Statistical Consistency of Kernel PCA with Random Features

Bharath K. Sriperumbudur

Department of Statistics, Pennsylvania State University

École Polytechnique
CMAP, July 6, 2017

(Joint work with Nicholas Sterge, PSU)

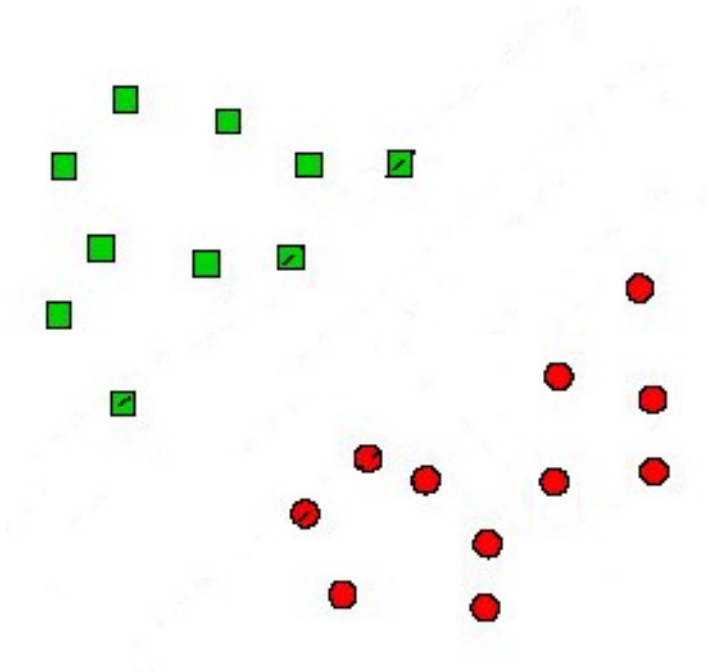
Outline

- Motivating example
  - Binary classification
  - Ridge regression
- Approximation methods
- Kernel PCA
  - Computational vs. statistical trade off

Motivating Example: Binary Classification

- Given: $D := \{(x_j, y_j)\}_{j=1}^{n}$, $x_j \in \mathcal{X}$, $y_j \in \{-1, +1\}$
- Goal: Learn a function $f : \mathcal{X} \to \mathbb{R}$ such that

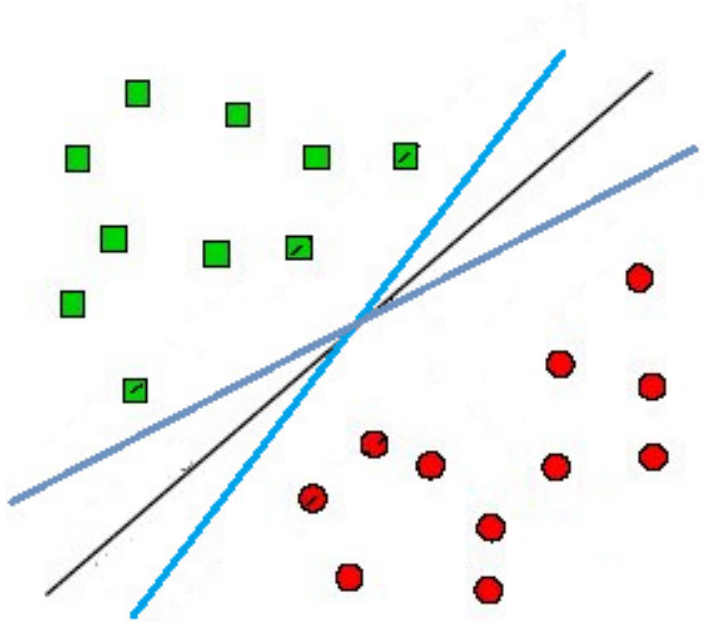
$$y_j = \text{sign}(f(x_j)), \forall\, j = 1, \ldots, n.$$



Linear Classifiers

- Linear classifier: $f_{w,b}(x) = \langle w, x \rangle_2 + b,\ w, x \in \mathbb{R}^d,\ b \in \mathbb{R}$
- Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y_j \left(\langle w, x_j \rangle_2 + b \right) \geq 0,\ \forall j = 1, \ldots, n.$$

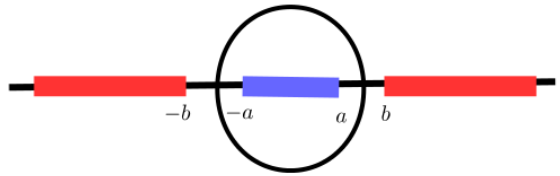


- Fisher discriminant analysis, Support vector machine, Perceptron, ...

Nonlinear Classification: 1

- There is no linear classifier that separates red and blue regions.

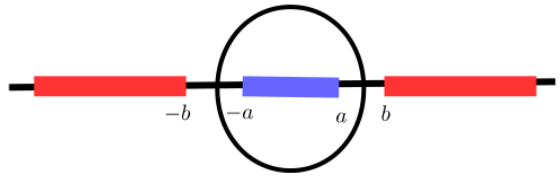
Nonlinear Classification: 1



- There is no linear classifier that separates red and blue regions.
- However, the following function perfectly separates red and blue regions

$$f(x) = x^2 - r = \left\langle \underbrace{(1, -r)}_{w}, \underbrace{(x^2, 1)}_{\Phi(x)} \right\rangle_2 , \; a < r < b.$$

Nonlinear Classification: 1

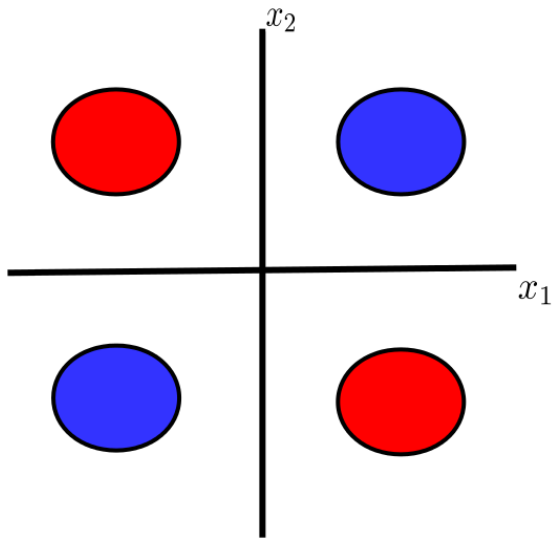


- There is no linear classifier that separates red and blue regions.
- However, the following function perfectly separates red and blue regions

$$f(x) = x^2 - r = \left\langle \underbrace{(1, -r)}_{w}, \underbrace{(x^2, 1)}_{\Phi(x)} \right\rangle_2 , \; a < r < b.$$

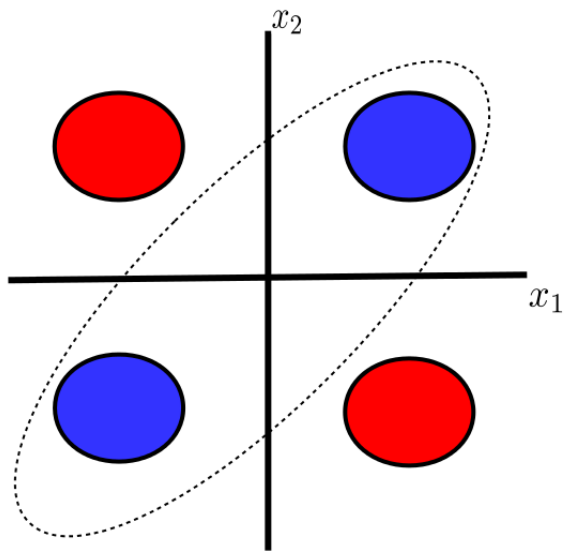
- By mapping $x \in \mathbb{R}$ to $\Phi(x) = (x^2, 1) \in \mathbb{R}^2$, the nonlinear classification problem is turned into a linear problem.
- We call Φ as the feature map.

Nonlinear Classification: 2



- There is no linear classifier that separates red and blue regions.

Nonlinear Classification: 2



- There is no linear classifier that separates red and blue regions.
- A conic section, however, perfectly separates them

$$f(x) = ax_1^2 + bx_1x_2 + cx_2^2 + dx_1 + ex_2 + g$$
$$= \left\langle \underbrace{(a, b, c, d, e, g)}_{w}, \underbrace{(x_1^2, x_1x_2, x_2^2, x_1, x_2, 1)}_{\Phi(x)} \right\rangle_2 .$$

- $\Phi(x) \in \mathbb{R}^6$.

Motivating Example: Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{primal}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Motivating Example: Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{\textit{primal}}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{\textit{dual}}$$

Motivating Example: Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{\textit{primal}}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{\textit{dual}}$$

Motivating Example: Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{\textit{primal}}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{\textit{dual}}$$

Ridge regression

- Prediction: Given $t \in \mathbb{R}^d$

$$f(t) = \langle w, t \rangle_2 = \mathbf{y}^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + n\lambda I_d \right)^{-1} t$$
$$= \mathbf{y}^\top \left(\mathbf{X}^\top \mathbf{X} + n\lambda I_n \right)^{-1} \mathbf{X}^\top t$$

- How does $\mathbf{X}^\top \mathbf{X}$ look like?

$$\mathbf{X}^\top \mathbf{X} = \underbrace{\begin{bmatrix} \langle x_1, x_1 \rangle_2 & \langle x_1, x_2 \rangle_2 & \cdots & \langle x_1, x_n \rangle_2 \\ \langle x_2, x_1 \rangle_1 & \langle x_2, x_2 \rangle_2 & \cdots & \langle x_2, x_n \rangle_2 \\ \vdots & \langle x_i, x_j \rangle_2 & \ddots & \vdots \\ \langle x_n, x_1 \rangle_1 & \langle x_n, x_2 \rangle_2 & \cdots & \langle x_n, x_n \rangle_2 \end{bmatrix}}_{\text{Matrix of inner products: Gram Matrix}}$$

Ridge regression

- Prediction: Given $t \in \mathbb{R}^d$

$$f(t) = \langle w, t \rangle_2 = \mathbf{y}^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + n\lambda I_d \right)^{-1} t$$
$$= \mathbf{y}^\top \left(\mathbf{X}^\top \mathbf{X} + n\lambda I_n \right)^{-1} \mathbf{X}^\top t$$

- How does $\mathbf{X}^\top \mathbf{X}$ look like?

$$\mathbf{X}^\top \mathbf{X} = \underbrace{\begin{bmatrix} \langle x_1, x_1 \rangle_2 & \langle x_1, x_2 \rangle_2 & \cdots & \langle x_1, x_n \rangle_2 \\ \langle x_2, x_1 \rangle_1 & \langle x_2, x_2 \rangle_2 & \cdots & \langle x_2, x_n \rangle_2 \\ \vdots & \langle x_i, x_j \rangle_2 & \ddots & \vdots \\ \langle x_n, x_1 \rangle_1 & \langle x_n, x_2 \rangle_2 & \cdots & \langle x_n, x_n \rangle_2 \end{bmatrix}}_{\text{Matrix of inner products: Gram Matrix}}$$

Kernel Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n} \mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$

$$= \frac{1}{n} \mathbf{y}^\top \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top \Phi(t) = [\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}}]^\top$$

Kernel Ridge regression

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n} \mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$

$$= \frac{1}{n} \mathbf{y}^\top \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top \Phi(t) = [\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}}]^\top$$

Remarks

- Kernel: A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall\, x, x' \in \mathcal{H}.$$

- A complete characterization is provided by Moore-Aronszajn Theorem (Aronszajn, 1950)

  A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.

- The dual formulation is entirely determined by kernel evaluations, Gram matrix and $(k(x_i, t))_i$. But poor scalability: $O(n^3)$.

Remarks

- Kernel: A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall\, x, x' \in \mathcal{H}.$$

- A complete characterization is provided by Moore-Aronszajn Theorem (Aronszajn, 1950)

  A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.

- The dual formulation is entirely determined by kernel evaluations, Gram matrix and $(k(x_i, t))_i$. But poor scalability: $O(n^3)$.

Remarks

- Kernel: A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall\, x, x' \in \mathcal{H}.$$

- A complete characterization is provided by Moore-Aronszajn Theorem (Aronszajn, 1950)

  A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.

- The dual formulation is entirely determined by kernel evaluations, Gram matrix and $(k(x_i, t))_i$. But poor scalability: $O(n^3)$.

Approximation Schemes

- Incomplete Cholesky factorization (e.g., (Fine and Scheinberg, JMLR 2001)
- Sketching (Yang et al., 2015)
- Sparse greedy approximation (Smola and Schölkopf, NIPS 2000)
- Nyström method (e.g., Williams and Seeger, NIPS 2001)
- Random Fourier features (e.g., Rahimi and Recht, NIPS 2008), ...

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x, y) = \psi(x - y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x, y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle \omega, x-y \rangle_2} \, d\Lambda(\omega),$$

  where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.
- Therefore

$$k(x, y) = \int_{\mathbb{R}^d} \cos(\langle \omega, x - y \rangle_2) \, d\Lambda(\omega).$$

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x, y) = \psi(x - y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x, y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle \omega, x-y \rangle_2} \, d\Lambda(\omega),$$

  where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.
- Therefore

$$k(x, y) = \int_{\mathbb{R}^d} \cos(\langle \omega, x - y \rangle_2) \, d\Lambda(\omega).$$

Random Feature Approximation

(Rahimi and Recht, NIPS 2008): Draw $(\omega_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$.

$$k_m(x,y) = \frac{1}{m}\sum_{j=1}^{m} \cos\left(\langle \omega_j, x-y\rangle_2\right) = \langle \Phi_m(x), \Phi_m(y)\rangle_{\mathbb{R}^{2m}},$$

where

$$\Phi_m(x) = (\cos(\langle \omega_1, x\rangle_2), \ldots, \cos(\langle \omega_m, x\rangle_2), \sin(\langle \omega_1, x\rangle_2), \ldots, \sin(\langle \omega_m, x\rangle_2))^\top.$$

- Use the feature map idea with Φ_m, i.e., $x \mapsto \Phi_m(x)$ and apply your favorite linear method.

How good is the approximation?

(S and Szabó, NIPS 2016):

$$\sup_{x,y \in \mathscr{S}} |k_m(x,y) - k(x,y)| = O_{a.s.}\left(\sqrt{\frac{\log|\mathscr{S}|}{m}}\right)$$

Optimal convergence rate

- Other results are known but they are non-optimal (Rahimi and Recht, NIPS 2008; Sutherland and Schneider, UAI 2015).

Ridge Regression: Random Feature Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $w \in \mathbb{R}^{2m}$ s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|_{\mathbb{R}^{2m}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

Ridge Regression: Random Feature Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $w \in \mathbb{R}^{2m}$ s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|_{\mathbb{R}^{2m}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

Ridge Regression: Random Feature Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $w \in \mathbb{R}^{2m}$ s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|_{\mathbb{R}^{2m}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

What happens statistically?

- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of the unknown true regressor, f^*, then

$$\mathcal{R}_{L,\mathbf{P}}(f_{m,n}) - \mathcal{R}_{L,\mathbf{P}}(f^*)$$

achieves the minimax optimal rate as obtained in the case with no approximation. Here L is the squared loss.

Computational gain with no statistical loss!!

- Goal: "Learn" a function $f : X \to Y$ given a set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs drawn independently from an unknown probability distribution $\mathbf{P}$ on $X \times Y$.
- Loss function: $L : Y \times Y \to [0, \infty)$
- Risk functional:

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y).$$

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall\, x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X},\ \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \Big\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \Big\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall\, x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X},\ \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \left\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \right\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.

- Every r.k. is a symmetric and positive definite function.

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall\, x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X},\ \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \left\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \right\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall\, x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X},\ \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \left\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \right\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

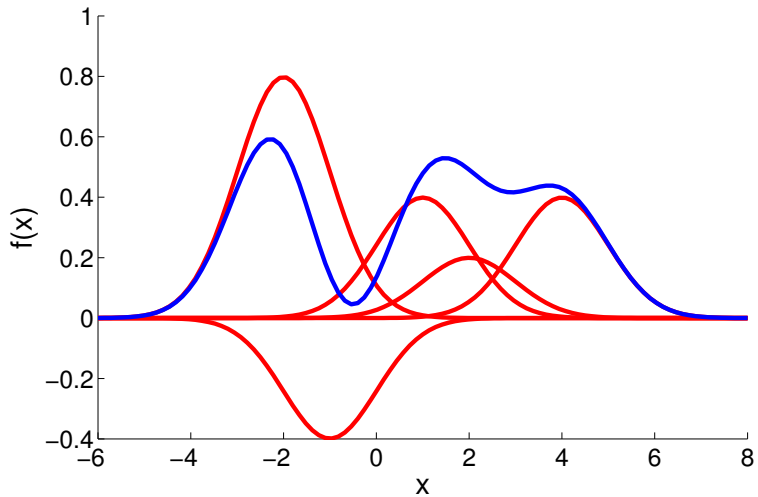
  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Functions in the RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$ (linear span of kernel functions)
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for arbitrary $m \in \mathbb{N}$, $\{\alpha_i\} \subset \mathbb{R}$, $x \in \mathcal{X}$ and $\{x_i\} \subset \mathcal{X}$.

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$



Picture credit: A. Gretton

Properties of RKHS

- k is bounded if and only every $f \in \mathcal{H}$ is bounded.
- If $\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mu(x) < \infty$, then for every $f \in \mathcal{H}$, $\int_{\mathcal{X}} f(x)\, d\mu(x) < \infty$.
- Every $f \in \mathcal{H}$ is continuous if and only if $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$.
- Every $f \in \mathcal{H}$ is m-times continuously differentiable if k is m-times continuously differentiable.

k controls the properties of $\mathcal{H}$

Explicit Realization of RKHS

- $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$ where ψ is a positive definite function.
- Assume ψ satisfies $\int_{\mathbb{R}^d} |\psi(x)|\, dx < \infty$. Denote $\hat{\psi}$ to be the Fourier transform of ψ.
- $$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \Big| \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\psi}(\omega)}\, d\omega < \infty \right\}$$

  endowed with

$$\langle f, g \rangle_{\mathcal{H}} = (2\pi)^{-d/2} \int \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\psi}(\omega)}\, d\omega$$

is an RKHS with k as the r.k.

(Wendland, 2005)

Gaussian RKHS

- Gaussian kernel:

$$k(x, y) = \psi(x - y) = e^{-\|x-y\|_2^2/\gamma^2},\ x, y \in \mathbb{R}^d$$

- Fourier transform:

$$\hat{\psi}(\omega) = \left(\frac{\gamma^2}{2}\right)^{d/2} e^{-\frac{\gamma^2\|\omega\|_2^2}{4}},\ \omega \in \mathbb{R}^d$$

-

$$\mathcal{H}_\gamma(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) : \underbrace{\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{\frac{\gamma^2\|\omega\|_2^2}{4}} d\omega}_{\|f\|_{\mathcal{H}_\gamma}^2} < \infty \right\}$$

Fast decay of $\hat{\psi} \Rightarrow$ Smooth $\mathcal{H}$

Sobolev RKHS

- Laplacian kernel:

$$k(x,y) = \psi(x-y) = \sqrt{\frac{\pi}{2}} e^{-|x-y|},\ x,y \in \mathbb{R}$$

- Fourier transform:

$$\hat{\psi}(\omega) = \frac{1}{1+|\omega|^2},\ \omega \in \mathbb{R}$$

-

$$\mathcal{H}_1^2(\mathbb{R}) := \left\{ f \in L^2(\mathbb{R}) \,:\, \underbrace{\int_{\mathbb{R}} |\hat{f}(\omega)|^2 (1+|\omega|^2)\, d\omega}_{\|f\|^2_{\mathcal{H}_1^2}} < \infty \right\}$$

Extension to $\mathbb{R}^d$: Matérn Kernel

Principal Component Analysis (PCA)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that

$$\mathrm{Var}[\langle w, X \rangle_2]$$

  is maximized.
- Find a direction $w \in \mathbb{R}^d$ such that

$$\mathbb{E}\|(X - \mu) - \langle w, (X - \mu) \rangle_2 w\|_2^2$$

  is minimized.
- The formulations are equivalent and the solution is the eigen vector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Principal Component Analysis (PCA)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that

$$\mathrm{Var}[\langle w, X \rangle_2]$$

  is maximized.
- Find a direction $w \in \mathbb{R}^d$ such that

$$\mathbb{E}\|(X - \mu) - \langle w, (X - \mu) \rangle_2 w\|_2^2$$

  is minimized.
- The formulations are equivalent and the solution is the eigen vector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Principal Component Analysis (PCA)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that

$$\mathrm{Var}[\langle w, X \rangle_2]$$

  is maximized.
- Find a direction $w \in \mathbb{R}^d$ such that

$$\mathbb{E}\|(X - \mu) - \langle w, (X - \mu) \rangle_2 w\|_2^2$$

  is minimized.
- The formulations are equivalent and the solution is the eigen vector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Kernel PCA

- Nonlinear generalization of PCA (Schölkopf et al., 1998).
- $X \mapsto \Phi(X)$ and apply PCA.
- Provides a low dimensional manifold (curves) in $\mathbb{R}^d$ to efficiently represent the data.
- Kernel PCA: Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathbb{E}[f^2(X)] - \mathbb{E}^2[f(X)]$$

- Using the reproducing property $f(X) = \langle f, k(\cdot, X)\rangle_{\mathcal{H}}$, we obtain

$$\begin{aligned}\mathbb{E}[f^2(X)] = \mathbb{E}\langle f, k(\cdot, X)\rangle^2_{\mathcal{H}} &= \mathbb{E}\langle f, (k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X))f\rangle_{\mathcal{H}} \\ &= \langle f, \mathbb{E}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)]f\rangle_{\mathcal{H}},\end{aligned}$$

  assuming $\int_{\mathcal{X}} k(x, x)\, d\mathbb{P}(x) < \infty$.

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ where $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$.

Kernel PCA

- Nonlinear generalization of PCA (Schölkopf et al., 1998).
- $X \mapsto \Phi(X)$ and apply PCA.
- Provides a low dimensional manifold (curves) in $\mathbb{R}^d$ to efficiently represent the data.
- Kernel PCA: Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathbb{E}[f^2(X)] - \mathbb{E}^2[f(X)]$$

- Using the reproducing property $f(X) = \langle f, k(\cdot, X)\rangle_{\mathcal{H}}$, we obtain

$$\begin{aligned}\mathbb{E}[f^2(X)] = \mathbb{E}\langle f, k(\cdot, X)\rangle^2_{\mathcal{H}} &= \mathbb{E}\langle f, (k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X))f\rangle_{\mathcal{H}} \\ &= \langle f, \mathbb{E}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)]f\rangle_{\mathcal{H}},\end{aligned}$$

  assuming $\int_{\mathcal{X}} k(x, x)\, d\mathbb{P}(x) < \infty$.

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ where $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$.

Kernel PCA

- Nonlinear generalization of PCA (Schölkopf et al., 1998).
- $X \mapsto \Phi(X)$ and apply PCA.
- Provides a low dimensional manifold (curves) in $\mathbb{R}^d$ to efficiently represent the data.
- Kernel PCA: Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathbb{E}[f^2(X)] - \mathbb{E}^2[f(X)]$$

- Using the reproducing property $f(X) = \langle f, k(\cdot, X)\rangle_{\mathcal{H}}$, we obtain

$$\begin{aligned}\mathbb{E}[f^2(X)] = \mathbb{E}\langle f, k(\cdot, X)\rangle^2_{\mathcal{H}} &= \mathbb{E}\langle f, (k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X))f\rangle_{\mathcal{H}} \\ &= \langle f, \mathbb{E}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)]f\rangle_{\mathcal{H}},\end{aligned}$$

  assuming $\int_{\mathcal{X}} k(x, x)\, d\mathbb{P}(x) < \infty$.

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ where $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$.

Kernel PCA

- Nonlinear generalization of PCA (Schölkopf et al., 1998).
- $X \mapsto \Phi(X)$ and apply PCA.
- Provides a low dimensional manifold (curves) in $\mathbb{R}^d$ to efficiently represent the data.
- Kernel PCA: Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathbb{E}[f^2(X)] - \mathbb{E}^2[f(X)]$$

- Using the reproducing property $f(X) = \langle f, k(\cdot, X)\rangle_{\mathcal{H}}$, we obtain

$$\begin{aligned}\mathbb{E}[f^2(X)] = \mathbb{E}\langle f, k(\cdot, X)\rangle^2_{\mathcal{H}} &= \mathbb{E}\langle f, (k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X))f\rangle_{\mathcal{H}} \\ &= \langle f, \mathbb{E}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)]f\rangle_{\mathcal{H}},\end{aligned}$$

  assuming $\int_{\mathcal{X}} k(x, x)\, d\mathbb{P}(x) < \infty$.

- $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ where $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$.

Kernel PCA

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \Sigma f \rangle_{\mathcal{H}},$$

where

$$\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) \, d\mathbb{P}(x) - \mu_{\mathbb{P}} \otimes_{\mathcal{H}} \mu_{\mathbb{P}}$$

is the covariance operator (symmetric, positive and Hilbert-Schmidt) on $\mathcal{H}$.

- Spectral theorem:

$$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$$

  where I is either countable ($\lambda_i \to 0$ as $i \to \infty$) or finite.

- Similar to PCA, the solution is the eigen function corresponding to the largest eigenvalue of Σ.

Kernel PCA

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \Sigma f \rangle_{\mathcal{H}},$$

where

$$\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) \, d\mathbb{P}(x) - \mu_{\mathbb{P}} \otimes_{\mathcal{H}} \mu_{\mathbb{P}}$$

is the covariance operator (symmetric, positive and Hilbert-Schmidt) on $\mathcal{H}$.

- Spectral theorem:

$$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$$

  where I is either countable ($\lambda_i \to 0$ as $i \to \infty$) or finite.

- Similar to PCA, the solution is the eigen function corresponding to the largest eigenvalue of Σ.

Empirical Kernel PCA

In practice, $\mathbb{P}$ is unknown but have access to $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

$$\hat{f}^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f \rangle_{\mathcal{H}},$$

where

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) - \mu_n \otimes_{\mathcal{H}} \mu_n$$

is the empirical covariance operator (symmetric, positive and Hilbert-Schmidt) on $\mathcal{H}$ and

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i).$$

.

- $\hat{\Sigma}$ is an infinite dimensional operator but with finite rank (what is its rank?)
- Spectral theorem:

$$\hat{\Sigma} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i.$$

- $\hat{f}^*$ is the eigen function corresponding to the largest eigenvalue of $\hat{\Sigma}$.

Empirical Kernel PCA

In practice, $\mathbb{P}$ is unknown but have access to $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

$$\hat{f}^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f \rangle_{\mathcal{H}},$$

where

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) - \mu_n \otimes_{\mathcal{H}} \mu_n$$

is the empirical covariance operator (symmetric, positive and Hilbert-Schmidt) on $\mathcal{H}$ and

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i).$$

.

- $\hat{\Sigma}$ is an infinite dimensional operator but with finite rank (what is its rank?)
- Spectral theorem:

$$\hat{\Sigma} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i.$$

- $\hat{f}^*$ is the eigen function corresponding to the largest eigenvalue of $\hat{\Sigma}$.

Empirical Kernel PCA

- Since $\hat{\Sigma}$ is an infinite dimensional operator, we are solving an infinite dimensional eigen system,

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i.$$

- How to solve find $\hat{\phi}_i$ in practice?
- Sampling operator: $S : \mathcal{H} \to \mathbb{R}^n,\ f \mapsto \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))$.
- Reconstruction operator: $S^* : \mathbb{R}^n \to \mathcal{H},\ \boldsymbol{\alpha} \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_i k(\cdot, X_i)$.

(Smale and Zhou et al., 2007)

- S^* is the adjoint (transpose) of S and they satisfy

$$\langle f, S^*\boldsymbol{\alpha}\rangle_{\mathcal{H}} = \langle Sf, \boldsymbol{\alpha}\rangle_{\mathbb{R}^n},\ \forall f \in \mathcal{H}, \boldsymbol{\alpha} \in \mathbb{R}^n.$$

- $\hat{\phi}_i = \frac{1}{\hat{\lambda}_i} S^*\mathbf{H}_n\hat{\boldsymbol{\alpha}}_i$ where $\mathbf{H}_n := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\hat{\boldsymbol{\alpha}}_i$ are the eigenvectors of $\mathbf{K}\mathbf{H}_n$ with $\hat{\lambda}_i$ being the corresponding eigenvalues.

Empirical Kernel PCA

- Since $\hat{\Sigma}$ is an infinite dimensional operator, we are solving an infinite dimensional eigen system,

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i \hat{\phi}_i.$$

- How to solve find $\hat{\phi}_i$ in practice?
- Sampling operator: $S : \mathcal{H} \to \mathbb{R}^n$, $f \mapsto \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))$.
- Reconstruction operator: $S^* : \mathbb{R}^n \to \mathcal{H}$, $\boldsymbol{\alpha} \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i k(\cdot, X_i)$.
  (Smale and Zhou et al., 2007)
- S^* is the adjoint (transpose) of S and they satisfy

$$\langle f, S^*\boldsymbol{\alpha} \rangle_{\mathcal{H}} = \langle Sf, \boldsymbol{\alpha} \rangle_{\mathbb{R}^n}, \ \forall f \in \mathcal{H}, \boldsymbol{\alpha} \in \mathbb{R}^n.$$

- $\hat{\phi}_i = \frac{1}{\hat{\lambda}_i} S^* \mathbf{H}_n \hat{\boldsymbol{\alpha}}_i$ where $\mathbf{H}_n := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\hat{\boldsymbol{\alpha}}_i$ are the eigenvectors of $\mathbf{K}\mathbf{H}_n$ with $\hat{\lambda}_i$ being the corresponding eigenvalues.

Empirical Kernel PCA: Random Features

- Empirical kernel PCA solves an $n \times n$ linear system: Complexity is $O(n^3)$.
- Use the idea of random features: $X \mapsto \Phi_m(X)$ and apply PCA.

$$\hat{w}^* := \arg \sup_{\|w\|=1} \widehat{\mathrm{Var}}[\langle w, \Phi_m(X) \rangle_2] = \langle w, \hat{\Sigma}_m w \rangle_2$$

where

$$\hat{\Sigma}_m := \frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \Phi_m^\top(X_i) - \left(\frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \right)^\top$$

is a covariance matrix of size $2m \times 2m$.

- Eigen decomposition: $\hat{\Sigma}_m = \sum_{i=1}^{2m} \hat{\lambda}_{i,m} \hat{\phi}_{i,m} \hat{\phi}_{i,m}^\top$
- $\hat{w}^*$ is obtained by solving a $2m \times 2m$ eigensystem: Complexity is $O(m^3)$.

What happens statistically?

Empirical Kernel PCA: Random Features

- Empirical kernel PCA solves an $n \times n$ linear system: Complexity is $O(n^3)$.
- Use the idea of random features: $X \mapsto \Phi_m(X)$ and apply PCA.

$$\hat{w}^* := \arg \sup_{\|w\|=1} \widehat{\mathrm{Var}}[\langle w, \Phi_m(X) \rangle_2] = \langle w, \hat{\Sigma}_m w \rangle_2$$

where

$$\hat{\Sigma}_m := \frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \Phi_m^\top(X_i) - \left(\frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \Phi_m(X_i) \right)^\top$$

is a covariance matrix of size $2m \times 2m$.

- Eigen decomposition: $\hat{\Sigma}_m = \sum_{i=1}^{2m} \hat{\lambda}_{i,m} \hat{\phi}_{i,m} \hat{\phi}_{i,m}^\top$
- $\hat{w}^*$ is obtained by solving a $2m \times 2m$ eigensystem: Complexity is $O(m^3)$.

What happens statistically?

Empirical Kernel PCA: Random Features

Two notions:

- Reconstruction error (what does this depend on?)
- Convergence of eigenvectors (more generally, eigenspaces)

Eigenspace Convergence

What we have?

- Eigenvectors after approximation, $(\hat{\phi}_{i,m})_{i=1}^{2m}$: these form a subspace in $\mathbb{R}^{2m}$
- Population eigenfunctions $(\phi_i)_{i \in I}$ of Σ: these form a subspace in $\mathcal{H}$.
- How do we compare?
- We embed them in a common space before comparing. The common space is $L^2(\mathbb{P})$.

Eigenspace Convergence

- $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- $\mathfrak{U} : \mathbb{R}^{2m} \to L^2(\mathbb{P})$, $\boldsymbol{\alpha} \mapsto \sum_{i=1}^{2m} \alpha_i \left(\Phi_{m,i} - \int_{\mathcal{X}} \Phi_{m,i}(x)\, d\mathbb{P}(x) \right)$ where

$$\Phi_m = (\cos(\langle \omega_1, \cdot \rangle_2), \ldots, \cos(\langle \omega_m, \cdot \rangle_2), \sin(\langle \omega_1, \cdot \rangle_2), \ldots, \sin(\langle \omega_m, \cdot \rangle_2))^\top.$$

Main results: (S and Sterge, 2017)

- $$\sum_{i=1}^{\ell} \| \mathfrak{I}\hat{\phi}_i - \mathfrak{I}\phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

  Similar result holds in $\| \cdot \|_{\mathcal{H}}$ (Zwald et al., NIPS 2005)
- $$\sum_{i=1}^{\ell} \| \mathfrak{U}\hat{\phi}_{i,m} - \mathfrak{I}\phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right).$$

- Random features are not helpful from the point of view of eigenvector convergence (also holds for eigenspaces).
- May be useful in the reconstruction error convergence (in progress).

Eigenspace Convergence

- $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- $\mathfrak{U} : \mathbb{R}^{2m} \to L^2(\mathbb{P})$, $\boldsymbol{\alpha} \mapsto \sum_{i=1}^{2m} \alpha_i \left(\Phi_{m,i} - \int_{\mathcal{X}} \Phi_{m,i}(x)\, d\mathbb{P}(x) \right)$ where

$$\Phi_m = (\cos(\langle \omega_1, \cdot \rangle_2), \ldots, \cos(\langle \omega_m, \cdot \rangle_2), \sin(\langle \omega_1, \cdot \rangle_2), \ldots, \sin(\langle \omega_m, \cdot \rangle_2))^\top.$$

Main results: (S and Sterge, 2017)

- $$\sum_{i=1}^{\ell} \| \mathfrak{I} \hat{\phi}_i - \mathfrak{I} \phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{n}} \right).$$

  Similar result holds in $\| \cdot \|_{\mathcal{H}}$ (Zwald et al., NIPS 2005)

- $$\sum_{i=1}^{\ell} \| \mathfrak{U} \hat{\phi}_{i,m} - \mathfrak{I} \phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right).$$

- Random features are not helpful from the point of view of eigenvector convergence (also holds for eigenspaces).
- May be useful in the reconstruction error convergence (in progress).

Eigenspace Convergence

- $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- $\mathfrak{U} : \mathbb{R}^{2m} \to L^2(\mathbb{P})$, $\boldsymbol{\alpha} \mapsto \sum_{i=1}^{2m} \alpha_i \left(\Phi_{m,i} - \int_{\mathcal{X}} \Phi_{m,i}(x)\, d\mathbb{P}(x) \right)$ where

$$\Phi_m = (\cos(\langle \omega_1, \cdot \rangle_2), \ldots, \cos(\langle \omega_m, \cdot \rangle_2), \sin(\langle \omega_1, \cdot \rangle_2), \ldots, \sin(\langle \omega_m, \cdot \rangle_2))^\top.$$

Main results: (S and Sterge, 2017)

- $$\sum_{i=1}^{\ell} \| \mathfrak{I} \hat{\phi}_i - \mathfrak{I} \phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{n}} \right).$$

  Similar result holds in $\| \cdot \|_{\mathcal{H}}$ (Zwald et al., NIPS 2005)

- $$\sum_{i=1}^{\ell} \| \mathfrak{U} \hat{\phi}_{i,m} - \mathfrak{I} \phi_i \|_{L^2(\mathbb{P})} = O_p\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right).$$

- Random features are not helpful from the point of view of eigenvector convergence (also holds for eigenspaces).
- May be useful in the reconstruction error convergence (in progress).

References I

Aronszajn, N. (1950).
Theory of reproducing kernels.
Trans. Amer. Math. Soc., 68:337–404.

Fine, S. and Scheinberg, K. (2001).
Efficient SVM training using low-rank kernel representations.
Journal of Machine Learning Research, 2:243–264.

Rahimi, A. and Recht, B. (2008).
Random features for large-scale kernel machines.
In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rudi, A. and Rosasco, L. (2016).
Generalization properties of learning with random features.
https://arxiv.org/pdf/1602.04474.pdf.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998).
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319.

Smale, S. and Zhou, D.-X. (2007).
Learning theory estimates via integral operators and their approximations.
Constructive Approximation, 26:153–172.

Smola, A. J. and Schölkopf, B. (2000).
Sparse greedy matrix approximation for machine learning.
In *Proc. 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, San Francisco, CA.

Sriperumbudur, B. K. and Sterge, N. (2017).
Statistical consistency of kernel PCA with random features.
arXiv:1706.06296.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc.

Sutherland, D. and Schneider, J. (2015).
On the error of random fourier features.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871.

References II

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press, Cambridge, UK.

Williams, C. and Seeger, M. (2001).
Using the Nyström method to speed up kernel machines.
In T. K. Leen, T. G. Diettrich, V. T., editor, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA. MIT Press.

Yang, Y., Pilanci, M., and Wainwright, M. J. (2015).
Randomized sketches for kernels: Fast and optimal non-parametric regression.
Technical report.
`https://arxiv.org/pdf/1501.06195.pdf`.

Zwald, L. and Blanchard, G. (2006).
On the convergence of eigenspaces in kernel principal component analysis.
In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press.