Gaussian process and reproducing kernel Hilbert space (RKHS)

STATS 303 Statistical Machine Learning

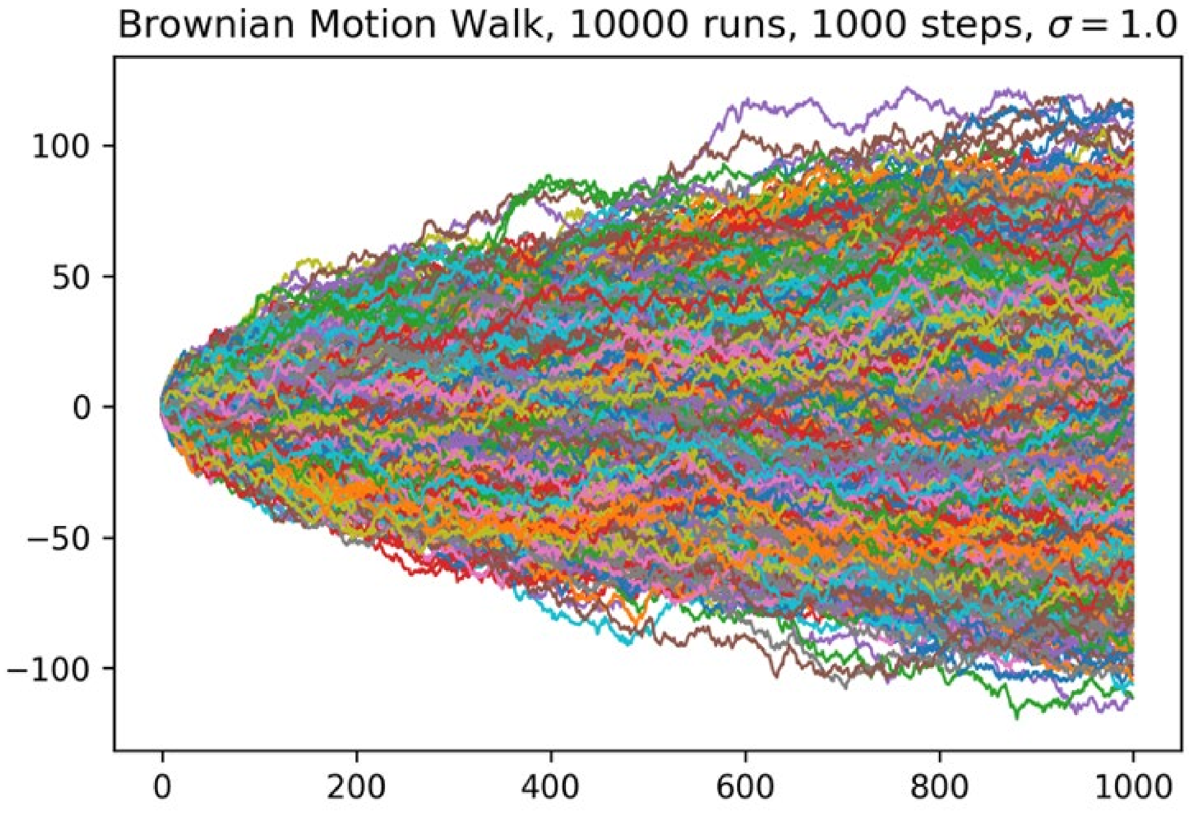
Spring 2022

Lecture 12

Gaussian processes

Gaussian processes

- A Gaussian process (GP) is a probability distribution over functions $y(\mathbf{x})$ such that the set of values of $y(\mathbf{x})$ evaluated at an arbitrary set of points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ jointly have a Gaussian distribution.
- In high-dim, also called a Gaussian random field.
- example: Brownian motion



Gaussian processes

- Fact: a Gaussian process is completely determined by
  - expectation $\mathbb{E}[y(\mathbf{x})] = \mu(\mathbf{x})$
  - covariance $\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}')) = \mathbf{K}(\mathbf{x}, \mathbf{x}')$
- We can denote such a Gaussian process as $\mathcal{GP}(\mu, \mathbf{K})$

Gaussian processes

- Let's stick with the regression case with the model
$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$$
- Consider a prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- Then we get a joint distribution for any collection $y(\mathbf{x}_1), \cdots, y(\mathbf{x}_N)$. Let $\mathbf{y} = (y(\mathbf{x}_1), \cdots, y(\mathbf{x}_N))^{\mathrm{T}}$:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w}$$

$$\begin{aligned}
\mathbb{E}[\mathbf{y}] &= \mathbf{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0} \\
\operatorname{cov}[\mathbf{y}] &= \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\mathbf{\Phi}^{\mathrm{T}} = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{K}
\end{aligned}$$

GP for regression

- One widely used kernel

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m$$

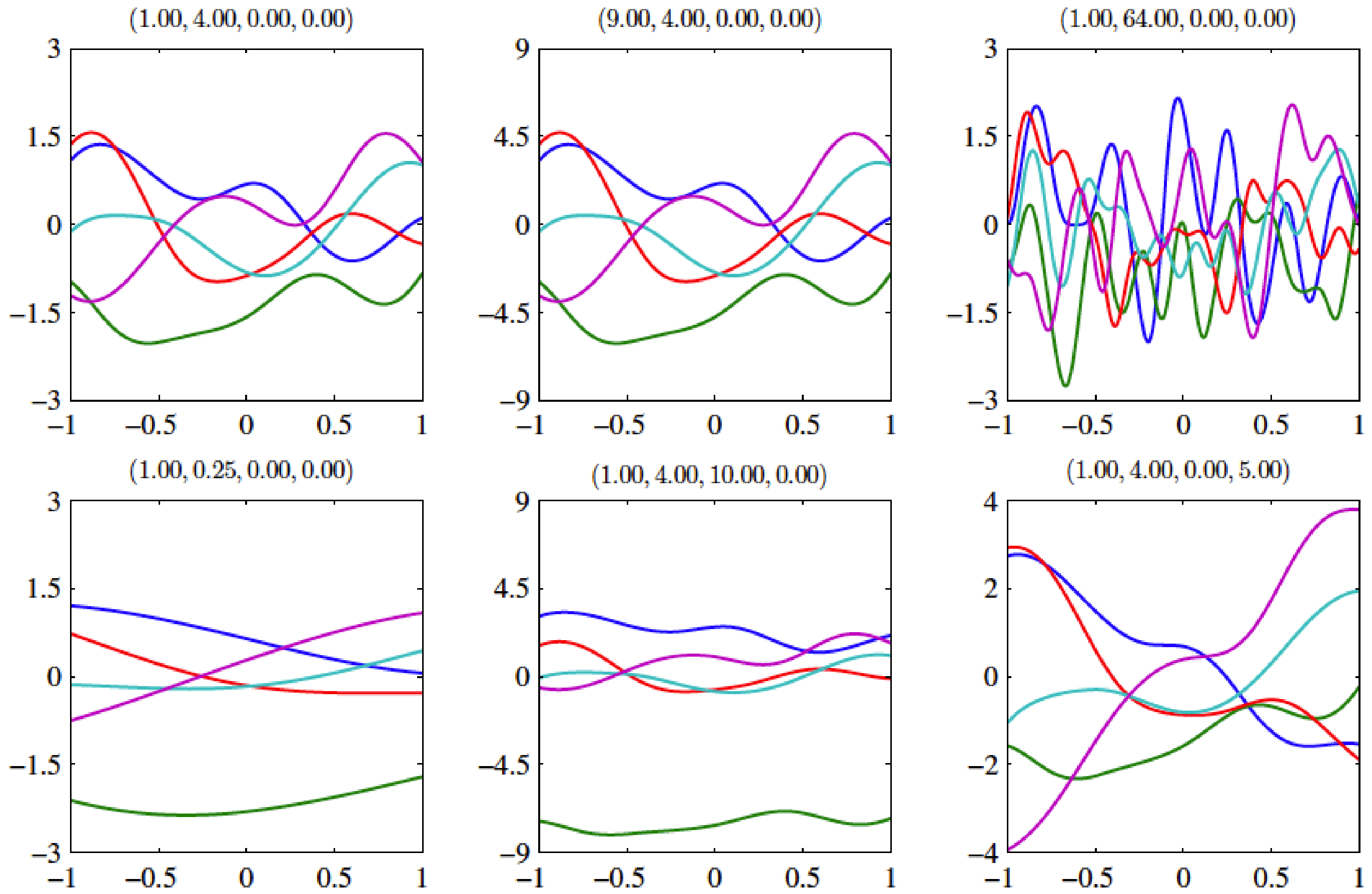


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

GP for regression

- Consider $t_n = y_n + \epsilon_n$ with $p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1})$
- Assume the random variable y follows the GP described above.
- The joint distribution of the target values $\mathbf{t} = (t_1, \cdots, t_N)^{\mathrm{T}}$ conditioned on $\mathbf{y} = (y_1, \cdots, y_N)^{\mathrm{T}}$

$$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}_N)$$

- The prior is, according to the GP,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_N)$$

GP for regression

- By a fact of Gaussian, the marginal distribution of t is

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}_N)$$

where the (n, m)-th entry of $\mathbf{C}_N$ is given by

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$$

- This also holds if we replace N by $N + 1$.

GP for regression

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C})$$

- Consider a test point $\mathbf{x}_{N+1}$. Want to find $p(t_{N+1}|\mathbf{t})$.
- First consider the joint distribution $p(\mathbf{t}_{N+1})$ where

$$\mathbf{t}_{N+1} = (t_1, \cdots, t_N, t_{N+1})^{\mathrm{T}}.$$

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^{\mathrm{T}} & c \end{pmatrix} \quad \boxed{\mathbf{k} \text{ has elements } k(\mathbf{x}_n, \mathbf{x}_{N+1}) \text{ for } n = 1, \ldots, N}$$

$$\boxed{c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}}$$

GP for regression

- Then from analysis of Gaussian, $p(t_{N+1}|\mathbf{t})$ is also Gaussian with mean and variance given by

$$\begin{aligned} m(\mathbf{x}_{N+1}) &= \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} \\ \sigma^2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} \end{aligned}$$

GP for classification*

- Consider a binary classification problem with $t \in \{0,1\}$.
- Define a GP $a(\mathbf{x})$ and then the Bernoulli distribution

$$p(t|a) = \sigma(a)^t(1-\sigma(a))^{1-t}$$

where σ is the sigmoid function $\sigma(a) = \frac{1}{1+e^{-a}}$.

*: optional and not required for exams.

GP for classification *

- Again, consider the training inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and the corresponding targets $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$
- Consider also a test point $\mathbf{x}_{N+1}$ with target value t_{N+1}
- Want to determine $p(t_{N+1}|\mathbf{t})$

GP for classification *

- The Gaussian prior for $\mathbf{a}_{N+1}$, the vector composed of $a(\mathbf{x}_1), \ldots, a(\mathbf{x}_{N+1})$, is

$$p(\mathbf{a}_{N+1}) = \mathcal{N}(\mathbf{a}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

where

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \nu\delta_{nm}$$

- It is sufficient to predict

$$p(t_{N+1} = 1|\mathbf{t}_N) = \int \underbrace{p(t_{N+1} = 1|a_{N+1})}_{\sigma(a_{N+1})} \underbrace{p(a_{N+1}|\mathbf{t}_N)}_{?} \, \mathrm{d}a_{N+1}$$

GP for classification *

- Among all possible solutions to estimate the integral on the previous slide, we introduce the one based on Laplace approximation. Note that

$$\begin{aligned} p(a_{N+1}|\mathbf{t}_N) &= \int p(a_{N+1}, \mathbf{a}_N|\mathbf{t}_N)\, \mathrm{d}\mathbf{a}_N \\ &= \frac{1}{p(\mathbf{t}_N)} \int p(a_{N+1}, \mathbf{a}_N) p(\mathbf{t}_N|a_{N+1}, \mathbf{a}_N)\, \mathrm{d}\mathbf{a}_N \\ &= \frac{1}{p(\mathbf{t}_N)} \int p(a_{N+1}|\mathbf{a}_N) p(\mathbf{a}_N) p(\mathbf{t}_N|\mathbf{a}_N)\, \mathrm{d}\mathbf{a}_N \\ &= \int p(a_{N+1}|\mathbf{a}_N) p(\mathbf{a}_N|\mathbf{t}_N)\, \mathrm{d}\mathbf{a}_N \end{aligned}$$

GP for classification *

$$\boxed{p(a_{N+1}|\mathbf{t}_N) = \int p(a_{N+1}|\mathbf{a}_N)\underbrace{p(\mathbf{a}_N|\mathbf{t}_N)}_{}\, \mathrm{d}\mathbf{a}_N}$$

use a Gaussian approximation

- Also,

$$p(a_{N+1}|\mathbf{a}_N) = \mathcal{N}(a_{N+1}|\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N, c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k})$$

$$p(\mathbf{t}_N|\mathbf{a}_N) = \prod_{n=1}^{N} \sigma(a_n)^{t_n}(1-\sigma(a_n))^{1-t_n} = \prod_{n=1}^{N} e^{a_n t_n}\sigma(-a_n)$$

$$\begin{aligned}\Psi(\mathbf{a}_N) &= \ln p(\mathbf{a}_N) + \ln p(\mathbf{t}_N|\mathbf{a}_N) \\ &= -\frac{1}{2}\mathbf{a}_N^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{C}_N| + \mathbf{t}_N^{\mathrm{T}}\mathbf{a}_N \\ &\quad - \sum_{n=1}^{N}\ln(1+e^{a_n}) + \text{const.}\end{aligned}$$

by Taylor expansion

Exercise: What Gaussian gives a similar log posterior?

reproducing kernel Hilbert space (RKHS)

kernel

- A function $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a kernel over $\mathcal{X}$.

- We want to define kernels that can be written as $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ for any pair of $\mathbf{x}, \mathbf{x}'$ in $\mathcal{X}$.

- Here $\phi: \mathcal{X} \rightarrow \mathbb{F}$ is a mapping where $\mathbb{F}$ is a Hilbert space (inner-product space) called a feature space.

kernel

Theorem [Mercer's Theorem]

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then K admits the following expansion

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{x}')$$

with nonnegative numbers $\{\lambda_j\}_{j=1}^{\infty}$ and orthonormal functions $\psi_j(\cdot): \mathcal{X} \to \mathbb{R}$ if and only if it satisfies the Mercer condition:

$$\int K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

for any integrable function f.

positive definite symmetric kernel

- A function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite symmetric (PDS) if for any $\{\mathbf{x}_n\}_{n=1}^{N} \subset \mathcal{X}$, the matrix $\mathbf{K}$, whose (n, m)-th entry is given by $K(\mathbf{x}_n, \mathbf{x}_m)$, is a symmetric positive semidefinite matrix.

- That is to say, the eigenvalues of $\mathbf{K}$ are all non-negative.

- This $\mathbf{K}$ is called the kernel matrix, or the Gram matrix associated with K.

positive definite symmetric kernel

- (exercise) Let $\mathbf{K}$ be a positive definite symmetric kernel. Then for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$K(\mathbf{x}, \mathbf{x}')^2 \leq K(\mathbf{x}, \mathbf{x}) K(\mathbf{x}', \mathbf{x}').$$

(Hint: Consider the 2-by-2 kernel matrix whose data points are $\mathbf{x}, \mathbf{x}'$. What can you derive given that the eigenvalues of the kernel matrix have to be all nonnegative?)

reproducing kernel Hilbert space (RKHS)

- Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. There exists a Hilbert space (inner-product space) $\mathcal{H}$ and a mapping $\phi: \mathcal{X} \to \mathcal{H}$, such that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$. Furthermore, this $\mathcal{H}$ satisfies the reproducing property: for any $f \in \mathcal{H}$, $f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$.

- This $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS).

reproducing kernel Hilbert space (RKHS)

- In general, a Hilbert space $\mathcal{H}$ is said to be an RKHS if there exists a function $K(\cdot, \cdot): \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

1. $K(\mathbf{x}, \cdot) \in \mathcal{H}$ for any $\mathbf{x} \in \mathcal{X}$
2. $f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$

- Nevertheless, if $\mathcal{H}$ is an RKHS, then $K(\cdot, \cdot)$ must be PDS (why?). That is, there is no need to distinguish RKHS from PDS kernels.

uniqueness of RKHS

- An RKHS contains a unique reproducing kernel.
- Conversely, a reproducing kernel defines a unique RKHS.

review of Mercer's Theorem

Theorem [Mercer's Theorem]

Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then K admits the following expansion

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{x}')$$

with nonnegative numbers $\{\lambda_j\}_{j=1}^{\infty}$ and orthonormal functions $\psi_j(\cdot)\colon \mathcal{X} \to \mathbb{R}$ if and only if it satisfies the <u>Mercer condition</u>:

$$\int K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

for any integrable function f.

This K is a reproducing kernel for $\mathcal{H} = \left\{ f \,\middle|\, \sum_{j=1}^{\infty} \frac{\langle f, \psi_j \rangle^2}{\lambda_j} < \infty \right\}$ such that

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\langle f, \psi_j \rangle \langle g, \psi_j \rangle}{\lambda_j}$$

understanding Mercer's theorem

- Let $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, $\mathbf{K} = [K(\mathbf{x}_n, \mathbf{x}_m)]_{n,m=1}^{N}$ and $\mathbf{f}: \mathcal{X} \to \mathbb{R}^N$ with $\mathrm{f}_n = f(\mathbf{x}_n)$

- Then $\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} \geq 0$ and $\mathbf{K} = \sum \lambda_j \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}}$

- $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{K}_{nm} = (\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}})_{nm} = \sum_j \lambda_j v_{jn} v_{mj} = \sum_j \lambda_j \phi_j(\mathbf{x}_n)\phi_k(\mathbf{x}_m)$

general regularization problem

$$\min_{f \in \mathcal{H}} H[f] = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

Search for the best map f from a RKHS $\mathcal{H}$ which contains all possible solutions.

general regularization problem

$$\min_{f\in\mathcal{H}} H[f] = \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

Search for the best map f from a RKHS $\mathcal{H}$ which contains all possible solutions.

$$\min_{f\in\mathcal{H}} H[f] = \frac{1}{N}\sum_{n=1}^{N} L\left(y_n, \sum_{j=1}^{\infty} c_j\psi_j(x_n)\right) + \lambda\sum_{j=1}^{\infty}\frac{c_j^2}{\gamma_j}$$

representer theorem

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel whose RKHS is $\mathcal{H}$. Then, for any non-decreasing function $G: \mathbb{R} \to \mathbb{R}$ and any loss function L, the optimization problem

$$\min_{f \in \mathcal{H}} H[f] = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) + G(\|f\|_{\mathcal{H}})$$

admits a solution of the form $f^* = \sum_{n=1}^{N} c_n K(x_n, \cdot)$.

representer theorem

- For instance, in SVM, the optimizer is given by

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n K(\mathbf{x}_n, \mathbf{x}) + b$$

Questions?

Reference

- *Gaussian processes:*
    - *[Bi] Ch.6.4.1-6.4.2, 6.4.5*
    - *[HaTF] Ch.5.8.1-5.8.2*
- *RKHS:*
    - *[HaTF] Ch.5.8*