Kernelized algorithms (Lecture 7)

Zoltán Szabó

December 13, 2016

One-page summary

- Until now:
  - Regularized least-squares problems.

$$J(f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \|Lf\|_2^2 \to \min_f,$$

$$f(x) = \sum_{j=1}^{B} c_j \phi_j(x).$$

  - Basis expansion (finite).

One-page summary

- Until now:
  - Regularized least-squares problems.

$$J(f) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \left\| Lf \right\|_2^2 \to \min_f,$$
$$f(x) = \sum_{j=1}^{B} c_j \phi_j(x).$$

  - Basis expansion (finite).
- Today:
  - kernel → kernel ridge regression, kernel PCA.

Kernel

Hilbert space

Definition (Inner product space)

$\mathcal{F}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is an inner product on $\mathcal{F}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{F}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),

Hilbert space

Definition (Inner product space)

$\mathcal{F}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is an inner product on $\mathcal{F}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{F}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),

Hilbert space

Definition (Inner product space)

$\mathcal{F}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is an inner product on $\mathcal{F}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{F}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Hilbert space

Definition (Inner product space)

$\mathcal{F}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is an inner product on $\mathcal{F}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{F}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Notes: 1, 2 $\Rightarrow$ bilinearity.

- Norm induced by the inner product: $\|f\| = \sqrt{\langle f, f \rangle}$.
- CBS: $|\langle f, g \rangle| \leqslant \|f\| \|g\| \Rightarrow \cos(f, g)$.

Hilbert space

Definition (Inner product space)

$\mathcal{F}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is an inner product on $\mathcal{F}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{F}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Notes: 1, 2 $\Rightarrow$ bilinearity.

- Norm induced by the inner product: $\|f\| = \sqrt{\langle f, f \rangle}$.
- CBS: $|\langle f, g \rangle| \leqslant \|f\| \|g\| \Rightarrow \cos(f, g)$.

Definition (Hilbert space)

Nice ('complete') inner product space.

Kernel: inner product of features

Let $\mathcal{X}$ be a set. A $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ function is called kernel if

- there exists a Hilbert space $\mathcal{H}$, and
- $\phi : \mathcal{X} \to \mathcal{H}$ feature map such that

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}}.$$

- $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^D}$, $\phi(x) = [\phi_1(x); \ldots; \phi_B(x)]$.

Kernel: examples ($\mathcal{X} = \mathbb{R}^d$)

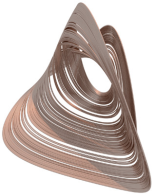
- $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^D}$, $\phi(x) = [\phi_1(x); \ldots; \phi_B(x)]$.
- Polynomial, Gaussian, Laplacian kernel ($\theta > 0, d \in \mathbb{Z}^+$):

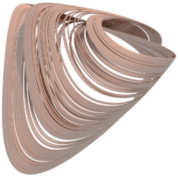
$$k(x, x') = \left(\langle x, x' \rangle + \theta\right)^d, \qquad k(x, x') = e^{-\frac{\|x - x'\|_2^2}{2\theta^2}},$$

$$k(x, x') = e^{-\theta \|x - x'\|_2}.$$

Kernel: example domains ($\mathcal{X}$)

- Euclidean space: $\mathcal{X} = \mathbb{R}^d$.
- Graphs, texts, time series, dynamical systems, probability distributions.





Kernel: properties

Construction from old kernels ($c \geqslant 0$):

- k_1, k_2, k: kernel on $\mathcal{X}$ $\Rightarrow$ $k_1 + k_2$, ck: kernel on $\mathcal{X}$.

Kernel: properties

Construction from old kernels ($c \geqslant 0$):

- k_1, k_2, k: kernel on $\mathcal{X} \Rightarrow k_1 + k_2$, ck: kernel on $\mathcal{X}$.
- Composition: Let $\tilde{k}$ kernel on $\tilde{\mathcal{X}}$, $M : \mathcal{X} \to \tilde{\mathcal{X}}$ mapping. Then

$$k(x, x') = \tilde{k}\left(M(x), M(x')\right)$$

  is kernel on $\mathcal{X}$.

Kernel: properties – continued

Product:

- k_i kernel on $\mathcal{X}_i$ ($i = 1, 2$). Then

$$(k_1 \times k_2)\left((x, y), (x', y')\right) = k_1(x, x')k_2(y, y')$$

is kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.

Product:

- k_i kernel on $\mathcal{X}_i$ ($i = 1, 2$). Then

$$(k_1 \times k_2)\left((x, y), (x', y')\right) = k_1(x, x')k_2(y, y')$$

  is kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.
- For $\mathcal{X} := \mathcal{X}_1 = \mathcal{X}_2$:

$$k(x, x') = k_1(x, x')k_2(x, x')$$

  is kernel on $\mathcal{X}$.

Kernel: properties – Taylor series construction

- Let

$$f(z) = \sum_{j=0}^{\infty} b_j z^j \quad (|z| < r)$$

with $r \in (0, \infty]$, $b_j \geqslant 0$ $(\forall j)$. Then

$$k(x, x') = f\left(\langle x, x' \rangle\right)$$

is kernel with $\|x\| < r$.

Kernel: properties – Taylor series construction

- Let

$$f(z) = \sum_{j=0}^{\infty} b_j z^j \quad (|z| < r)$$

with $r \in (0, \infty]$, $b_j \geqslant 0$ $(\forall j)$. Then

$$k(x, x') = f\left(\langle x, x' \rangle\right)$$

is kernel with $\|x\| < r$.

- Example (exponential kernel, $b_j = \frac{1}{j!}$):

$$f(z) = e^z = \sum_{j=0}^{\infty} \frac{z^j}{j!}, \qquad k(x, x') = e^{\langle x, x' \rangle}.$$

Kernel: properties – usage example

1. Sum, multiplication with a non-negative scalar, product $\Rightarrow$ polynomial kernel.

Kernel: properties – usage example

1. Sum, multiplication with a non-negative scalar, product $\Rightarrow$ polynomial kernel.
2. Gaussian kernel:

$$k(x,x') = e^{-\theta\|x-x'\|_2^2} = e^{-\theta\sum_i \left(x_i-x_i'\right)^2} = \prod_i e^{-\theta(x_i-x_i')^2},$$

Kernel: properties – usage example

1. Sum, multiplication with a non-negative scalar, product $\Rightarrow$ polynomial kernel.
2. Gaussian kernel:

$$k(x,x') = e^{-\theta\|x-x'\|_2^2} = e^{-\theta\sum_i\left(x_i-x_i'\right)^2} = \prod_i e^{-\theta(x_i-x_i')^2},$$

$$k(x,x') = e^{-\theta(x-x')^2} = e^{-\theta\left(x^2+x'^2-2xx'\right)} = \left[e^{-\theta x^2}e^{-\theta x'^2}\right]\times e^{2\theta xx'}$$

by the product rule, composition rule $[M(x) = e^{-\theta x^2}]$ and Taylor-rule the result follows.

Kernel: reproducing view

Reproducing view $\Rightarrow$ elements, kernel trick.

- Let $\mathcal{H}$ be a Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if for $\forall x \in \mathcal{X}$
  1. $k(\cdot, x) \in \mathcal{H}$ ('generators'),

Kernel: reproducing view

Reproducing view $\Rightarrow$ elements, kernel trick.

- Let $\mathcal{H}$ be a Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if for $\forall x \in \mathcal{X}, f \in \mathcal{H}$
  1. $k(\cdot, x) \in \mathcal{H}$ ('generators'),
  2. $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (reproducing property).

Kernel: reproducing view

Reproducing view ⇒ elements, kernel trick.

- Let $\mathcal{H}$ be a Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if for $\forall x \in \mathcal{X}, f \in \mathcal{H}$
  1. $k(\cdot, x) \in \mathcal{H}$ ('generators'),
  2. $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (reproducing property).

  Specifically: $\forall x, y \in \mathcal{X}$,

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

Kernel: positive-definite view

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function.

Kernel: positive-definite view

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function.
- $\mathbf{G} := [k(x_i, x_j)]_{i,j=1}^n$: Gram matrix.

Kernel: positive-definite view

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function.
- $\mathbf{G} := [k(x_i, x_j)]_{i,j=1}^n$: Gram matrix.
- k is called positive definite, if

$$\mathbf{a}^T \mathbf{G} \mathbf{a} \geqslant 0$$

for $\forall n \geqslant 1$, $\forall \mathbf{a} \in \mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$.

Kernel: positive-definite view

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function.
- $\mathbf{G} := [k(x_i, x_j)]_{i,j=1}^n$: Gram matrix.
- k is called positive definite, if

$$\mathbf{a}^T \mathbf{G} \mathbf{a} \geqslant 0$$

for $\forall n \geqslant 1$, $\forall \mathbf{a} \in \mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$.

The different views are *equivalent*!

Convergence in RKHS norm

Convergence in RKHS $\Rightarrow$ uniform convergence! (k: bounded).

Indeed:

$$|f(x)| \overset{k:\text{ r.k.}}{=} |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}| \overset{CBS}{\leqslant} \|k(\cdot,x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$$
$$\overset{k:\text{ r.k.}}{=} \sqrt{k(x,x)}\,\|f\|_{\mathcal{H}}.$$

Convergence in RKHS norm

Convergence in RKHS $\Rightarrow$ uniform convergence! (k: bounded).

Indeed:

$$|f(x)| \overset{k:\text{ r.k.}}{=} |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}| \overset{CBS}{\leqslant} \|k(\cdot,x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \overset{k:\text{ r.k.}}{=} \sqrt{k(x,x)}\, \|f\|_{\mathcal{H}}.$$

Kernel k is called bounded if

$$\sup_{x\in\mathcal{X}} k(x,x) < \infty.$$

Kernel algorithms

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} = \mathcal{H}(k)$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} = \mathcal{H}(k)$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f\in\mathcal{H}}.$$

- Analytical solution:

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)](\mathbf{G} + \lambda n I)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n.$$

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} = \mathcal{H}(k)$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f\in\mathcal{H}}.$$

- Analytical solution:

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)](\mathbf{G} + \lambda n I)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n.$$

Question

How do we get this solution?

Kernel ridge regression

Assume for a moment (representer theorem):

$$f(x) = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Kernel ridge regression

Assume for a moment (representer theorem):

$$f(x) = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{G}\mathbf{a} + \mathbf{a}^T \left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Kernel ridge regression

Assume for a moment (representer theorem):

$$f(x) = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T\mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{G}\mathbf{a} + \mathbf{a}^T\left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I})^{-1}\mathbf{y}$

Kernel ridge regression

Assume for a moment (representer theorem):

$$f(x) = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{G}\mathbf{a} + \mathbf{a}^T [\mathbf{G}^2 + (\lambda n)\mathbf{G}]\mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I})^{-1}\mathbf{y}$ by

$$\frac{\partial \mathbf{a}^T \mathbf{B} \mathbf{a}}{\partial \mathbf{a}} = \left(\mathbf{B} + \mathbf{B}^T\right)\mathbf{a}, \frac{\partial \mathbf{c}^T \mathbf{a}}{\partial \mathbf{a}} = \mathbf{c}.$$

Representer theorem

Let $r : [0, \infty) \to \mathbb{R}$ be monotonically increasing. Then $\exists f \in \mathcal{H}(k)$ minimizer of

$$J(f) = c\left(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)\right) + r\left(\|f\|_{\mathcal{H}}\right)$$

admitting the form

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad a_i \in \mathbb{R}.$$

Representer theorem - proof

Decompose f to $span\left(\{k(\cdot, x_i)\}_{i=1}^n\right)$ and its orthogonal complement:

$$f = f_D + f_\perp, \qquad f_D = \sum_{i=1}^n a_i k(\cdot, x_i).$$

Representer theorem - proof

Decompose f to $span\left(\{k(\cdot,x_i)\}_{i=1}^n\right)$ and its orthogonal complement:

$$f = f_D + f_\perp, \qquad f_D = \sum_{i=1}^n a_i k(\cdot, x_i).$$

Objective terms:

$$\begin{aligned} f(x_i) &= \langle f, k(\cdot,x_i)\rangle_{\mathcal{H}} = \langle f_D + f_\perp, k(\cdot,x_i)\rangle_{\mathcal{H}} \\ &= \langle f_D, k(\cdot,x_i)\rangle_{\mathcal{H}} + \underbrace{\langle f_\perp, k(\cdot,x_i)\rangle_{\mathcal{H}}}_{=0}, \end{aligned}$$

Representer theorem - proof

Decompose f to $span\left(\{k(\cdot,x_i)\}_{i=1}^n\right)$ and its orthogonal complement:

$$f = f_D + f_\perp, \qquad f_D = \sum_{i=1}^n a_i k(\cdot, x_i).$$

Objective terms:

$$\begin{aligned} f(x_i) &= \langle f, k(\cdot,x_i)\rangle_{\mathcal{H}} = \langle f_D + f_\perp, k(\cdot,x_i)\rangle_{\mathcal{H}} \\ &= \langle f_D, k(\cdot,x_i)\rangle_{\mathcal{H}} + \underbrace{\langle f_\perp, k(\cdot,x_i)\rangle_{\mathcal{H}}}_{=0}, \end{aligned}$$

$$r\left(\|f\|_{\mathcal{H}}\right) = r\left(\|f_D + f_\perp\|_H\right) \geqslant r\left(\|f_D\|_{\mathcal{H}}\right).$$

Kernel principal component analysis

Kernel PCA

Let $\mathcal{H} = \mathcal{H}(k)$, $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{H}}$.

- Objective function:

$$J(f) = \frac{1}{n} \sum_{i=1}^{n} \left\langle f, \underbrace{\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)}_{=:\tilde{\phi}(x_i)} \right\rangle^2 = var(f) \to \max_{f:\|f\|_{\mathcal{H}} \leqslant 1}.$$

Kernel PCA

Let $\mathcal{H} = \mathcal{H}(k)$, $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{H}}$.

- Objective function:

$$J(f) = \frac{1}{n} \sum_{i=1}^{n} \left\langle f, \underbrace{\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)}_{=:\tilde{\phi}(x_i)} \right\rangle^2 = var(f) \to \max_{f:\|f\|_{\mathcal{H}} \leqslant 1}.$$

- The solution can be searched in the form ($f \leftrightarrow \mathbf{a}$):

$$f = \sum_{i=1}^{n} a_i \tilde{\phi}(x_i)$$

since component $\perp$ $span\left(\{\phi(x_i)\}_{i=1}^{n}\right)$ has no contribution.

Kernel PCA

Let $\mathcal{H} = \mathcal{H}(k)$, $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{H}}$.

- Objective function:

$$J(f) = \frac{1}{n} \sum_{i=1}^{n} \left\langle f, \underbrace{\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)}_{=:\tilde{\phi}(x_i)} \right\rangle^2 = var(f) \to \max_{f:\|f\|_{\mathcal{H}} \leqslant 1}.$$

- The solution can be searched in the form ($f \leftrightarrow \mathbf{a}$):

$$f = \sum_{i=1}^{n} a_i \tilde{\phi}(x_i)$$

since component $\perp$ $span\left(\{\phi(x_i)\}_{i=1}^{n}\right)$ has no contribution.

- We will get an eigenvalue problem for $\mathbf{a}$.

(Empirical) covariance operator

$$C := \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i).$$

$c \otimes d$ is the analogue of cd^T:

$$(c \otimes d)(e) = c \langle d, e \rangle_{\mathcal{H}} .$$

Similarly to the finite-dimensional case:

$$Cf_j = \lambda_j f_j.$$

Challenge

How do we solve this eigenvalue problem?

Computation of Cf_j

Assume j is fixed ($Cf = \lambda f$):

$$Cf = \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i)\right] f$$

$$\overset{\otimes\text{ def}}{=} \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\left\langle\tilde{\phi}(x_i),\sum_{j=1}^{n}a_j\tilde{\phi}(x_j)\right\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\sum_{j=1}^{n}a_j\tilde{k}(x_i,x_j)$$

with $\tilde{\mathbf{G}} = \mathbf{HGH} = \left[\tilde{k}(x_i,x_j)\right]_{i,j=1}^{n}$, $\mathbf{H} = \mathbf{I}_n - \frac{\mathbf{1}}{n}$.

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}(x_i) \sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\phi}(x_r)$

$$\left\langle \tilde{\phi}(x_r), \lambda f \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_{rj}},$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\sum_{j=1}^{n} a_j \tilde{k}(x_i,x_j)$.
- Idea: multiple by $\tilde{\phi}(x_r)$

$$\left\langle \tilde{\phi}(x_r), \lambda f \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_{rj}},$$

$$\left\langle \tilde{\phi}(x_r), Cf \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_r), \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}(x_i) \sum_{j=1}^{n} a_j \tilde{k}(x_i,x_j) \right\rangle_{\mathcal{H}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{G}_{ri} \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{ij}}_{(\tilde{\mathbf{G}}\mathbf{a})_{ij}} = \frac{1}{n}(\tilde{\mathbf{G}}^2\mathbf{a})_{rj}.$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\sum_{j=1}^{n}a_j\tilde{k}(x_i,x_j)$.
- Idea: multiple by $\tilde{\phi}(x_r)$

$$\left\langle \tilde{\phi}(x_r), \lambda f \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_{rj}},$$

$$\left\langle \tilde{\phi}(x_r), Cf \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_r), \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\sum_{j=1}^{n}a_j\tilde{k}(x_i,x_j) \right\rangle_{\mathcal{H}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{ri}\underbrace{\sum_{j=1}^{n}a_j\tilde{G}_{ij}}_{(\tilde{\mathbf{G}}\mathbf{a})_{ij}} = \frac{1}{n}(\tilde{\mathbf{G}}^2\mathbf{a})_{rj}.$$

- Eigenvalue problem: $\tilde{\mathbf{G}}^2\mathbf{a} = n\lambda\tilde{\mathbf{G}}\mathbf{a}$, i.e. $\tilde{\mathbf{G}}\mathbf{a} = (n\lambda)\mathbf{a}$.

Ortogonal eigenvectors in kernel PCA

Taking two (eigenvector, eigenvalue) pairs:

$$f_1 = \sum_{i=1}^{n} a_{1i}\tilde{\phi}(x_i), \qquad \tilde{\mathbf{G}}\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

$$f_2 = \sum_{j=1}^{n} a_{2j}\tilde{\phi}(x_j), \qquad \tilde{\mathbf{G}}\mathbf{a}_2 = \lambda_2\mathbf{a}_2.$$

one has

$$0 \stackrel{?}{=} \langle f_1, f_2\rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_{1i}\tilde{\phi}(x_i), \sum_{j=1}^{n} a_{2j}\tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} = \mathbf{a}_1^T\tilde{\mathbf{G}}\mathbf{a}_2 = \mathbf{a}_1^T\lambda_2\mathbf{a}_2.$$

Orthogonality $\Rightarrow$ projection is easy

Projection of a new x^* to the first d-PCs:

$$\Pi[\tilde{\phi}(x^*)] = \sum_{j=1}^{d} \left\langle \tilde{\phi}(x^*), f_j \right\rangle_{\mathcal{H}} f_j.$$

Orthogonality $\Rightarrow$ projection is easy

Projection of a new x^* to the first d-PCs:

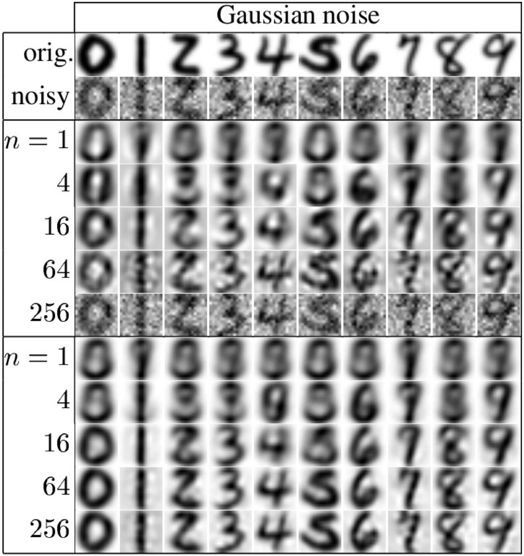
$$\Pi[\tilde{\phi}(x^*)] = \sum_{j=1}^{d} \left\langle \tilde{\phi}(x^*), f_j \right\rangle_{\mathcal{H}} f_j.$$

For fixed $f = f_j$, using $f = \sum_{i=1}^{n} a_i \tilde{\phi}(x_i)$:

$$\left\langle \tilde{\phi}(x^*), f \right\rangle_{\mathcal{H}} f = \sum_i a_i \tilde{k}(x_i, x^*) f = \sum_{i,j=1}^{n} a_i a_j \tilde{k}(x_i, x^*) \tilde{\phi}(x_j).$$

In denoising application

The pre-image problem to solve: $\widehat{x^*} = \arg\min_{x \in \mathcal{X}} \left\| \tilde{\phi}(x) - \Pi[\tilde{\phi}(x^*)] \right\|_{\mathcal{H}}^2$.



Summary

- We covered:
  - basic properties of kernels.
  - kernel ridge regression, representer theorem.
  - kernel PCA.
- References: [4-6].