Kernels
A Machine Learning Overview

S.V.N. “Vishy” Vishwanathan

vishy@axiom.anu.edu.au

National ICT of Australia
and
Australian National University

Thanks to Alex Smola, Stéphane Canu, Mike Jordan and Peter Bartlett

Overview

- (Really Really) Quick Review of Basics
- Functional Analysis Viewpoint of RKHS
  - Evaluation Functional
  - Kernels and RKHS
  - Mercer's Theorem
- Properties of Kernels
  - Positive Semi-Definiteness
  - Constructing Kernels of RKHS
- Regularization
  - Norm in a RKHS
  - Representer Theorem
  - Fourier Perspective

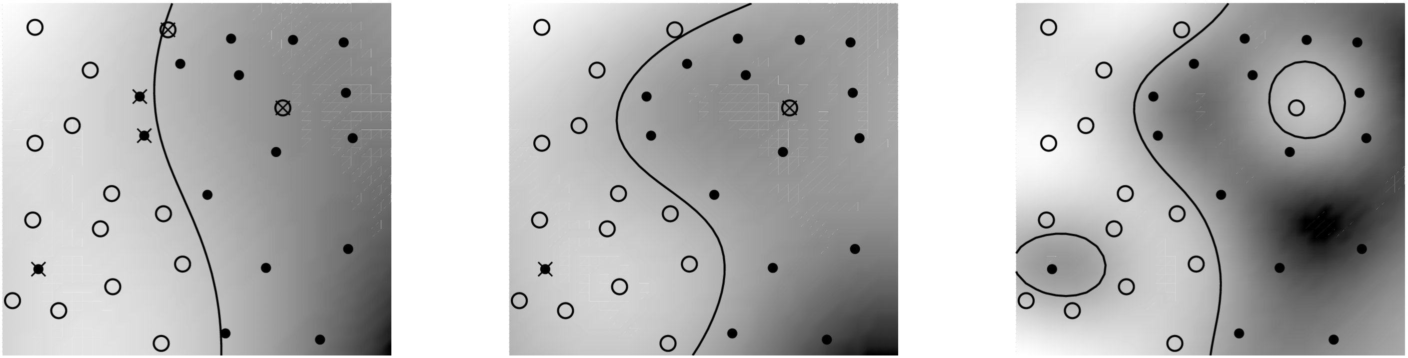
Machine Learning

Data:

- Pairs of observations $(\mathbf{x}_i, y_i)$
- Underlying distribution $\mathrm{P}(\mathbf{x}, y)$
- Examples (blood status, cancer), (transactions, fraud)

Task:

- Find a function $f(\mathbf{x})$ which predicts y given $\mathbf{x}$
- The function $f(\mathbf{x})$ must *generalize* well



What Are Kernels?

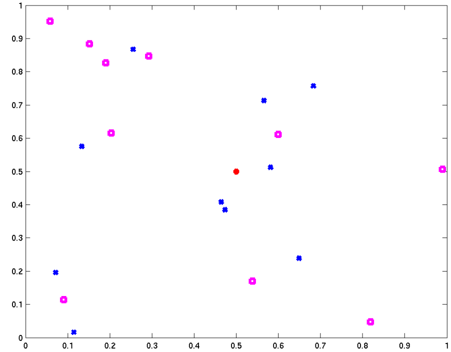
Problem:

- Pairs of observations $(\mathbf{x}_i, y_i)$
- We want to decide the label of point $\mathbf{x}$

Intuition:

- If $k(\mathbf{x}, \mathbf{x}_i)$ is a measure of influence then

$$y(\mathbf{x}) = \sum_i k(\mathbf{x}, \mathbf{x}_i) y_i$$



Linear Functions?

General Form:

- We typically want to use functions of the form

$$f(\mathbf{x}) = \Lambda(\langle \phi(\mathbf{x}), \theta \rangle - g(\theta))$$

- We map data to $\phi(\mathbf{x})$ and then apply Λ to output function

Special Cases:

- For Linear Regression $\Lambda = \mathbf{1}$
- For classification use $\Lambda = \mathrm{sign}$
- For density estimation use $\Lambda = \exp$

The RKHS Connection:

- We need a way to tie this to kernels
- The RKHS setting is suited for this purpose
- We implicitly map data to a high dimensional space

Vector Spaces

Vector Space:

A set $\mathcal{X}$ such that $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\forall \alpha \in \mathbb{R}$ we have

- $\mathbf{x} + \mathbf{y} \in \mathcal{X}$ (**Addition**)
- $\alpha\, \mathbf{x} \in \mathcal{X}$ (**Multiplication**)

Examples:

- Rational numbers $\mathbb{Q}$ over the rational field
- Real numbers $\mathbb{R}$
- Also true for $\mathbb{R}^n$

Counterexamples:

- $f : [0, 1] \to [0, 1]$ does not form a vector space!
- $\mathbb{Z}$ is not a vector space over the real field
- The alphabet $\{\mathrm{a}, \ldots, \mathrm{z}\}$ is not a vector space! (How do you define $+$ and $\times$ operators?)

Banach Spaces

Normed Space:

A pair $(\mathcal{X}, \|\cdot\|)$, where $\mathcal{X}$ is a vector space and $\|\cdot\| : \mathcal{X} \to \mathbb{R}_0^+$ is a normed space if $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$ it satisfies

- $\| \mathbf{x} \| = 0$ if and only if $\mathbf{x} = 0$
- $\|\alpha \mathbf{x} \| = |\alpha| \| \mathbf{x} \|$ **(Scaling)**
- $\| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|$ **(Triangle inequality)**

A norm not satisfying the first condition is called a pseudo norm

Norm and Metric:

A norm induces a metric via $d(\mathbf{x}, \mathbf{y}) := \| \mathbf{x} - \mathbf{y} \|$

Banach Space:

A complete (in the metric defined by the norm) vector space $\mathcal{X}$ together with a norm $\| \cdot \|$

Hilbert Spaces

Inner Product Space:

A pair $(\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$, where $\mathcal{X}$ is a vector space and $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a inner product space if $\forall\, \mathbf{x}, \mathbf{y}\, \mathbf{z} \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$ it satisfies

- $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ **(Additivity)**
- $\langle \alpha\, \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ **(Linearity)**
- $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ **(Symmetry)**
- $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$
- $\langle \mathbf{x}, \mathbf{y} \rangle = 0 \quad \forall\, \mathbf{y} \implies \mathbf{x} = 0$

Dot Product and Norm:

A dot product induces a norm via $\| \mathbf{x} \| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

Hilbert Space:

A complete (in the metric induced by the dot product) vector space $\mathcal{X}$, endowed with a dot product $\langle \cdot, \cdot \rangle$

Hilbert Spaces: Examples

Euclidean Spaces:

Take $\mathbb{R}^m$ endowed with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^m x_i y_i$

ℓ_2 Spaces:

- Infinite series of real numbers
- We define a dot product as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^\infty x_i y_i$

Function Spaces $L_2(\mathcal{X})$:

- A function is square integrable if $\int |f(x)|^2 dx < \infty$
- For square integrable functions $f, g : \mathcal{X} \to \mathbb{R}$ define $\langle f, g \rangle := \int_X f(x) g(x) dx$

Polarization Inequality:

To recover the dot product from the norm compute $\| \mathbf{x} + \mathbf{y} \|^2 - \| \mathbf{x} \|^2 - \| \mathbf{y} \|^2 = 2\langle \mathbf{x}, \mathbf{y} \rangle$

Positive Matrices

Positive Definite Matrix:

A matrix $M \in \mathbb{R}^{m \times m}$ for which for all $\mathbf{x} \in \mathbb{R}^m$ we have

$$\mathbf{x}^\top M \mathbf{x} \geq 0 \text{ if } \mathbf{x} \neq 0$$

This matrix has only positive eigenvalues since for all eigenvectors $\mathbf{x}$ we have $\mathbf{x}^\top M \mathbf{x} = \lambda \mathbf{x}^\top \mathbf{x} = \lambda \| \mathbf{x} \|^2 > 0$ and thus $\lambda > 0$.

Induced Norms and Metrics:

Every positive definite matrix induces a norm via

$$\| \mathbf{x} \|_M^2 := \mathbf{x}^\top M \mathbf{x}$$

- The triangle inequality can be seen by writing

$$\| \mathbf{x} + \mathbf{x}' \|_M^2 = (\mathbf{x} + \mathbf{x}')^\top M^{\frac{1}{2}} M^{\frac{1}{2}} (\mathbf{x} + \mathbf{x}') = \| M^{\frac{1}{2}} (\mathbf{x} + \mathbf{x}') \|^2$$

and using the triangle inequality for $M^{\frac{1}{2}} \mathbf{x}$ and $M^{\frac{1}{2}} \mathbf{x}'$.

Our Setting

Notation:

- Let $\mathcal{X}$ a learning domain and $\mathbb{R}^{\mathcal{X}} := \{f : \mathcal{X} \to \mathbb{R}\}$.
- Hypothesis set $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$

What We Want:

- There are many nasty functions in $\mathbb{R}^{\mathcal{X}}$
- We restrict our attention to *nice* hypothesis sets
- We want to *learn* a function which is *smooth*

Restriction:

- We look at functions of the form

$$\mathcal{H}_0 = \{f(\mathbf{x}) = \sum_{i \in I} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \mathbf{x}_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\}$$

- I is an index set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function

Making Dot Products

Definition:

- Let $g(\mathbf{x}) = \sum_{j \in J} \beta_j k(\mathbf{x}, \mathbf{x}_j)$ then

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j)$$

Properties:

- If k is symmetric then $\langle f, g \rangle_{\mathcal{H}_0} = \langle g, f \rangle_{\mathcal{H}_0}$
- If k is a positive semi definite then $\langle f, f \rangle_{\mathcal{H}_0} \geq 0$

Completion:

- $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ defines a dot-product space
- In order to obtain a Hilbert space $\mathcal{H}$ complete $\mathcal{H}_0$
- $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is *naturally* extended to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

RKHS

Definition:

- For every $f \in \mathcal{H}$ if there is a k such that

$$\langle f(.), k(., \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x})$$

then $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space

Evaluation Functional:

- A linear functional which maps f to $f(\mathbf{x})$

$$\delta_{\mathbf{x}}(f) := f(\mathbf{x})$$

- Observe that $\delta_{\mathbf{x}}$ is linear

Theorem:

- If $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a RKHS $\iff$ $\delta_{\mathbf{x}}$ is continuous
- Equivalently $\delta_{\mathbf{x}}$ is bounded i.e.

$$\forall \mathbf{x} \; \exists M_{\mathbf{x}} \forall f \quad |f(\mathbf{x})| \leq M_{\mathbf{x}} ||f||_{\mathcal{H}}$$

Constructing Kernels

Matrices:

- Let $\mathcal{X} = \{1, \ldots, d\}$, $f(i) = f_i$ $\mathcal{H} = \mathbb{R}^d$, $\langle f, g \rangle_{\mathcal{H}} = f^\top M g$, $M \succeq 0$

$$f = KMf \text{ and } K = M^{-1}$$

n^{th}-Order Polynomials:

- Let $\mathcal{X} = [a, b]$, $\mathcal{H} = \tau_n[a, b]$. Define

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=0}^{n} f^{(i)}(c) g^{(i)}(c) \text{ for } c \in [a, b]$$

then

$$k(x, y) = \sum_{i=0}^{n} \frac{(x - c)^i}{i!} \frac{(y - c)^i}{i!}$$

General Recipe

Define Γ_x:

- Let $c_{\mathbf{x}} \in L_2(\mathcal{X})$. For $f \in L_2(\mathcal{X})$ define

$$\Gamma_{\mathbf{x}}(f) = \langle c_{\mathbf{x}}, f \rangle_{L_2} := g(\mathbf{x})$$

Define Γ:

- Using the pointwise limit above define

$$\Gamma(f) := g$$

Define a RKHS:

- Now let $\mathcal{H} = \mathrm{image}(\Gamma)$ and observe

$$|g(\mathbf{x})| = |\langle c_{\mathbf{x}}, f \rangle_{L_2}| \leq ||c_{\mathbf{x}}||_{L_2} \cdot ||f||_{L_2}$$

- The kernel is

$$k(\mathbf{x}, \mathbf{y}) = \langle c_{\mathbf{x}}, c_{\mathbf{y}} \rangle_{L_2}$$

Mercer's Theorem

Statement:

- Let $k \in L_\infty(\mathcal{X}^2)$ be the kernel of a linear operator

$$T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$$
$$(T_k f)(\mathbf{x}) := \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mu_{\mathbf{y}}$$

such that

$$\int_{\mathcal{X}^2} k(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mu_{\mathbf{y}}\, d\mu_{\mathbf{x}} \geq 0.$$

Let (ψ_j, λ_j) be the normalized eigensystem of k then

- $\lambda_j \in \ell_1$
- Almost everywhere

$$k(\mathbf{x}, \mathbf{y}) = \sum_j \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{y})$$

RKHS from Mercer's Theorem

Construction:

- We define $\mathcal{H} := \{\sum_i c_i \psi_i(\mathbf{x})\}$ and

$$\langle f, g \rangle_{\mathcal{H}} := \sum_i \frac{c_i d_i}{\lambda_i}$$

where ψ_i are eigenfunctions and λ_i are eigenvalues of k

Validity:

- The series c_i^2 / λ_i must converge to 0

Reproducing Property:

- We can check

$$\begin{aligned} \langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_H &= \sum_i c_i \lambda_i \psi_i(\mathbf{x}) / \lambda_i \\ &= \sum_i c_i \psi_i(\mathbf{x}) = f(\mathbf{x}) \end{aligned}$$

Kernels in Practice

Intuition:

- Kernels are measures of similarity
- By the reproducing property

$$\langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{y})$$

- They define a dot product via the map

$$\begin{aligned} \phi : \mathcal{X} &\to \mathcal{H} \\ \mathbf{x} &\mapsto k(\cdot, \mathbf{x}) := \phi(\mathbf{x}) \end{aligned}$$

Why is this Interesting:

- No assumption about the input domain $\mathcal{X}$!!
- Meaningful dot products in $\mathcal{X} \implies$ we are in business
- Kernel methods successfully applied for discrete data
- Strings, trees, graphs, automata, transducers etc.

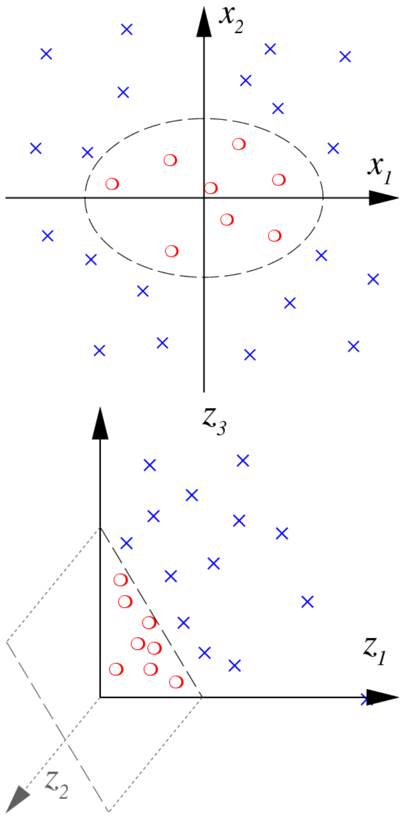
Kernels and Nonlinearity

Problem: Linear functions are often too simple to provide good estimators

Idea 1: Map to a higher dimensional feature space via $\Phi : \mathbf{x} \to \Phi(\mathbf{x})$ and solve the problem there Replace every $\langle \mathbf{x}, \mathbf{y} \rangle$ by $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$

Idea 2: Instead of computing $\Phi(\mathbf{x})$ explicitly use a **kernel function** $k(\mathbf{x}, \mathbf{y}) := \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$

- A large class of functions are admissible as kernels
- Non-vectorial data can be handled if we can compute meaningful $k(\mathbf{x}, \mathbf{y})$



A Few Kernels

Gaussian Kernel:

- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}^+$ then

$$k(\mathbf{x}, \mathbf{y}) := \exp\left(\frac{||\,\mathbf{x} - \mathbf{y}\,||^2}{\sigma^2}\right)$$

Polynomial Kernel:

- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $d \in \mathbb{N}$ then

$$k(\mathbf{x}, \mathbf{y}) := (\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\sigma^2} + 1)^d$$

- Computes all monomials up to degree d

String Kernel:

- Let $x, y \in \mathcal{A}^*$ and w_s be weights

$$k(x, y) := \sum_{s \sqsubseteq x, s' \sqsubseteq y} w_s \delta_{s,s'} = \sum_{s \in \mathcal{A}^*} \#_s(x) \#_s(y) w_s$$

Norms in RKHS

Occams Razor:

- Of all functions which explain data pick the simplest one

Simple Functions:

- We need a way to characterize a *simple* function
- Low function norm in RKHS $\implies$ smooth function

Regularization:

- To encourage simplicity we minimize

$$f_s = \mathrm{argmin}_{f \in H} \frac{1}{m} \sum_{i=1}^{m} c(f(\mathbf{x}_i), y_i) + \lambda ||f||_H^2$$

- $c(\cdot, \cdot)$ is any loss function
- λ is a trade-off parameter

Representer Theorem

Statement:

- Let $c : (\mathcal{X} \times \mathbb{R} \times \mathbb{R})^m \to \mathbb{R} \cup \{\infty\}$ denote a loss function
- Let $\Omega : [0, \infty) \to \mathbb{R}$ be a strictly increasing function
- The objective function (regularized risk) is

$$c((\mathbf{x}_1, \mathbf{y}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_m, \mathbf{y}_m, f(\mathbf{x}_m))) + \Omega(||f||_H)$$

- Each minimizer $f \in H$ of the above admits a representation

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

- The solution is the span of m particular kernels
- Those points for which $\alpha_i > 0$ are Support Vectors

Proof

Sketch:

- Replace $\Omega(||f||_{\mathcal{H}})$ by $\bar{\Omega}(||f||^2_{\mathcal{H}}))$
- Decompose f as

$$f(\mathbf{x}) = f_{||}(\mathbf{x}) + f_{\perp}(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + f_{\perp}(\mathbf{x})$$

- Since $\langle f_{\perp}, k(\mathbf{x}_i, \cdot)\rangle = 0$ we have

$$f(\mathbf{x}_j) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_j)$$

- Now observe that

$$\Omega(||f||_{\mathcal{H}}) \geq \bar{\Omega}\left(||\sum_i \alpha_i k(\mathbf{x}_i, \cdot)||^2_{\mathcal{H}}\right)$$

- The objective function is minimized when $f_{\perp} = 0$

Green's Function

Adjoint Operator:

- Linear operators T and T^* are adjoint if

$$\langle Tf, g \rangle = \langle f, T^* g \rangle$$

- A differential operator is a linear operator

Green's Function:

- Let L linear and k be the kernel of an integral operator
- If $Lk(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y})$ then k is the Green's function of L

Intuition:

- Given f and $Lu = f$ find u
- The Green's function is the kernel of L^{-1}
- You can verify $u = L^{-1} f$ since

$$LL^{-1} f(\mathbf{x}) = L \int k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\,\mathbf{y} = \int \delta(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\,\mathbf{y} = f(\mathbf{x}$$

Kernels are Green's Function

Objective Function:

- Suppose L is the linear differential operator
- We impose smoothing by minimizing

$$c((\mathbf{x}_1, \mathbf{y}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_m, \mathbf{y}_m, f(\mathbf{x}_m))) + ||Lf||^2_{L_2}$$

Define a RKHS:

- Define $\mathcal{H}$ as the completion of $\mathcal{H}_0 = \{f \in L^*LL_2 : ||Lf||_{L_2} < \infty\}$
- The dot product is defined as

$$\langle f, g \rangle_{\mathcal{H}} := \langle Lf, Lg \rangle_{L_2}$$

- Green's function of L^*L is a reproducing kernel for $\mathcal{H}$

Generalization:

- Let L be any linear mapping into a dot product space

Fourier Transform

Definition:

- For $f : \mathbb{R}^n \to \mathbb{R}$ the Fourier transform is

$$\tilde{\mathbf{f}}(\omega) = (2\pi)^{\frac{n}{2}} \int f(\mathbf{x}) \exp(-i\langle \omega, \mathbf{x} \rangle) d\,\mathbf{x}$$

  and the inverse Fourier transform is

$$f(\mathbf{x}) = (2\pi)^{\frac{n}{2}} \int \tilde{\mathbf{f}}(\omega) \exp(i\langle \omega, \mathbf{x} \rangle) d\omega$$

Parseval's Theorem:

- For a function f we have $\langle f, f \rangle_{L_2} = \langle \tilde{\mathbf{f}}, \tilde{\mathbf{f}} \rangle_{L_2}$

Properties:

- For function f and differential operator L we have

$$||Lf||^2 = (2\pi)^{-\frac{n}{2}} \int \frac{|\,\tilde{\mathbf{f}}\,|^2}{\mu(\omega)} d\omega$$

Green's Function

Dot Products:

- We define $\mathcal{H}_0 = \{f : ||Lf||^2 < \infty\}$
- The dot product is defined as

$$\langle f, g \rangle_{\mathcal{H}_0} = (2\pi)^{-\frac{n}{2}} \int \frac{\tilde{\mathbf{f}}(\omega)\overline{\tilde{\mathbf{g}}(\omega)}}{\mu(\omega)} d\omega$$

- The RKHS $\mathcal{H}$ is the completion of $\mathcal{H}_0$

Green's Function:

- We guess the Green's function (kernel) for L^*L as

$$k(\mathbf{x}, \mathbf{y}) = (2\pi)^{-\frac{n}{2}} \int \exp(i\langle \omega, \mathbf{x} - \mathbf{y} \rangle)\mu(\omega) d\omega$$

- Verify that

$$\tilde{k}(\cdot, \mathbf{x}) = \mu(\omega) \exp(-i\langle \omega, \mathbf{x} \rangle)$$

Questions?