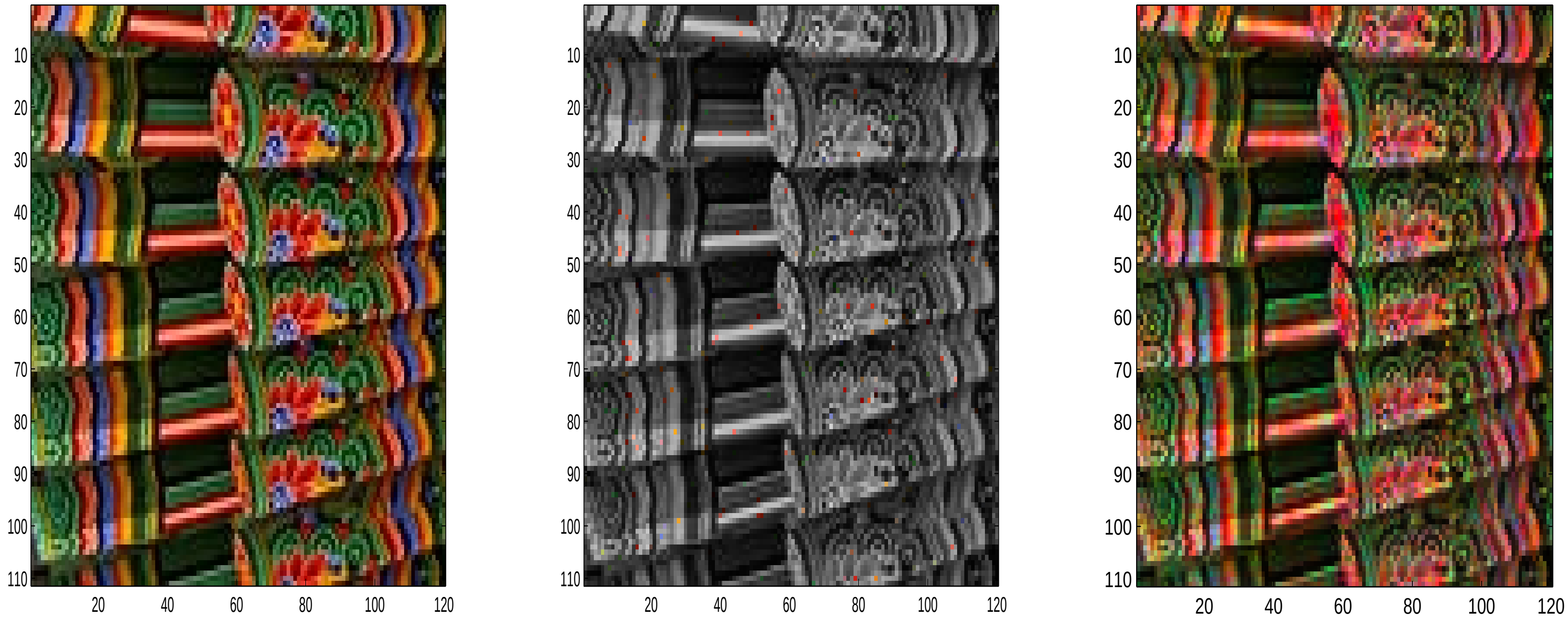


Figure 9: The colorization result with $r = 10$, $p = 1.5$ $\sigma_1 = 0.4$, $\sigma_2 = 10$. Less than 2% of color is given

Conclusion - Main Part

- Operator-valued positive definite kernels and their induced vector-valued RKHS
- Use of RKHS for the problem of function extension (vector-valued)
- An application in Image Colorization
- Full preprint of paper is: Minh Ha Quang, Sung Ha Kang, and Triet Le, *Image and video colorization using vector-valued reproducing kernel Hilbert spaces*, available on my website (or UCLA CAM reports)

Some questions

- Is stereographic projection optimal? More general method?
- How to incorporate geometry of the images (manifold structure)?
- Example: the eye

Outline of the Talk

- Brief Review of Scalar-valued RKHS
- Vector-valued RKHS
- Function Extension
- Application: Image Colorization
- **Learning Theory Estimates** (if time permits)

Error Estimates - Learning Theory

- Input space $X \subset \mathbb{R}^n$ closed (complete separable metric space)
- Output space $Y \subset [-M, M]$ (finite dimensional inner product space)
- $Z = X \times Y$ equipped with an unknown probability measure ρ.
- $\rho(x, y) = \rho_X(x)\rho(y|x)$
- ρ determines a correspondence between X and Y.
- Learning algorithms: find functions $f : X \to Y$ to capture this correspondence.

Least Square Regression

- $\varepsilon_\rho(f) = \int_{X \times Y} (f(x) - y)^2 d\rho$ is minimized by the **regression function**

$$f_\rho(x) = \int_Y y d\rho(y|x)$$

- **Assumption**: $f_\rho \in L^2_{\rho_X}$
- $\varepsilon_\rho(f) - \varepsilon_\rho(f_\rho) = ||f - f_\rho||^2_{L^2_{\rho_X}}$, $f \in L^2_{\rho_X}$
- We want a function $f_{\mathbf{z}}$ that approximates f_ρ in the $||\ ||_{L^2_{\rho_X}}$ norm.

Learning from Sample Data

- ε_ρ is not computable, since ρ is unknown.
- Access to sample $\mathbf{z} = (x_i, y_i)_{i=1}^m \in (X \times Y)^m$, drawn IID according to ρ, thus can construct functions $f_\mathbf{z}$ based on this sample data, to approximate f_ρ or $sgn(f_\rho)$.

Learning Algorithms with Kernel

- Construct a function

$$f_{\mathbf{z},\lambda} = \arg\min_{\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i) + \lambda\Omega(f)$$

  where $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Spaces with norm $|| \; ||_K$, $\lambda > 0$
- $\Omega(f)$ is a regularization term characterizing the smoothness/capacity of f
- Typically $\Omega(f) = ||f||_K^2$.

Examples

- $V(f(x), y) = \max(0, 1 - f(x)y)$: Support Vector Machine
- $V(f(x), y) = (f(x) - y)^2$: Regularized Least Square

Regularized Least Square (RLS)

$$f_{\mathbf{z},\lambda} = \arg\min_{\mathcal{H}_K} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda ||f||_K^2$$

is uniquely given by

$$f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} a_i K(x_i, .)$$

where

$$(K[\mathbf{x}] + m\lambda I)\mathbf{a} = \mathbf{y}$$

with $K[\mathbf{x}] = m \times m$ matrix having entries $K[\mathbf{x}]_{ij} = K(x_i, x_j)$.

Integral Operators induced by Kernels

- Consider $L_K : L^2_\mu \to L^2_\mu$, μ a finite Borel measure, K continuous, positive definite,

$$(L_K f)(x) = \int_X K(x,t) f(t) d\mu(t)$$

- L_K is compact, positive, with eigenvalues $\{\gamma_k\}_{k=0}^\infty$ and eigenfunctions $\{\phi_k\}_{k=0}^\infty$
- $\gamma_{k+1} \leq \gamma_k$ and $\lim_{k\to\infty} \gamma_k = 0$

$$\sum_{k=0}^\infty \gamma_k \leq \kappa^2$$

  where $\kappa^2 = \max_{x\in X} K(x,x)$.
- $\{\phi_k\}_{k=0}^\infty$ form an orthonormal basis in L^2_μ

Integral Operators

- **Mercer's Theorem** (1909): K continuous, positive definite, μ a finite, strictly positive Borel measure on X

$$K(x,t) = \sum_{k=1}^{\infty} \gamma_k \phi_k(x)\phi_k(t)$$

where the convergence is absolute for each pair (x,t) and uniform on compact subsets.

$$\mathcal{H}_K = \{f \in L^2_\mu(X) : ||f||^2_K = \sum_{k=0}^{\infty} \frac{|\langle f, \phi_k \rangle|^2}{\gamma_k} < \infty\}$$

Spectra and Convergence

Theorem 1 *Suppose* $|y| \leq M$ *almost surely. Assume that* $f_\rho \in \mathcal{H}_K$*. Then for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\varepsilon_\rho(f_{\mathbf{z},\lambda_0}) - \varepsilon_\rho(f_\rho) \leq 144(\log \frac{4}{\delta}) \left[M + \kappa ||f_\rho||_K\right]^2 \left(\frac{D(\lambda_0)}{m}\right),$$

where λ_0 *is the unique positive number satisfying*

$$\lambda_0 = 144(\log \frac{4}{\delta}) \left(\frac{M + \kappa ||f_\rho||_K}{||f_\rho||_K}\right)^2 \frac{D(\lambda_0)}{m}$$

$$D(\lambda) = \sum_{k=1}^{\infty} \frac{\gamma_k}{\lambda + \gamma_k} \leq \frac{\kappa^2}{\lambda}$$

Effective Dimensionality

$$D(\lambda_0) \leq \min\{\dim(\mathcal{H}_K), \frac{\sqrt{m}}{12\sqrt{\log \frac{4}{\delta}}}\}$$

- For $\delta = 0.05$ (so that we have a confidence level of 95%), we have

$$D(\lambda_0) \leq \min\{\dim(\mathcal{H}_K), 0.0398\sqrt{m}\},$$

For $m = 1000$ and $m = 1000, 000$, one has

$$D(\lambda_0) \leq \min\{\dim(\mathcal{H}_K), 1.26\}$$

$$D(\lambda_0) \leq \min\{\dim(\mathcal{H}_K), 39.81\}$$

Effective Dimensionality

$$D(\lambda_0) \leq \min\{\dim(\mathcal{H}_K), \frac{\sqrt{m}}{12\sqrt{\log\frac{4}{\delta}}}\}$$

- The order $\sqrt{m}$ for the upper bound is tight.

Convergence Analysis Framework

- Sample Error/Approximation Error Decomposition
- Inverse Problem Formulation
- Law of Large Numbers for Vector-Valued Random Variables

Sample Error and Approximation Error

- Theoretical version of $f_{\mathbf{z},\lambda}$:

$$f_\lambda = \arg\min_{\mathcal{H}_K} \int_Z (f(x) - y)^2 + \lambda ||f||_K^2$$

- Error Decomposition

$$||f_{\mathbf{z},\lambda} - f_\rho||_{L^2_{\rho_X}} \leq ||f_{\mathbf{z},\lambda} - f_\lambda||_{L^2_{\rho_X}} + ||f_\lambda - f_\rho||_{L^2_{\rho_X}}$$

- For $\lambda > 0$ fixed

$$||f_{\mathbf{z},\lambda} - f_\lambda||_{L^2_{\rho_X}} \to 0 \quad \text{as } m \to \infty$$

- As $\lambda \to 0$

$$||f_\lambda - f_\rho||_{L^2_{\rho_X}} \to 0$$

Inverse Problem Formulation

- Solve an ill-posed operator equation

$$Af = F$$

$f \in H_1$, $F \in H_2$, H_1, H_2 Hilbert spaces, by regularization.

- Find f^* that mimimizes

$$||Af - F||_2^2 + \lambda ||f||_1^2$$

- Normal Equation

$$f^* = (A^*A + \lambda I)^{-1} A^* F$$

Inverse Problem Formulation

$$f_{\mathbf{z},\lambda} = \arg\min \quad ||S_{\mathbf{x}} f - \mathbf{y}||^2_{\mathbb{R}^m} + m\lambda||f||^2$$

where $S_{\mathbf{x}} : f \in \mathcal{H}_K \to (f(x_1), \ldots, f(x_m)) \in \mathbb{R}^m$

$$S^*_{\mathbf{x}} : \mathbf{a} \in \mathbb{R}^m \to \sum_{i=1}^m a_i K_{x_i} \in \mathcal{H}_K$$

$$f_\lambda = \arg\min_{\mathcal{H}_K} ||Jf - f_\rho||^2_{L^2_{\rho_X}} + \lambda||f||^2_K$$

where $J : \mathcal{H}_K \to L^2_{\rho_X} =$ inclusion operator and $J^* = L_K : L^2_{\rho_X} \to \mathcal{H}_K$:

$$(L_K f)(t) = \int_Z K(x,t) f(x) d\rho_X(x)$$

Inverse Problem Formulation

$$f_{\mathbf{z},\lambda} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + m\lambda I)^{-1} S_{\mathbf{x}}^* \mathbf{y} = (\frac{1}{m} S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1} \frac{1}{m} S_{\mathbf{x}}^* \mathbf{y}$$

$$\frac{1}{m} S_{\mathbf{x}}^* S_{\mathbf{x}} f = \frac{1}{m} \sum_{i=1}^{m} f(x_i) K_{x_i} = \frac{1}{m} \sum_{i=1}^{m} \langle f, K_{x_i} \rangle_K K_{x_i}$$

$$\frac{1}{m} S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m} \sum_{i=1}^{m} y_i K_{x_i}$$

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$$

$$L_K f = \int_X \langle f, K_x \rangle_K K_x d\rho_X(x)$$

$$L_K f_\rho = \int_Z y K_x d\rho(x, y)$$

Law of Large Numbers

Theorem 2 (Pinelis, 1994) *Let* H *be a Hilbert space with norm* $|| \ ||$ *and* ξ *be a random variable on* (Z, ρ) *with values in* H*. Assume that* $||\xi|| \leq M < \infty$ *almost surely for a fixed constant* $M > 0$*. Let* $\sigma^2(\xi) = E(||\xi||^2)$*. Let* $\{z_i\}_{i=1}^m$ *be independently sampled according to* ρ*. Then for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - E\xi \right\| \leq \frac{2M \log \frac{2}{\delta}}{m} + \sqrt{\frac{2\sigma^2(\xi) \log \frac{2}{\delta}}{m}}.$$

- Apply to estimate $||f_{\mathbf{z},\lambda} - f_\lambda||_{L^2_{\rho_X}}$.

Acknowledgement

- Institute for Pure and Applied Mathematics (IPAM)
- German Research Foundation (DFG)
- Hausdorff Institute for Mathematics (HIM) in Bonn
  (Junior Program in Analysis, September-October 2008)

Thank you

for listening!

Feature Maps

Typical intuition of learning with kernels (for classification):

- Kernels map data **implicitly** into (high dimensional) **feature spaces** via **feature maps**, by Mercer's theorem
- Nonlinearly separable data in input space become linearly separable in feature space
- Linear classifiers are constructed in feature space

Feature Maps via Mercer's Theorem

- Standard feature map in learning literature $\Phi : X \to \ell^2$:

$$\Phi(x) = (\sqrt{\gamma_k}\phi_k(x))_k$$

- Φ depends on the measure μ
- Φ is **not unique**: there is a different map for each measure μ
- Φ is difficult to compute in general

Non-Mercer Feature Maps

- A kernel K on X induces a mapping $\Phi : X \to H_K$

$$\Phi : x \to K_x$$

- By definition of $\langle , \rangle_K$

$$K(x,t) = \langle K_x, K_t \rangle_K = \langle \Phi(x), \Phi(t) \rangle_K$$

- Φ: **feature map**, H_K: **feature space**
- Φ is **explicit**, not **implicit**
- Φ depends only on K and the domain X

Other Non-Mercer Feature Maps

- The map $\Phi : x \to K_x \in H_K$ is universal, true for any positive definite kernel K
- Other maps, for specific kernels:
  - Polynomial kernel $K(x,t) = \langle x, t\rangle^2$

$$\Phi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathbb{R}^3$$

  - Gaussian kernel $K(x,t) = e^{-\frac{||x-t||^2}{\sigma^2}}$

$$\Phi : x \to e^{-\frac{||x||^2}{\sigma^2}}\left(\sqrt{\frac{(2/\sigma^2)^k C_\alpha^k}{k!}}x^\alpha\right)_{|\alpha|=k,k=0}^{\infty} \in \ell^2$$

  - See also Steinwart et al (2005)

Equivalence of Feature Maps

- Invariance of geometry: if $\Phi_1, \Phi_2 : X \to H$ are two feature maps, then for $i = 1, 2$

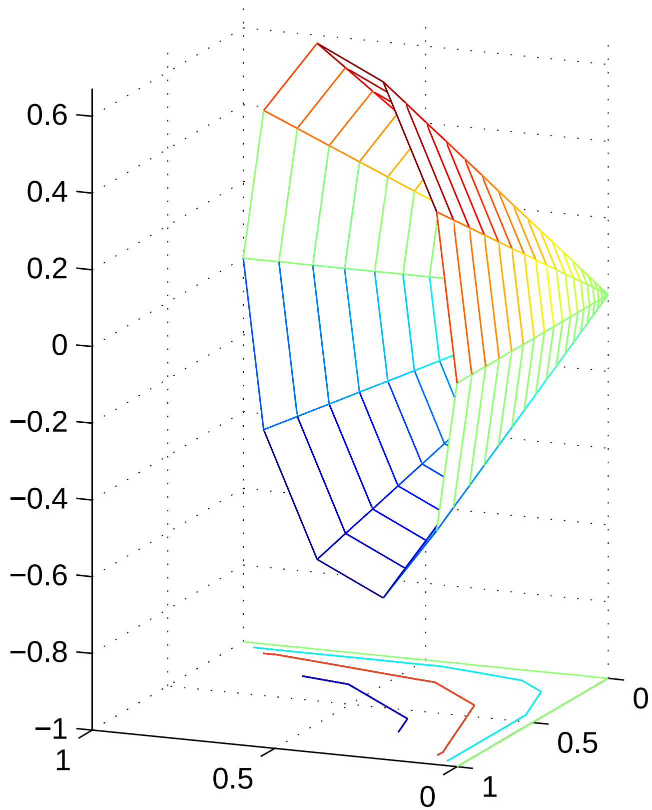
$$||\Phi_i(x) - \Phi_i(t)||^2 = K(x,x) + K(t,t) - 2K(x,t)$$

- Each choice of $\Phi : X \to H_\Phi$ is equivalent to a factorization of $\Phi_K : x \to K_x$

$$\begin{tikzcd}
x \in X \arrow[rr, "\Phi_K"] \arrow[dr, "\Phi"'] & & K_x \in H_K \\
& \Phi_x \in H_\Phi \arrow[ur, "L_\Phi"'] &
\end{tikzcd}$$

Image of Mapped Data

Image of $x_1^2 + x_2^2 \leq 1$ under $\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$



Basic Semi-supervised Learning

- Encounter when we have abundant **unlabeled** data, but not much **labeled** data.
- We wish to utilize the unlabeled data to gain some knowledge of the geometry or underlying marginal distribution of the input data.
- Following material is research carried out by Niyogi, Belkin, Sindhwani, and others.

Basic Semi-supervised Learning

- Labeled data: $(x_i, y_i)_{i=1}^{l}$.
- Unlabeled data: $(x_i)_{i=l+1}^{l+u}$.
- If the input data x_i's actually lie on or close to a low dimensional manifold (in a much higher dimensional ambient space), then we should try to reflect this.
- The new optimization problem is

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \lambda_A ||f||_K^2 + \lambda_I ||f||_I^2.$$

Graph Laplacian

- A major concept from Spectral Graph Theory (see for example Fan Chung's book). From the input data points x_i, one can create a graph.
- W is the weight matrix of the graph.
- D is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$.
- The graph Laplacian is $L = D - W$.
- L has many applications in machine learning.
- If $\mathbf{f} = [f(x_1), \ldots, f(x_{l+u})]$, then

$$\mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij}.$$

Laplacian RLS

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda_A ||f||_K^2 + \frac{\lambda_I}{(l+u)^2} \mathbf{f}^T L \mathbf{f}.$$

The solution has the form

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x).$$

$$\alpha = (JK[\mathbf{x}] + \lambda_A l I + \frac{\lambda_I l}{(l+u)^2} LK[\mathbf{x}])^{-1} \mathbf{y},$$

with $J = diag(1, \ldots, 1, 0, \ldots, 0)$.