Quantization with Maximum Mean Discrepancies

Are Reproducing Kernel Hilbert Spaces (RKHS) of any use in Stochastic Optimization?

Alois Pichler
with Zahra Mehraban and Paul Dommel

Faculty of Mathematics, TU Chemnitz

ICSP, Paris
July 29, 2025

Mathematik!
TU Chemnitz

Kernels
Densities and bandwidth selection

- The kernel density estimator of the density f is

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} k_h\left(x - X_i\right),$$

where

$$k_h(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right).$$

- Typical examples for kernels k are $k(x) \simeq \max\left\{1 - x^2; 0\right\}$, the Gaussian kernel, or the logistic kernel $k(x) \simeq \frac{1}{(e^x + e^{-x})^2}$.
- The Cauchy kernel $k(x) \simeq \frac{1}{1+x^2}$ does not have moments.

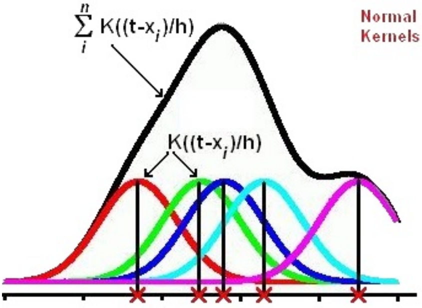


Figure: Normal Kernels

Density estimator
Traditional kernels

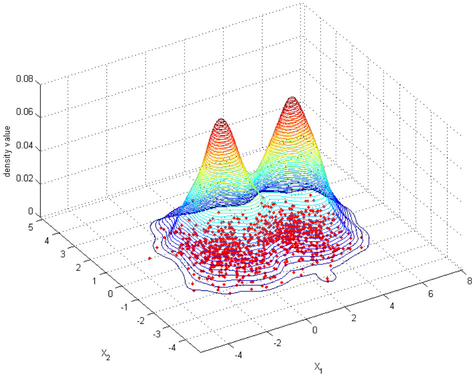


Figure: density

Examples (Density estimation)

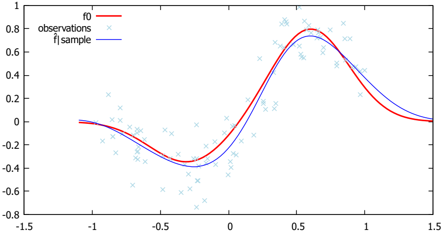
Parzen–Rosenblatt estimator associates with the empirical measure

$$\mu := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

the function

$$\mu_k(x) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, x)$$

in the space $\mathcal{H}_k$.



Problem (Stochastic optimization)

Solve

$$\min_{x\in\mathcal{X}} f_0(x) := \mathbb{E}\, f(x, Y).$$

Motivation (cont.)

Conditional expectation and stochastic optimization

Problem (Optimal control: Hamilton–Jacobi–Bellman)

$$v_t(x) = \sup_u \mathbb{E}\left(\begin{array}{l} c(x, X_{t+1}, u) \\ +\gamma\, v_{t+1}(X_{t+1}) \end{array} \,\middle|\, X_t = x \right);$$

Problem (Time series, learning)

Predict the next X_{t+1}, given the history $X_t, \dots, X_{t-\ell}$.

Outline

Outline

Consider the space $\mathcal{H}$ of functions on $\mathcal{X}$ (usually $\mathcal{X} \subset \mathbb{R}^d$),

$$\begin{aligned} f \colon \mathcal{X} &\to \mathbb{R} \\ z &\mapsto f(z). \end{aligned}$$

In RKHS $(\mathcal{H}, \langle \cdot \mid \cdot \rangle_{\mathcal{H}})$, there is $k_x \in \mathcal{H}$ such that

$$f(x) = \langle f \mid k_x \rangle_{\mathcal{H}} = \langle f(\cdot) \mid k_x(\cdot) \rangle_{\mathcal{H}} \qquad \text{for all} \qquad f \in \mathcal{H},\ x \in \mathcal{X}.$$

We define the kernel

$$\begin{aligned} k \colon \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle k_x \mid k_y \rangle_{\mathcal{H}}. \end{aligned}$$

Consider the space $\mathcal{H}$ of functions on $\mathcal{X}$ (usually $\mathcal{X} \subset \mathbb{R}^d$),

$$\begin{aligned} f\colon \mathcal{X} &\to \mathbb{R} \\ z &\mapsto f(z). \end{aligned}$$

In RKHS $(\mathcal{H}, \langle \cdot \mid \cdot \rangle_{\mathcal{H}})$, there is $k_x \in \mathcal{H}$ such that

$$f(x) = \langle f \,|\, k_x \rangle_{\mathcal{H}} = \langle f(\cdot) |\; k_x(\cdot) \rangle_{\mathcal{H}} \qquad \text{for all} \qquad f \in \mathcal{H},\ x \in \mathcal{X}.$$

We define the kernel

$$\begin{aligned} k\colon \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle k_x \,|\, k_y \rangle_{\mathcal{H}}. \end{aligned}$$

The characterizing property

Consider the space $\mathcal{H}$ of functions on $\mathcal{X}$ (usually $\mathcal{X} \subset \mathbb{R}^d$),

$$\begin{aligned} f\colon \mathcal{X} &\to \mathbb{R} \\ z &\mapsto f(z). \end{aligned}$$

In RKHS $(\mathcal{H}, \langle \cdot \mid \cdot \rangle_{\mathcal{H}})$, there is $k_x \in \mathcal{H}$ such that

$$f(x) = \langle f \,|\, k_x \rangle_{\mathcal{H}} = \langle f(\cdot) |\; k_x(\cdot) \rangle_{\mathcal{H}} \qquad \text{for all} \qquad f \in \mathcal{H},\; x \in \mathcal{X}.$$

We define the kernel

$$\begin{aligned} k\colon \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle k_x |\, k_y \rangle_{\mathcal{H}}. \end{aligned}$$

Examples

A. Berlinet and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US. DOI: 10.1007/978-1-4419-9096-9

A sequence space

$$f=(f_i)_{i=1}^\infty\in\mathbb{R}^{\mathbb{N}},\ \mathcal{X}=\mathbb{N},$$

$$\|f\|:=\sum_{i=1}^\infty\frac{(f_{i+1}-f_i)^2}{\theta^i}$$

$$k(i,j)=\frac{\theta^{\max(i,j)}}{1-\theta}$$

Dirichlet kernel — trigonometric polynomials

$$f(x)=\sum_{k=-N}^N a_k\,e^{ikx},$$

$$\|f\|^2:=\int_{-\pi}^{\pi}f(x)^2\mathrm{d}x$$

$$k(x,y)=\frac{\sin\left(N+\frac12\right)(x-y)}{2\sin\frac{x-y}{2}}$$

Examples

A. Berlinet and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US. DOI: 10.1007/978-1-4419-9096-9

A sequence space

$$f = (f_i)_{i=1}^\infty \in \mathbb{R}^{\mathbb{N}},\ \mathcal{X} = \mathbb{N},$$

$$\|f\| := \sum_{i=1}^\infty \frac{(f_{i+1} - f_i)^2}{\theta^i}$$

$$k(i,j) = \frac{\theta^{\max(i,j)}}{1-\theta}$$

Dirichlet kernel — trigonometric polynomials

$$f(x) = \sum_{k=-N}^{N} a_k\, e^{ikx},$$

$$\|f\|^2 := \int_{-\pi}^{\pi} f(x)^2 \mathrm{d}x$$

$$k(x,y) = \frac{\sin\left(N + \frac{1}{2}\right)(x-y)}{2\sin\frac{x-y}{2}}$$

Examples (cont.)

Nice inner product on $[0,1]$ (Payley–Wiener space)

$f, g \in L^2([-1,1])$ and Fourier transform supported in $[0,1]$

$$\langle f | g \rangle_{\mathcal{H}} := \int_{-1}^{1} f(x) g(x) \,\mathrm{d}x$$

$$k(x,y) = \frac{\sin(t-s)}{t-s}$$

Periodic functions

$$\|f\|_{\mathcal{H}}^2 := \int_0^1 f(x)^2 + f(x)^{(m)2} \mathrm{d}x$$

$$k(x,y) = B_{2m}(|x-y|)$$

where B_{2m} is the Bernoulli polynomial of degree $2m$.

Examples (cont.)

Nice inner product on $[0,1]$ (Payley–Wiener space)

$f, g \in L^2([-1,1])$ and Fourier transform supported in $[0,1]$

$$\langle f | g \rangle_{\mathcal{H}} := \int_{-1}^{1} f(x) g(x) \,\mathrm{d}x$$

$$k(x,y) = \frac{\sin(t-s)}{t-s}$$

Periodic functions

$$\|f\|_{\mathcal{H}}^2 := \int_0^1 f(x)^2 + f(x)^{(m)2} \mathrm{d}x$$

$$k(x,y) = B_{2m}(|x-y|)$$

where B_{2m} is the Bernoulli polynomial of degree $2m$.

Examples (cont.)

Laplace kernel

$f, g \in L^2(\mathbb{R})$,

$$k(x,y) := \frac{\tau}{2} e^{-\tau|x-y|},$$

$$\|f\|_{\mathcal{H}} := \int_{\mathbb{R}} f(x)^2 + \frac{1}{\tau^2} f'(x)^2 \mathrm{d}x$$

Connections to ReLU functions

$$f, g \in L^2((0,\infty))$$

$$\langle f | g \rangle_{\mathcal{H}} := \int_{-1}^{1} f'(x) g'(x) \, \mathrm{d}x$$

$$k(x,y) = \min(x,y)$$

Examples (cont.)

Laplace kernel

$f, g \in L^2(\mathbb{R})$,

$$k(x,y) := \frac{\tau}{2} e^{-\tau|x-y|},$$

$$\|f\|_{\mathcal{H}} := \int_{\mathbb{R}} f(x)^2 + \frac{1}{\tau^2} f'(x)^2 \mathrm{d}x$$

Connections to ReLU functions

$$f, g \in L^2\big((0,\infty)\big)$$

$$\langle f | g \rangle_{\mathcal{H}} := \int_{-1}^{1} f'(x) g'(x) \,\mathrm{d}x$$

$$k(x,y) = \min(x,y)$$

Moore–Aronszajn

k is positive definite

With $k_y := k(\cdot, y)$, $\langle k_x | k_y \rangle_{\mathcal{H}} = \langle k_x | k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$, and

$$\left\langle \sum_{i=1}^{n} \alpha_i k_{x_i} \middle| \sum_{j=1}^{m} \beta_j k_{x_j} \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \cdot k(x_i, x_j),$$

thus $\big(k(x_i, x_j)\big)_{i,j=1}^{n}$ is a positive definite matrix.

Theorem (Moore–Aronszajn)

If k is positive definite, then there is a space $\mathcal{H}_k$ with kernel k.

The Gaussian kernel

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

is a kernel function (for $\ell > 0$ some fixed length parameter).

With $k_y := k(\cdot, y)$, $\langle k_x | k_y \rangle_{\mathcal{H}} = \langle k_x | k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$, and

$$\left\langle \sum_{i=1}^n \alpha_i k_{x_i} \middle| \sum_{j=1}^m \beta_j k_{x_j} \right\rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \cdot k(x_i, x_j),$$

thus $\big(k(x_i, x_j)\big)_{i,j=1}^n$ is a positive definite matrix.

Theorem (Moore–Aronszajn)

If k is positive definite, then there is a space $\mathcal{H}_k$ with kernel k.

The Gaussian kernel

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

is a kernel function (for $\ell > 0$ some fixed length parameter).

With $k_y := k(\cdot, y)$, $\langle k_x | k_y \rangle_{\mathcal{H}} = \langle k_x | k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$, and

$$\left\langle \sum_{i=1}^{n} \alpha_i k_{x_i} \,\middle|\, \sum_{j=1}^{m} \beta_j k_{x_j} \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \cdot k(x_i, x_j),$$

thus $\big(k(x_i, x_j)\big)_{i,j=1}^{n}$ is a positive definite matrix.

Theorem (Moore–Aronszajn)

If k is positive definite, then there is a space $\mathcal{H}_k$ with kernel k.

The Gaussian kernel

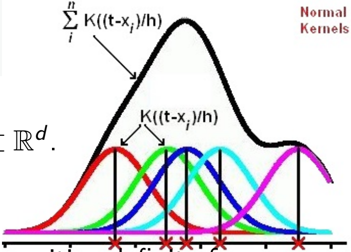
$$k(x, y) := e^{-\|x-y\|^2/\ell^2}$$

is a kernel function (for $\ell > 0$ some fixed length parameter).

Outline

Stochastic Optimization Problems Involve Uncertainties

Embedding probabilities in RKHS



Consider probabilities on $(\mathcal{X}, \Sigma, P)$ with $\mathcal{X} \subset \mathbb{R}^d$.

$$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is a kernel, if every $\boldsymbol{K}$ with $\boldsymbol{K}_{ij} := k(x_i, x_j)$ is positive definite for every $x_i, \dots, x_n \in \mathcal{X}$.

Definition (Embedding)

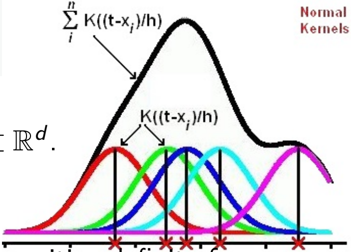
Let μ be a measure on $\mathcal{X}$.
Define the function

$$\mu_k(\cdot) := \int_{\mathcal{X}} k(\cdot, y)\, \mu(\mathrm{d}y).$$

For a probability measure P, then $P_k(\cdot) = \mathbb{E}\, k(\cdot, \xi)$ with $\xi \sim P$.

Stochastic Optimization Problems Involve Uncertainties

Embedding probabilities in RKHS



Consider probabilities on $(\mathcal{X}, \Sigma, P)$ with $\mathcal{X} \subset \mathbb{R}^d$.

$$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is a kernel, if every $\boldsymbol{K}$ with $\boldsymbol{K}_{ij} := k(x_i, x_j)$ is positive definite for every $x_i, \dots, x_n \in \mathcal{X}$.

Definition (Embedding)

Let μ be a measure on $\mathcal{X}$.
Define the function

$$\mu_k(\cdot) := \int_{\mathcal{X}} k(\cdot, y)\, \mu(\mathrm{d}y).$$

Embedding:

$$\iota\colon \mathcal{M}(\mathcal{X}) \hookrightarrow \mathcal{H}_k(\mathcal{X})$$

$$\mu \mapsto \mu_k$$

For a probability measure P, then $P_k(\cdot) = \mathbb{E}\, k(\cdot, \xi)$ with $\xi \sim P$.

Outline

Bounded kernel

We shall typically assume that

$$k(x,x) = \|k(x,\cdot)\|^2_{\mathcal{H}_k} \leq C < \infty, \qquad x \in \mathcal{X}.$$

The embedding function

$$\mu_k(\cdot) = \int_{\mathcal{X}} k(\cdot, y)\,\mu(\mathrm{d}y) \in \mathcal{H}_k(\mathcal{X})$$

then is well defined (as Bochner integral, cf. Diestel and Uhl[1]).

[1] J. Diestel and J. J. Uhl Jr. (1977). *Vector Measures*. Mathematical Surveys 15. American Mathematical Society. URL: http://books.google.com/books?id=EQFjD90fXWAC.

Key property of the space $\mathcal{H}_k$ is its inner product, satisfying

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

Reproducing property
Characterization

Key property of the space $\mathcal{H}_k$ is its inner product, satisfying

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

If $\mu_k(\cdot) = \int_{\mathcal{X}} k(\cdot,y)\,\mu(\mathrm{d}y)$, by linearity

$$\begin{aligned}
\langle k(\cdot,x) | \mu_k(\cdot)\rangle_k &= \int_{\mathcal{X}} \langle k(\cdot,x) | k(\cdot,y)\rangle_{\mathcal{H}_k}\,\mu(\mathrm{d}y) \\
&= \int_{\mathcal{X}} k(x,y)\,\mu(\mathrm{d}y) \\
&= \mu_k(x).
\end{aligned}$$

Maximum Mean Discrepancy

A new distance for probability measures

Definition

The *Maximum Mean Discrepancy* of measures $\mu, \gamma \in \mathcal{M}(\mathcal{X})$ is

$$\mathrm{MMD}(\mu,\gamma) := d_k(\mu,\gamma) := \|\mu_k - \gamma_k\|_{\mathcal{H}_k}.$$

Definition

The *Maximum Mean Discrepancy* of measures $\mu, \gamma \in \mathcal{M}(\mathcal{X})$ is

$$\mathrm{MMD}(\mu,\gamma) := d_k(\mu,\gamma) := \|\mu_k - \gamma_k\|_{\mathcal{H}_k}.$$

- MMD is a distance
- on $\mathcal{M}(\mathcal{X})$, not just $\mathcal{P}(\mathcal{X})$;
- $$\begin{aligned}\mathrm{MMD}(\mu,\eta)^2 &= \langle \mu_k - \gamma_k |\ \mu_k - \gamma_k \rangle_k \\ &= \iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &\qquad - 2\iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\mu(\mathrm{d}x)\gamma(\mathrm{d}y) \\ &\qquad + \iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\gamma(\mathrm{d}x)\gamma(\mathrm{d}y).\end{aligned}$$

Let $f \in \mathcal{H}_k(\mathcal{X})$. Then

$$\begin{aligned}
\mathbb{E}_P f - \mathbb{E}_Q f &= \int_{\mathcal{X}} f(x)\, P(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\, Q(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot)\rangle_k\, P(\mathrm{d}x) - \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot)\rangle_k\, Q(\mathrm{d}x) \\
&= \langle P_k(\cdot) | f(\cdot)\rangle_k - \langle Q_k(\cdot) | f(\cdot)\rangle_k \\
&\le \|P_k - Q_k\|_k \cdot \|f\|_k \qquad (1) \\
&= \|f\|_k \cdot \mathrm{MMD}(P,Q),
\end{aligned}$$

with Cauchy–Schwartz in (1).

Of Interest In Stochastic optimization?
Optimal Transport and MMD

Let $f \in \mathcal{H}_k(\mathcal{X})$. Then

$$\begin{aligned}
\mathbb{E}_P f - \mathbb{E}_Q f &= \int_{\mathcal{X}} f(x)\, P(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\, Q(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \langle k(x,\cdot) |\, f(\cdot)\rangle_k\, P(\mathrm{d}x) - \int_{\mathcal{X}} \langle k(x,\cdot) |\, f(\cdot)\rangle_k\, Q(\mathrm{d}x) \\
&= \langle P_k(\cdot) |\, f(\cdot)\rangle_k - \langle Q_k(\cdot) |\, f(\cdot)\rangle_k \\
&\le \|P_k - Q_k\|_k \cdot \|f\|_k \qquad (1) \\
&= \|f\|_k \cdot \mathrm{MMD}(P, Q),
\end{aligned}$$

with Cauchy–Schwartz in (1).

Of Interest In Stochastic optimization?

Optimal Transport and MMD

Let $f \in \mathcal{H}_k(\mathcal{X})$. Then

$$\begin{aligned}
\mathbb{E}_P f - \mathbb{E}_Q f &= \int_{\mathcal{X}} f(x)\, P(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\, Q(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot) \rangle_k\, P(\mathrm{d}x) - \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot) \rangle_k\, Q(\mathrm{d}x) \\
&= \langle P_k(\cdot) | f(\cdot) \rangle_k - \langle Q_k(\cdot) | f(\cdot) \rangle_k \\
&\leq \| P_k - Q_k \|_k \cdot \| f \|_k \qquad (1) \\
&= \| f \|_k \cdot \mathrm{MMD}(P, Q),
\end{aligned}$$

with Cauchy–Schwartz in (1).

Remark (Wasserstein)

$$\mathbb{E}_P f - \mathbb{E}_Q f \leq Lip(f) \cdot w(P, Q).$$

Let $f \in \mathcal{H}_k(\mathcal{X})$. Then

$$\begin{aligned}
\mathbb{E}_P f - \mathbb{E}_Q f &= \int_{\mathcal{X}} f(x)\, P(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\, Q(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot) \rangle_k\, P(\mathrm{d}x) - \int_{\mathcal{X}} \langle k(x,\cdot) | f(\cdot) \rangle_k\, Q(\mathrm{d}x) \\
&= \langle P_k(\cdot) | f(\cdot) \rangle_k - \langle Q_k(\cdot) | f(\cdot) \rangle_k \\
&\le \| P_k - Q_k \|_k \cdot \| f \|_k \qquad (1) \\
&= \| f \|_k \cdot \mathrm{MMD}(P, Q),
\end{aligned}$$

with Cauchy–Schwartz in (1).

Remark (Wasserstein)

$$\mathbb{E}_P f - \mathbb{E}_Q f \le Lip(f) \cdot w(P, Q).$$

Let $f\in\mathcal{H}_k(\mathcal{X})$. Then

$$\begin{aligned}\mathbb{E}_P f-\mathbb{E}_Q f &= \int_{\mathcal{X}} f(x)\,P(\mathrm{d}x)-\int_{\mathcal{X}} f(x)\,Q(\mathrm{d}x)\\ &= \int_{\mathcal{X}} \langle k(x,\cdot)|\,f(\cdot)\rangle_k\,P(\mathrm{d}x)-\int_{\mathcal{X}} \langle k(x,\cdot)|\,f(\cdot)\rangle_k\,Q(\mathrm{d}x)\\ &= \langle P_k(\cdot)|\,f(\cdot)\rangle_k-\langle Q_k(\cdot)|\,f(\cdot)\rangle_k\\ &\le \|P_k-Q_k\|_k\cdot\|f\|_k \qquad (1)\\ &= \|f\|_k\cdot \mathrm{MMD}(P,Q),\end{aligned}$$

with Cauchy–Schwartz in (1).

Remark (Wasserstein)

$$\mathbb{E}_P f-\mathbb{E}_Q f\le Lip(f)\cdot w(P,Q).$$

Remark

For $x \in \mathcal{X}$,

$$\delta_x(A) := \mathbb{1}_A(x)$$

is the Dirac measure. Its embedding is

$$\delta_{xk}(\cdot) = k(x,\cdot).$$

Remark

For $x, y \in \mathcal{X}$,

$$\begin{aligned}\mathrm{MMD}(\delta_x, \delta_y) \\ &= \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}_k} \\ &=: d_k(x,y)\end{aligned}$$

is a pseudodistance on $\mathcal{X}$.

It holds that

$$\mathrm{MMD}(\delta_x, \delta_y) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}.$$

Remark

For $x \in \mathcal{X}$,

$$\delta_x(A) := \mathbb{1}_A(x)$$

is the Dirac measure. Its embedding is

$$\delta_{xk}(\cdot) = k(x, \cdot).$$

Remark

For $x, y \in \mathcal{X}$,

$$\begin{aligned}\operatorname{MMD}(\delta_x, \delta_y) \\ &= \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}_k} \\ &=: d_k(x,y)\end{aligned}$$

is a pseudodistance on $\mathcal{X}$.

It holds that

$$\operatorname{MMD}(\delta_x, \delta_y) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}.$$

Remark

Recall

$$\mathrm{MMD}(\delta_x,\delta_y)=\sqrt{k(x,x)-2k(x,y)+k(y,y)}.$$

For the Gaussian kernel $k(x,y)=e^{-\|x-y\|^2/\ell^2}$,

$$d(x,y)=\sqrt{2-2e^{-\|x-y\|^2/\ell^2}}\le\frac{\sqrt{2}}{\ell}\cdot\|x-y\|.$$

For the Laplacian kernel $k(x,y)=e^{-\|x-y\|/\ell}$,

$$d(x,y)=\sqrt{2-2e^{-\|x-y\|/\ell}}\le\sqrt{\frac{2}{\ell}}\cdot\|x-y\|^{1/2}.$$

Remark

Recall

$$\mathrm{MMD}(\delta_x,\delta_y)=\sqrt{k(x,x)-2k(x,y)+k(y,y)}.$$

For the Gaussian kernel $k(x,y)=e^{-\|x-y\|^2/\ell^2}$,

$$d(x,y)=\sqrt{2-2e^{-\|x-y\|^2/\ell^2}}\le\frac{\sqrt{2}}{\ell}\cdot\|x-y\|.$$

For the Laplacian kernel $k(x,y)=e^{-\|x-y\|/\ell}$,

$$d(x,y)=\sqrt{2-2e^{-\|x-y\|/\ell}}\le\sqrt{\frac{2}{\ell}}\cdot\|x-y\|^{1/2}.$$

Optimal transport and MMD
Relation

Proposition

Let P and Q be probability measures. Suppose that

$$k(x,x) - 2k(x,y) + k(y,y) \le c^2 \cdot d(x,y)^{2\alpha}, \qquad x, y \in \mathcal{X},$$

for some $c > 0$ and $\alpha \ge 1/2$. Then it holds that[a]

$$\mathrm{MMD}(P,Q) = \|P_k - Q_k\|_{\mathcal{H}_k} \le c \cdot w_\alpha(P,Q)^\alpha.$$

[a] R. Lakshmanan and A. P. (2024). "Unbalanced Optimal Transport and Maximum Mean Discrepancies: Interconnections and Rapid Evaluation". In: *Journal of Scientific Computing* 100.72. DOI: 10.1007/s10915-024-02586-2.

	Kernel	Hölder α	constant c
Gauss	$e^{-\|x-y\|^2/\ell^2}$	1	$2/\ell^2$
Laplace	$e^{-\|x-y\|/\ell}$	$1/2$	$2/\ell$
Inverse multiquadratic	$1/\sqrt{\|x-y\|^2+c^2}$	1	$2/c^4$

Outline

Problem statement

Problem (General measures)

$$\text{minimize } \operatorname{MMD}(\mu, \mu^n) = \|\mu_k - \mu_k^n\|_k$$
$$\text{in } \boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$$
$$\text{and } \boldsymbol{x} = (x_1, \dots, x_n),$$

where $\mu^n = \sum_{i=1}^n \mu_i \cdot \delta_{x_i}$.

Here,
$\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$
are the weights, and
$\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$
are the locations.

Problem statement

Problem (General measures)

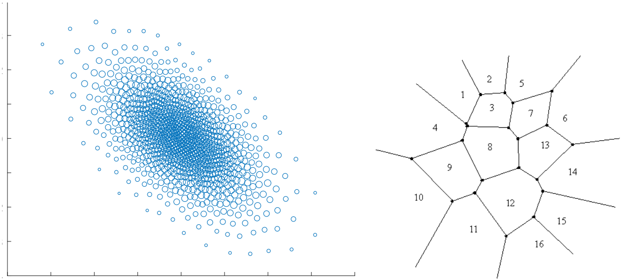
$$\text{minimize } \operatorname{MMD}(\mu, \mu^n) = \|\mu_k - \mu_k^n\|_k$$

$$\text{in } \boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$$

$$\text{and } \boldsymbol{x} = (x_1, \dots, x_n),$$

where $\mu^n = \sum_{i=1}^n \mu_i \cdot \delta_{x_i}$.

Here,
$\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$
are the weights, and
$\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$
are the locations.



(a) Optimal weights and locations **(b)** Voronoi tessellation

Figure: Quantization in Wasserstein distance

Locations given, optimal weights

Proposition

Let $x_1, \dots, x_n$ be fixed. Then the optimal weights $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ satisfy

$$\boldsymbol{K}\boldsymbol{\mu} = \boldsymbol{m},$$

where $\boldsymbol{K} = \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix}$ *and*

$$\boldsymbol{m}_i = \mathbb{E}\, k(x_i, \cdot) = \int_{\mathcal{X}} k(x_i, \xi)\, P(\mathrm{d}\xi).$$

Further, $\mathrm{MMD}_k(\mu, \mu^n)^2 = \iint_{\mathcal{X}^2} k(\xi, \xi')\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') - \boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m}$.

In general,

$$\mu(\mathcal{X}) \neq \mu^n(\mathcal{X}).$$

Locations given, optimal weights

Proposition

Let $x_1, \dots, x_n$ *be fixed. Then the optimal weights* $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ *satisfy*

$$\boldsymbol{K}\boldsymbol{\mu} = \boldsymbol{m},$$

where $\boldsymbol{K} = \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix}$ *and*

$$\boldsymbol{m}_i = \mathbb{E}\, k(x_i,\cdot) = \int_{\mathcal{X}} k(x_i,\xi)\, P(\mathrm{d}\xi).$$

Further, $\mathrm{MMD}_k(\mu,\mu^n)^2 = \iint_{\mathcal{X}^2} k(\xi,\xi')\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') - \boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m}.$

In general,

$$\mu(\mathcal{X}) \neq \mu^n(\mathcal{X}).$$

Optimal weights

Problem (Probability measures)

$$\text{minimize } \|P_k - P_k^n\|_k$$
$$\text{where } P^n = \sum_{i=1}^{n} p_i\, \delta_{x_i} \text{ and } p_1 + \dots + p_n = 1.$$

Proposition

Let $x_1, \dots, x_n$ be fixed. Then the optimal weights $(p_1, \dots, p_n)$ satisfy

$$\boldsymbol{p} = \boldsymbol{K}^{-1}\boldsymbol{m} - \frac{\boldsymbol{K}^{-1}\mathbb{1}\cdot\mathbb{1}^\top\boldsymbol{K}^{-1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}}\boldsymbol{m} + \frac{\boldsymbol{K}^{-1}\mathbb{1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}} \text{ and}$$

$$\begin{aligned}\mathrm{MMD}_k(P, P^n)^2 &= \iint_{\mathcal{X}^2} k(\xi,\xi')\,\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') \\ &\quad - \boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m} + \frac{\left(\mathbb{1}^\top\boldsymbol{K}^{-1}\boldsymbol{m} - 1\right)^2}{\mathbb{1}^\top\boldsymbol{K}^{-1}\mathbb{1}}.\end{aligned}$$

Optimal weights

Problem (Probability measures)

$$\text{minimize } \|P_k - P_k^n\|_k$$
$$\text{where } P^n = \sum_{i=1}^n p_i\,\delta_{x_i} \text{ and } p_1 + \cdots + p_n = 1.$$

Proposition

Let $x_1, \dots, x_n$ be fixed. Then the optimal weights $(p_1, \dots, p_n)$ satisfy

$$\boldsymbol{p} = \boldsymbol{K}^{-1}\boldsymbol{m} - \frac{\boldsymbol{K}^{-1}\mathbb{1}\cdot\mathbb{1}^\top\boldsymbol{K}^{-1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}}\boldsymbol{m} + \frac{\boldsymbol{K}^{-1}\mathbb{1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}} \text{ and}$$

$$\begin{aligned}\mathrm{MMD}_k(P, P^n)^2 &= \iint_{\mathcal{X}^2} k(\xi,\xi')\,\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') \\ &\quad - \boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m} + \frac{\left(\mathbb{1}^\top\boldsymbol{K}^{-1}\boldsymbol{m} - 1\right)^2}{\mathbb{1}^\top\boldsymbol{K}^{-1}\mathbb{1}}.\end{aligned}$$

Optimal weights

Problem (Probability measures)

$$\text{minimize } \|P_k - P_k^n\|_k$$
$$\text{where } P^n = \sum_{i=1}^{n} p_i\, \delta_{x_i} \text{ and } p_1 + \dots + p_n = 1.$$

Proposition

Let $x_1, \dots, x_n$ be fixed. Then the optimal weights $(p_1, \dots, p_n)$ satisfy

$$\boldsymbol{p} = \boldsymbol{K}^{-1}\boldsymbol{m} - \frac{\boldsymbol{K}^{-1}\mathbb{1}\cdot\mathbb{1}^\top\boldsymbol{K}^{-1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}}\boldsymbol{m} + \frac{\boldsymbol{K}^{-1}\mathbb{1}}{\mathbb{1}\boldsymbol{K}^{-1}\mathbb{1}} \text{ and}$$

$$\begin{aligned}\mathrm{MMD}_k(P, P^n)^2 &= \iint_{\mathcal{X}^2} k(\xi, \xi')\,\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') \\ &\quad - \boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m} + \frac{\left(\mathbb{1}^\top\boldsymbol{K}^{-1}\boldsymbol{m} - 1\right)^2}{\mathbb{1}^\top\boldsymbol{K}^{-1}\mathbb{1}}.\end{aligned}$$

Outline

Problem (General measures)

$$\text{minimize } \iint_{\mathcal{X}^2} k(\xi,\xi')\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') - \boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m} + \frac{\left(\mathbb{1}^\top \boldsymbol{K}^{-1}\boldsymbol{m} - 1\right)^2}{\mathbb{1}^\top \boldsymbol{K}^{-1}\mathbb{1}},$$

$$\text{where } \boldsymbol{m}_i = \mathbb{E}\, k(x_i,\cdot) = \int_{\mathcal{X}} k(x_i,\xi)\, P(\mathrm{d}\xi).$$

Define

$$c(\boldsymbol{x},\xi,\xi') := k(\xi,\xi') - k(\xi,\boldsymbol{x})^\top \boldsymbol{K}(\boldsymbol{x})^{-1} k(\xi',\boldsymbol{x}) + \frac{\left(1 - \mathbb{1}^\top \boldsymbol{K}(\boldsymbol{x})^{-1} k(\xi,\boldsymbol{x})\right)^2}{\mathbb{1}^\top \boldsymbol{K}(\boldsymbol{x})^{-1}\mathbb{1}}.$$

Then

$$\min_{\boldsymbol{x}=(x_1,\dots,x_n)} \mathbb{E}_{(\xi,\xi')\sim P\otimes P}\; c(\boldsymbol{x},\xi,\xi').$$

Problem (General measures)

$$\text{minimize} \iint_{\mathcal{X}^2} k(\xi,\xi')\mu(\mathrm{d}\xi)\mu(\mathrm{d}\xi') - \boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m} + \frac{\left(\mathbb{1}^\top \boldsymbol{K}^{-1}\boldsymbol{m} - 1\right)^2}{\mathbb{1}^\top \boldsymbol{K}^{-1}\mathbb{1}},$$

$$\text{where } \boldsymbol{m}_i = \mathbb{E}\, k(x_i,\cdot) = \int_{\mathcal{X}} k(x_i,\xi)\, P(\mathrm{d}\xi).$$

Define

$$c(\boldsymbol{x},\xi,\xi') := k(\xi,\xi') - k(\xi,\boldsymbol{x})^\top \boldsymbol{K}(\boldsymbol{x})^{-1} k(\xi',\boldsymbol{x}) + \frac{\left(1 - \mathbb{1}^\top \boldsymbol{K}(\boldsymbol{x})^{-1} k(\xi,\boldsymbol{x})\right)^2}{\mathbb{1}^\top \boldsymbol{K}(\boldsymbol{x})^{-1}\mathbb{1}}.$$

Then

$$\min_{\boldsymbol{x}=(x_1,\dots,x_n)} \mathbb{E}_{(\xi,\xi')\sim P\otimes P}\; c(\boldsymbol{x},\xi,\xi').$$

To solve

$$\min_{\boldsymbol{x}=(x_1,\dots,x_n)} \mathbb{E}_{(\xi,\xi')\sim P\otimes P}\; c(\boldsymbol{x},\xi,\xi'),$$

repeat

generate independent $\xi,\xi'\sim P\otimes P$,

$$\text{update } \boldsymbol{x} \leftarrow \boldsymbol{x} - \eta_t\cdot\nabla c(\boldsymbol{x},\xi,\xi') \text{ and}$$

$$\boldsymbol{m} \leftarrow \frac{1}{t+1}\left(t\cdot\boldsymbol{m} + \frac{1}{2}\big(k(\xi,\boldsymbol{x}) + k(\xi',\boldsymbol{x})\big)\right),$$

with $\sum_{t=1}^\infty \eta_t = \infty$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$.

Non-negativity constraints

Problem (Probability measures)

$$\begin{aligned} &\text{minimize } \|P_k - P_k^n\|_k \\ &\text{where } P^n = \sum_{i=1}^n p_i\,\delta_{x_i},\; p_1 + \dots + p_n = 1, \\ &\text{and } p_1 \geq 0, \dots, p_n \geq 0. \end{aligned}$$

Instead of solving

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

minimize $f\big(\pi_{\mathcal{X}}(\boldsymbol{x})\big) + \|\boldsymbol{x} - \pi_{\mathcal{X}}(\boldsymbol{x})\|$,

with projection $(\pi_{\mathcal{X}} \circ \pi_{\mathcal{X}} = \pi_{\mathcal{X}})$

$$\pi_{\mathcal{X}}(\boldsymbol{x}) \begin{cases} = \boldsymbol{x} & \text{if } \boldsymbol{x} \in \mathcal{X}, \\ \cdots & \text{if } \boldsymbol{x} \notin \mathcal{X} \end{cases}$$

for example (Norkin and P. (2024))

$\pi_{\mathcal{X}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}\|$.

Example

Suppose that $P \sim \mathcal{N}(\mu, \sigma^2)$, and $k(x,y) = \frac{1}{\sqrt{2\pi\ell^2}} e^{-\frac{1}{2}|x-y|^2}$, then

$$\mathrm{MMD}_k(P, P^n)^2 = \frac{1}{\sqrt{2\pi(2\sigma^2+\ell^2)}}.$$

$$- \frac{1}{\sqrt{2\pi(2\sigma^2+\ell^2)}} \sum_{i,j=1}^{n} e^{-\frac{(x_i-\mu)^2+(x_j-\mu)^2}{2(\sigma^2+\ell^2)}} \boldsymbol{K}(\boldsymbol{x})_{ij}^{-1}.$$

Outline

Proof of concept
Normal distribution

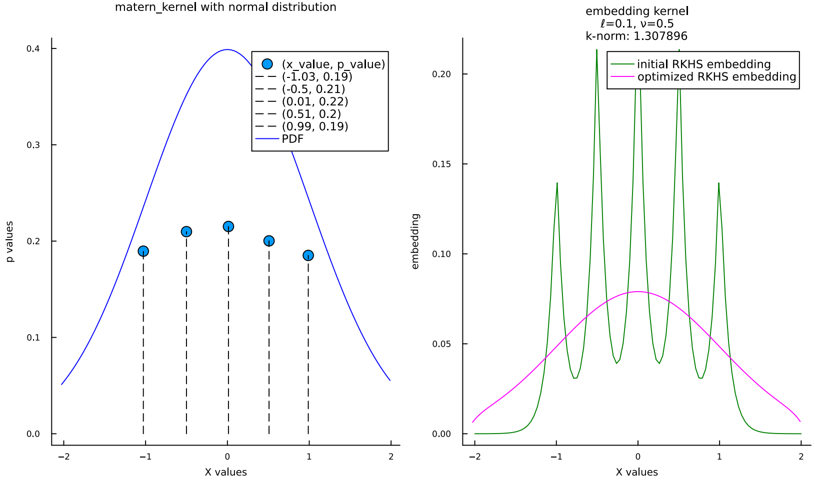


Figure: Normal distribution with Laplace kernel

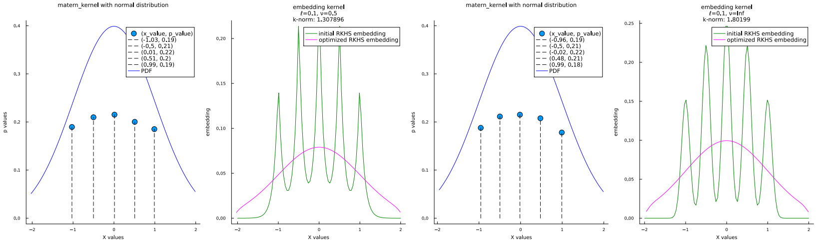


Figure: Normal distribution with smooth kernels

Numerical approximations
Uniform and exponential distribution

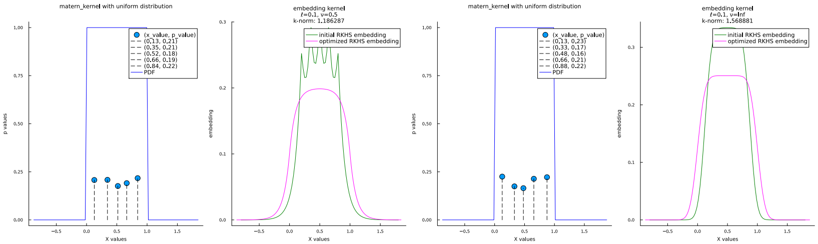


Figure: Uniform with varying kernels

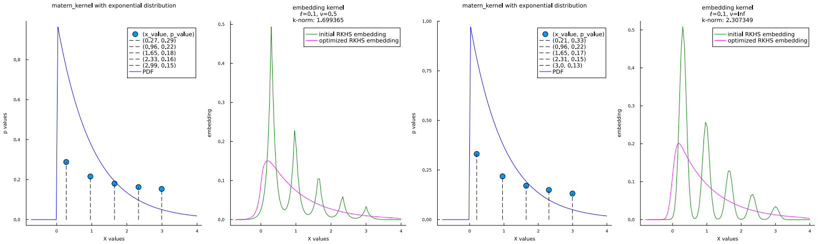
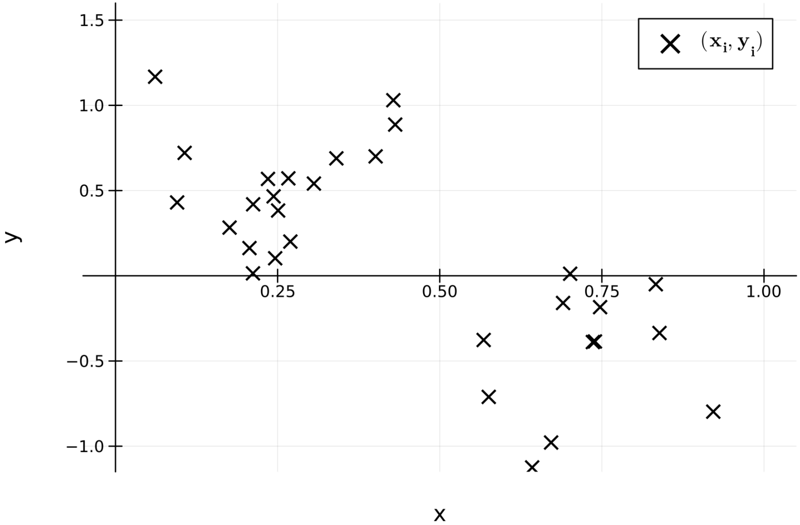


Figure: Exponential with varying kernels

Outline

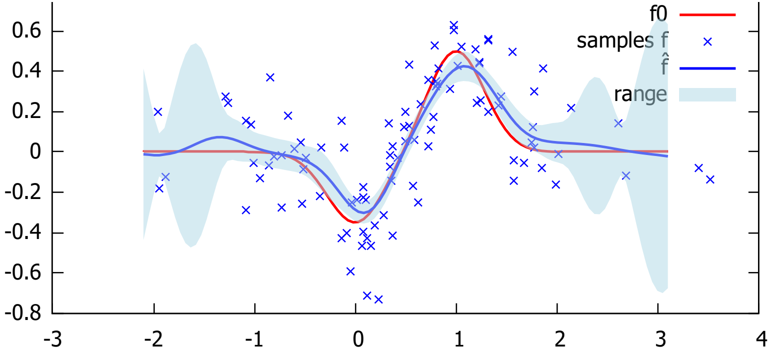
TECHNISCHE UNIVERSITÄT CHEMNITZ

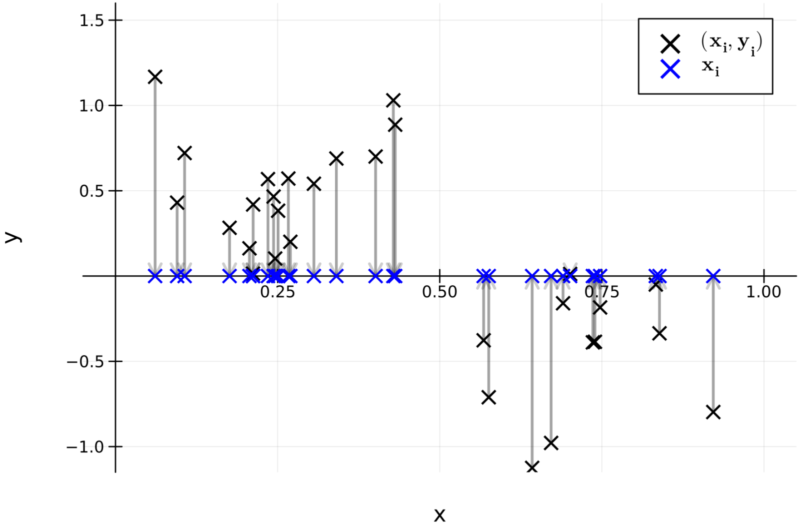
Quality of the predictor



random iid samples from $(x_1, y_1), \dots, (x_n, y_n) \sim (X, Y) \in \mathcal{X} \times \mathbb{R}$ with $\mathcal{X} \in \mathbb{R}^d$.

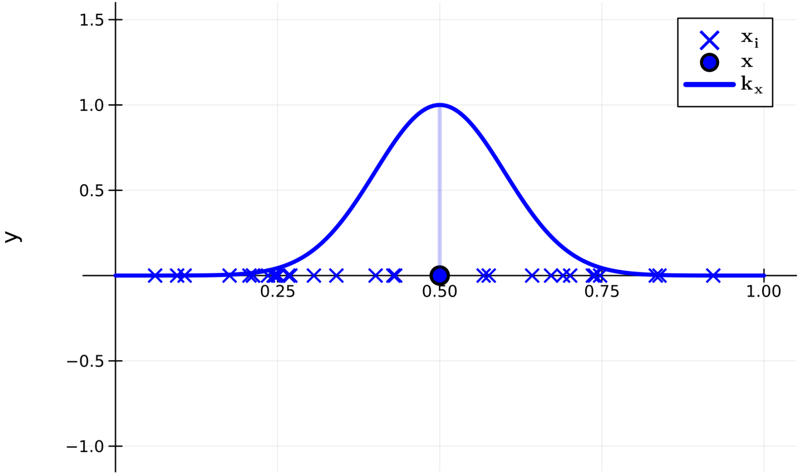
Quality of RKHS approximations
Probability measure on $\mathcal{X}$





The first component, $x_1, \ldots, x_n \sim X$ represent a measure on $\mathcal{X}$.

Approximating elementary functions
by combining translates



The first component, $x_1, \dots, x_n \sim X$ represent a measure on $\mathcal{X}$.

Approximate an elementary function
by combining translates

Best approximation

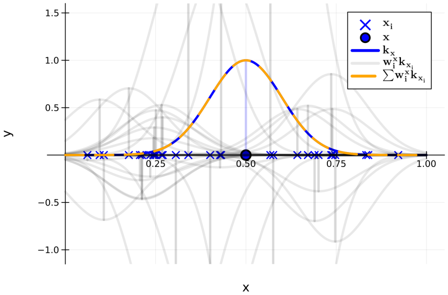
The solution of

$$\min_{w_1,\dots,w_n} \left\| k(\cdot,x) - \sum_{i=1}^{n} w_i\, k(\cdot,x_i) \right\|_k$$

is

$$\boldsymbol{K} w = f,$$

where $\boldsymbol{K}_{ij} = k(x_i,x_j)$ and $f_i = k(x,x_i)$.



Highly oscillating weights :-(

Smoothing splines

Theorem (Representer theorem Schölkopf, Herbrich, and Smola 2001)

The solution of the problem

$$\hat{\vartheta}_n := \min_{f_\lambda(\cdot)\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n \ell(f_i, f_\lambda(X_i)) + \lambda \|f_\lambda(\cdot)\|_k^2$$

takes the form

$$\hat{f}_n(\cdot) := \frac{1}{n}\sum_{j=1}^n \hat{w}_j \cdot k(\cdot, X_j).$$

For $\ell(x,y) = (x-y)^2$, *the weights are* $\hat{w} = \left(\lambda + \frac{1}{n}K\right)^{-1} f$.

Proposition ($\hat{\vartheta}_n$ is downwards biased)

It holds that (irrespective of $\ell(\cdot)$*)*

$$\mathbb{E}\,\hat{\vartheta}_n \le \mathbb{E}\,\hat{\vartheta}_{n+1} \le \vartheta^*.$$

Smoothing splines

Theorem (Representer theorem Schölkopf, Herbrich, and Smola 2001)

The solution of the problem

$$\hat{\vartheta}_n := \min_{f_\lambda(\cdot)\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n \ell(f_i, f_\lambda(X_i)) + \lambda \|f_\lambda(\cdot)\|_k^2$$

takes the form

$$\hat{f}_n(\cdot) := \frac{1}{n}\sum_{j=1}^n \hat{w}_j \cdot k(\cdot, X_j).$$

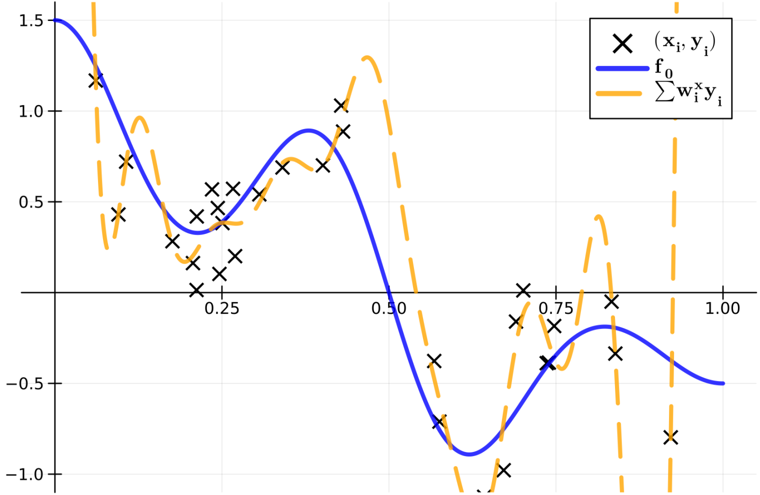
For $\ell(x,y) = (x-y)^2$*, the weights are* $\hat{w} = \left(\lambda + \frac{1}{n}K\right)^{-1} f$.

Proposition ($\hat{\vartheta}_n$ is downwards biased)

It holds that (irrespective of $\ell(\cdot)$*)*

$$\mathbb{E}\,\hat{\vartheta}_n \le \mathbb{E}\,\hat{\vartheta}_{n+1} \le \vartheta^*.$$

Noisy observations



Overfitting

If the data an had error. . .

Covariance operator

Discrete problem (random)

Analyzing the problem

$$\min_{w\in\mathbb{R}^n} \lambda\|w\|_2^2 + \left\| k(\cdot,x) - \sum_{i=1}^n k(\cdot,x_i)w_i \right\|_k^2$$

is difficult, due to the dependence of the support points $x_1,\dots,x_n$.

Discrete problem (random)

Continuous counterpart is

$$\min_{w\in\mathbb{R}^n} \lambda\|w\|_{L^2}^2 + \left\| k(\cdot,x) - L_k w \right\|_k^2$$

with

$$L_k\colon L^2 \to \mathcal{H}_k$$
$$f \mapsto \int_{\mathcal{X}} k(x,z)\, f(z)\, P(\mathrm{d}z)$$

Covariance operator

Discrete problem (random)

Analyzing the problem

$$\min_{w\in\mathbb{R}^n} \lambda\|w\|_2^2 + \left\| k(\cdot,x) - \sum_{i=1}^n k(\cdot,x_i)w_i \right\|_k^2$$

is difficult, due to the dependence of the support points $x_1,\dots,x_n$.

Discrete problem (random)

Continuous counterpart is

$$\min_{w\in\mathbb{R}^n} \lambda\|w\|_{L^2}^2 + \left\| k(\cdot,x) - L_k w \right\|_k^2$$

with

$$L_k\colon L^2 \to \mathcal{H}_k$$

$$f \mapsto \int_{\mathcal{X}} k(x,z)\, f(z)\, P(\mathrm{d}z)$$

Covariance operator

Definition (Covariance operator)

The covariance operator is

$$L_k\colon \mathcal{H}_k \to \mathcal{H}_k$$
$$L_k = \int_{\mathcal{X}} k(\cdot,x)\otimes k(\cdot,x)\,P(\mathrm{d}x),$$

where $(k_x\otimes k_x)(f) := \langle f \mid k_x\rangle\, k_x$.

Remark (Trace class)

It holds that

$$trace(k_x\otimes k_x) = k(x,x)$$

and $trace(L_k) = \int_{\mathcal{X}} k(x,x)\,P(\mathrm{d}x) \le \sup_{x\in\mathcal{X}} k(x,x)$.

Covariance operator

Definition (Covariance operator)

The covariance operator is

$$L_k\colon \mathcal{H}_k \to \mathcal{H}_k$$
$$L_k = \int_{\mathcal{X}} k(\cdot,x)\otimes k(\cdot,x)\, P(\mathrm{d}x),$$

where $(k_x \otimes k_x)(f) := \langle f \mid k_x\rangle\, k_x$.

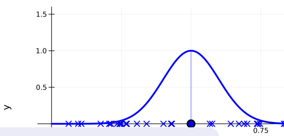
Remark (Trace class)

It holds that

$$trace(k_x\otimes k_x) = k(x,x)$$

and $trace(L_k) = \int_{\mathcal{X}} k(x,x)\, P(\mathrm{d}x) \le \sup_{x\in\mathcal{X}} k(x,x)$.

Maximal marginal degrees of freedom
Approximation quality



Definition

The *maximal marginal degrees of freedom* is the worst case value[a]

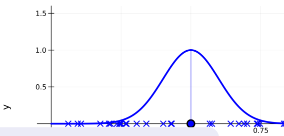
$$\mathcal{N}_\infty(\lambda) := \lambda^{-1} \sup_{x\in\mathcal{X}} \min_{w\in L^2} \lambda\|w\|_{L^2}^2 + \|k(\cdot,x) - L_k w\|_k^2 .$$

[a] A. Rudi, R. Camoriano, and L. Rosasco (2015). "Less is more: Nyström computational regularization". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1657–1665.

The quantity is related to the effective dimension,

$d_{\mathrm{eff}}(\lambda) := \mathbb{E}\left\|(L_k+\lambda)^{-1/2}\right\|^2.$

Maximal marginal degrees of freedom
Approximation quality



Definition

The *maximal marginal degrees of freedom* is the worst case value[a]

$$\mathcal{N}_\infty(\lambda) := \lambda^{-1} \sup_{x\in\mathcal{X}} \min_{w\in L^2} \lambda \|w\|_{L^2}^2 + \|k(\cdot,x) - L_k w\|_k^2 .$$

[a] A. Rudi, R. Camoriano, and L. Rosasco (2015). "Less is more: Nyström computational regularization". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1657–1665.

The quantity is related to the effective dimension,
$d_{\text{eff}}(\lambda) := \mathbb{E} \left\| (L_k+\lambda)^{-1/2} \right\|^2$.

Concentration inequality

Linking the discrete and the continuous problem

$$\mathcal{N}_\infty(\lambda) = \lambda^{-1} \sup_{x\in\mathcal{X}} \min_{w\in L^2} \lambda\|w\|_{L^2}^2 + \|k(\cdot,x) - L_k w\|_k^2.$$

Concentration inequality between discrete and continuous problem ("On the Approximation of Kernel functions")

Let $\tau > 0$. If $\lambda > 0$ is chosen such that

$$\frac{4\tau\mathcal{N}_\infty\left(\frac{\lambda}{n}\right)\ln\mathcal{N}_\infty\left(\frac{\lambda}{n}\right)}{3n} + \sqrt{\frac{2\tau\mathcal{N}_\infty\left(\frac{\lambda}{n}\right)\ln\mathcal{N}_\infty\left(\frac{\lambda}{n}\right)}{n}} \leq \frac{1}{2}$$

is satisfied, it holds, with probability at least $1-2e^{-\tau}$, that

$$\min_{w\in\mathbb{R}^n} \lambda\|w\|_2^2 + \left\|k(\cdot,x) - \sum_{i=1}^n w_i\, k(\cdot,x_i)\right\|_k^2 \leq \frac{4\lambda}{n}\mathcal{N}\left(\frac{\lambda}{n}\right).$$

Concentration inequality

Linking the discrete and the continuous problem

$$\mathcal{N}_\infty(\lambda) = \lambda^{-1} \sup_{x\in\mathcal{X}} \min_{w\in L^2} \lambda \|w\|_{L^2}^2 + \|k(\cdot,x) - L_k w\|_k^2.$$

Concentration inequality between discrete and continuous problem ("On the Approximation of Kernel functions")

Let $\tau > 0$. If $\lambda > 0$ is chosen such that

$$\frac{4\tau \mathcal{N}_\infty\left(\frac{\lambda}{n}\right) \ln \mathcal{N}_\infty\left(\frac{\lambda}{n}\right)}{3n} + \sqrt{\frac{2\tau \mathcal{N}_\infty\left(\frac{\lambda}{n}\right) \ln \mathcal{N}_\infty\left(\frac{\lambda}{n}\right)}{n}} \le \frac{1}{2}$$

is satisfied, it holds, with probability at least $1 - 2e^{-\tau}$, that

$$\min_{w\in\mathbb{R}^n} \lambda \|w\|_2^2 + \left\| k(\cdot,x) - \sum_{i=1}^n w_i\, k(\cdot,x_i) \right\|_k^2 \le \frac{4\lambda}{n} \mathcal{N}\left(\frac{\lambda}{n}\right).$$

Outline

Research questions

Question

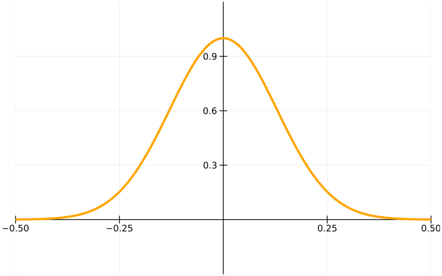
- What are powerful bounds on $\mathcal{N}_\infty(\lambda)$?
- $\hat{f}_{\lambda_n} \to f_0$ in L^2, if $n \to \infty$?
- $\hat{f}_{\lambda_n} \to f_0$ in L^∞, if $n \to \infty$?

Setting

- The sample space is $\mathcal{X} = [0,1]^d$?
- The kernel is

$$k(x,y) = e^{-a\|x-y\|_2^2}$$

with fixed precision parameter $a > 0$.



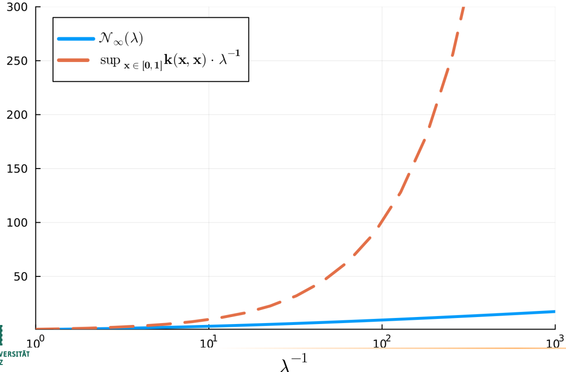
Behavior of $\mathcal{N}_\infty(\lambda)$

Proposition (A priori bound for Maximal marginal degrees of freedom)

The obvious bound is (cf. Rudi, Camoriano, and Rosasco (2015))

$$\mathcal{N}_\infty(\lambda) \le \lambda^{-1} \sup_{x\in\mathcal{X}} k(x,x),$$

which does not reflect the actual behavior.

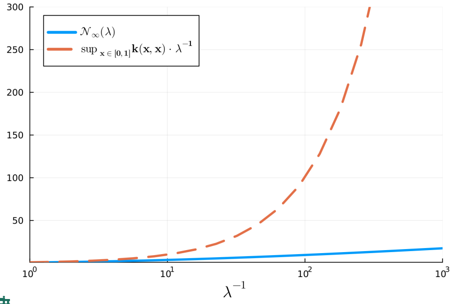


Proposition (A priori bound for Maximal marginal degrees of freedom)

The obvious bound is (cf. Rudi, Camoriano, and Rosasco (2015))

$$\mathcal{N}_\infty(\lambda) \le \lambda^{-1} \sup_{x\in\mathcal{X}} k(x,x),$$

which does not reflect the actual behavior.



Problem

This bound implies that

$$\lambda_n \ge c_0 > 0 \quad \text{for all } n \in \mathbb{N},$$

which gives convergence rates of $\left\|\hat f_{\lambda_n} - f_0\right\|_{L^2} = \mathcal{O}\left((\ln n)^{-\alpha}\right)$ for some $\alpha > 0$.

For the Gaussian kernel only

Proposition

The minimizer of

$$\min_{w\in L^2} \lambda \|w\|_{L^2}^2 + \|k_x - L_k w\|_k^2$$

is

$$w_\lambda^x = (\lambda + L_k)^{-1} k_x.$$

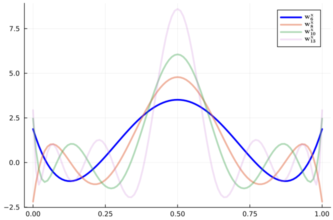


Figure: $x = 0.5$

Lemma

A good proxy is

$$w_\lambda^x(z) \approx \sum_{i=0}^{m-1} Q_i(x) Q_i(z)$$

with Q_i (scaled) Legendre polynomials.

Reason:

$$k(\cdot, a) = \sum_{i=0}^{\infty} \frac{(-a)^i}{i!} \|\cdot - x\|_2^2.$$

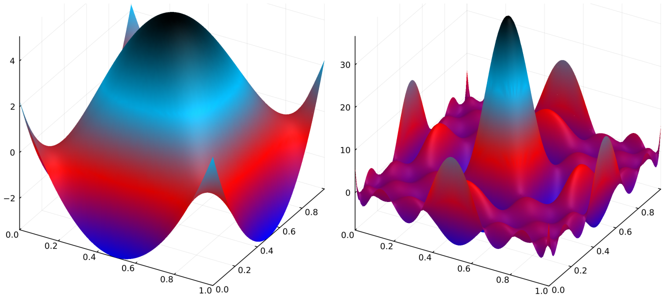
For the Gaussian kernel only

Proposition

For the Gaussian kernel, there is a constant $C>0$ such that[a]

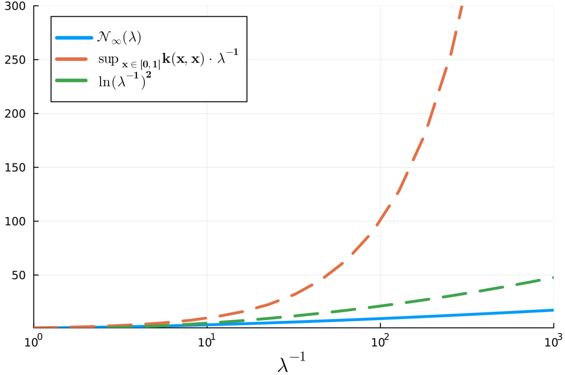
$$\mathcal{N}_\infty(\lambda) \le C \cdot \ln\left(\lambda^{-1}\right)^{2d}.$$

[a] Dommel and P., 2025.



(a) $m=4$, $x=(0.5,0.5)$ **(b)** $m=10$, $x=(0.5,0.5)$

Discussion



Notes

1. The new bound on $\mathcal{N}_\infty(\lambda)$ is significantly *tighter* than the initial bound;
2. The approach works for kernels with polynomial expansions;
3. The bound might still be improved.

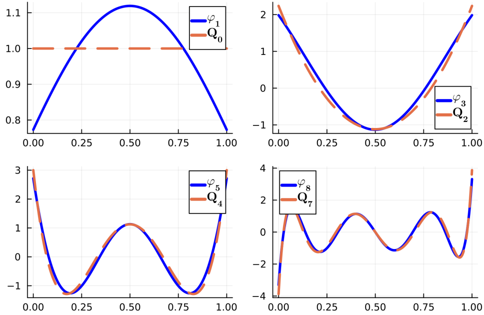
Outline

Consequences on Eigenfunctions

Proposition (Mercer)

The covariance operator L_k is compact and thus

$$(L_k f)(y) = \int_{\mathcal{X}} k(y,z)\,\mu(\mathrm{d}z) = \sum_{\ell=1}^{\infty} \mu_\ell \, \langle \varphi_\ell \mid f \rangle_{L^2} \cdot \varphi_\ell(y).$$



Why?

A bound on

$$\|\varphi_\ell\|_{L^\infty}$$

is essential for deriving uniform convergence for $f_0 \notin \mathcal{H}_k$.

Eigenfunctions
For Gaussian kernels

Relation

The eigenfunctions φ_ℓ and the eigenvalues μ_ℓ are related by

$$\sup_{x\in\mathcal{X}} |\varphi_\ell(x)| \le 4\sqrt{\mathcal{N}_\infty(\mu_\ell)} \le 4C\cdot\left(\ln \mu_\ell^{-1}\right)^d.$$

Proposition

There is a constant $C > 0$ such that

$$\|\varphi_\ell\|_{L^\infty} = \sup_{x\in[0,1]^d} |\varphi_\ell(x)| \le C\,\ell^2$$

for all $\ell \in \mathbb{N}$.

[2] P. Dommel and A. P. (2025). "On the Approximation of Kernel functions". In: *Journal of Machine Learning Research* 26.41, pp. 1–30. URL: http://jmlr.org/papers/v26/24-0270.html.

Eigenfunctions
For Gaussian kernels

Relation

The eigenfunctions φ_ℓ and the eigenvalues μ_ℓ are related by

$$\sup_{x\in\mathcal{X}} |\varphi_\ell(x)| \le 4\sqrt{\mathcal{N}_\infty(\mu_\ell)} \le 4C\cdot\left(\ln \mu_\ell^{-1}\right)^d.$$

Details[2] (2025)

Proposition

There is a constant $C>0$ such that

$$\|\varphi_\ell\|_{L^\infty} = \sup_{x\in[0,1]^d} |\varphi_\ell(x)| \le C\,\ell^2$$

for all $\ell\in\mathbb{N}$.

[2] P. Dommel and A. P. (2025). "On the Approximation of Kernel functions". In: *Journal of Machine Learning Research* 26.41, pp. 1–30. URL: http://jmlr.org/papers/v26/24-0270.html.

Relation of norms

$$\|f\|_{L^2} \le \|f\|_{L^\infty} \le C_k \cdot \|f\|_{\mathcal{H}_k}.$$

More precisely,

$$\|f\|_{L^2} \le \|L_k\|^{1/2} \cdot \|f\|_k \quad \text{and} \quad |f(x)| \le \sqrt{k(x,x)} \cdot \|f\|_{\mathcal{H}_k}.$$

Relation of norms

$$\|f\|_{L^2} \leq \|f\|_{L^\infty} \leq C_k \cdot \|f\|_{\mathcal{H}_k}.$$

More precisely,

$$\|f\|_{L^2} \leq \|L_k\|^{1/2} \cdot \|f\|_k \quad \text{and} \quad |f(x)| \leq \sqrt{k(x,x)} \cdot \|f\|_{\mathcal{H}_k}.$$

$$\|f\|_{L^2} \leq \|f\|_{L^\infty} \leq C_k \cdot \|f\|_{\mathcal{H}_k}$$

Theorem

For every function $f(\cdot) = \sum_{\ell=1}^{\infty} c_\ell \varphi_\ell(\cdot)$ *such that* $\sum_{\ell=1}^{\infty} \frac{c_\ell^2}{\ell^{-2\alpha}} < \infty$ *for some* $\alpha > 3$*, it holds that*

$$\|f\|_{L^\infty} \leq C \frac{\pi}{\sqrt{6}} \left(\sum_{\ell=1}^{\infty} \frac{c_\ell^2}{\ell^{-2\alpha}} \right)^{3/\alpha} \cdot \|f\|_{L^2}^{1-\frac{3}{\alpha}}.$$

Remark

1. The condition $\sum_{\ell=1}^{\infty} \frac{c_\ell^2}{\ell^{-2\alpha}} < \infty$ can be verified for functions $f \in \mathcal{H}_k$.
2. If $f \in H^s$ for $s > \frac{7d}{2}$, then $\sum_{\ell=1}^{\infty} \frac{c_\ell^2}{\ell^{-2\alpha}} < \infty$.

Theorem (Approximation of smooth functions)

It holds – with high probability – that

$$\left\|\hat{f}_{\lambda_n}-f_0\right\|_{L^2}^2=\mathcal{O}\left(n^{-\frac{2s}{2s+d}+\delta}\right)$$

and

$$\left\|\hat{f}_{\lambda_n}-f_0\right\|_{L^\infty}^2=\mathcal{O}\left(n^{-\frac{s-\frac{7d}{2}}{2s+d}+\delta}\right)$$

$(\delta>0)$ *where* $f_0\in H^s$*.*

Remark

Note the convergence for $s=\infty$, that is, $f\in C^\infty$.

Conclusions & Consequences

Theorem (Approximation of smooth functions)

It holds – with high probability – that

$$\left\|\hat{f}_{\lambda_n} - f_0\right\|_{L^2}^2 = \mathcal{O}\left(n^{-\frac{2s}{2s+d}+\delta}\right)$$

and

$$\left\|\hat{f}_{\lambda_n} - f_0\right\|_{L^\infty}^2 = \mathcal{O}\left(n^{-\frac{s-\frac{7d}{2}}{2s+d}+\delta}\right)$$

$(\delta > 0)$ *where* $f_0 \in H^s$.

Remark

Note the convergence for $s = \infty$, that is, $f \in C^\infty$.

Conclusions & Consequences, cont.

Approximations in lower dimensions: "Less is more: Nyström computational regularization"

The Nyström method

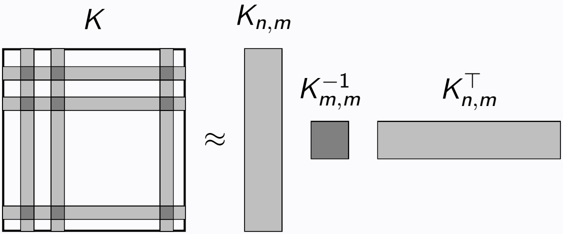


Figure: Approximation of the kernel matrix by lower rank matrices

Outline

RKHS. Useful in Stochastic Optimization?

Pros and potential future research

- Less functions (smooth)
- Faster empirical convergence
- Simpler to compute than Wasserstein
- Less curse of dimensionality
- Performance on real-world data
- Distributionally robust optimization: with MMD instead of Wasserstein
- What is the right kernel?
- Bandwidth selection à la Norkin
- Extend the results to multistage (nested) MMD

Thank you.
Literature

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US. DOI: 10.1007/978-1-4419-9096-9.

Diestel, J. and J. J. Uhl Jr. (1977). *Vector Measures*. Mathematical Surveys 15. American Mathematical Society. URL: http://books.google.com/books?id=EQFjD90fXWAC.

Dommel, P. and A. P. (2025). "On the Approximation of Kernel functions". In: *Journal of Machine Learning Research* 26.41, pp. 1–30. URL: http://jmlr.org/papers/v26/24-0270.html.

Lakshmanan, R. and A. P. (2024). "Unbalanced Optimal Transport and Maximum Mean Discrepancies: Interconnections and Rapid Evaluation". In: *Journal of Scientific Computing* 100.72. DOI: 10.1007/s10915-024-02586-2.

Norkin, V. and A. P. (2024). "Portfolio reshaping under 1st-order stochastic dominance constraints by the exact penalty function methods". In: *Optimization*, pp. 1–34. DOI: 10.1080/02331934.2024.2313086.

Rudi, A., R. Camoriano, and L. Rosasco (2015). "Less is more: Nyström computational regularization". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1657–1665.

Schölkopf, B., R. Herbrich, and A. J. Smola (2001). "A Generalized Representer Theorem". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 416–426. DOI: 10.1007/3-540-44581-1_27.

Outline

TECHNISCHE UNIVERSITÄT CHEMNITZ

Gaussian random fields

Method I: Feature map

Let $\varphi_k\colon \mathcal{X} \to \mathbb{R}$ be functions, $\sigma_k \in \mathbb{R}$. Set

$$f(x) := \sum_{k=0}^{\infty} \quad \sigma_k \varphi_k(x), \qquad x \in \mathcal{X}, \omega \in \Omega,$$

Gaussian random fields

Method I: Feature map

Let $\varphi_k\colon \mathcal{X} \to \mathbb{R}$ be functions, $\sigma_k \in \mathbb{R}$. Set

$$f(x) := \sum_{k=0}^{\infty} \xi_k \sigma_k \varphi_k(x), \qquad x \in \mathcal{X}, \omega \in \Omega,$$

with $\xi_k \sim \mathcal{N}(0,1)$ iid. Note, that $\mathbb{E}\,\xi_k = 0$ and $\mathbb{E}\,\xi_k \xi_\ell = \delta_{k\ell}$. It follows that $\mathbb{E}\,f(x) = 0$ and

$$\operatorname{cov}\big(f(x), f(y)\big) = \sum_{k=0}^{\infty} \sigma_k^2 \varphi_k(x) \varphi_k(y), \qquad x, y \in \mathcal{X}.$$

Let $\varphi_k\colon \mathcal{X} \to \mathbb{R}$ be functions, $\sigma_k \in \mathbb{R}$. Set

$$f(x) := \sum_{k=0}^{\infty} \xi_k \, \sigma_k \, \varphi_k(x), \qquad x \in \mathcal{X}, \omega \in \Omega,$$

with $\xi_k \sim \mathcal{N}(0,1)$ iid. Note, that $\mathbb{E}\,\xi_k = 0$ and $\mathbb{E}\,\xi_k \xi_\ell = \delta_{k\ell}$. It follows that $\mathbb{E}\,f(x) = 0$ and

$$\operatorname{cov}\big(f(x), f(y)\big) = \sum_{k=0}^{\infty} \sigma_k^2 \, \varphi_k(x) \, \varphi_k(y), \qquad x, y \in \mathcal{X}.$$

Let $\varphi_k\colon \mathcal{X} \to \mathbb{R}$ be functions, $\sigma_k \in \mathbb{R}$. Set

$$f(x) := \sum_{k=0}^{\infty} \xi_k \sigma_k \varphi_k(x), \qquad x \in \mathcal{X}, \omega \in \Omega,$$

with $\xi_k \sim \mathcal{N}(0,1)$ iid. Note, that $\mathbb{E}\,\xi_k = 0$ and $\mathbb{E}\,\xi_k\xi_\ell = \delta_{k\ell}$. It follows that $\mathbb{E}\, f(x) = 0$ and

$$\operatorname{cov}\left(f(x), f(y)\right) = \sum_{k=0}^{\infty} \sigma_k^2 \varphi_k(x)\varphi_k(y), \qquad x, y \in \mathcal{X}.$$

Let $\varphi_k \colon \mathcal{X} \to \mathbb{R}$ be functions, $\sigma_k \in \mathbb{R}$. Set

$$f(x) := \sum_{k=0}^{\infty} \xi_k \sigma_k \varphi_k(x), \qquad x \in \mathcal{X}, \omega \in \Omega,$$

with $\xi_k \sim \mathcal{N}(0,1)$ iid. Note, that $\mathbb{E}\,\xi_k = 0$ and $\mathbb{E}\,\xi_k \xi_\ell = \delta_{k\ell}$. It follows that $\mathbb{E}\,f(x) = 0$ and

$$k(x,y) := \operatorname{cov}\big(f(x), f(y)\big) = \sum_{k=0}^{\infty} \sigma_k^2 \varphi_k(x) \varphi_k(y), \qquad x, y \in \mathcal{X}.$$

Hence,

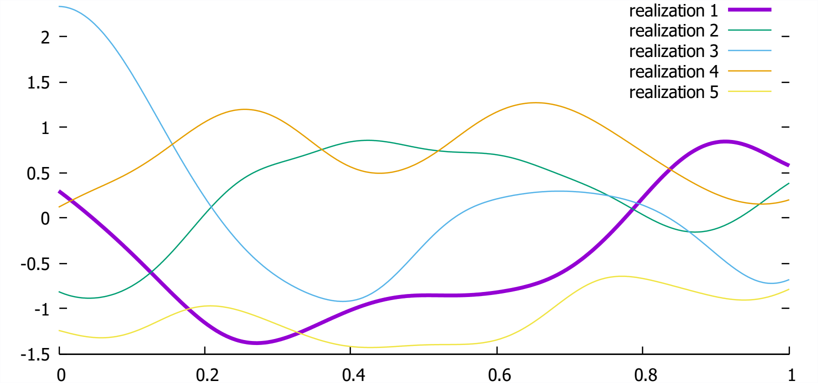
$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix} \right);$$

in particular, $f(x) \sim \mathcal{N}\big(0, \sum_{k=0}^{\infty} \sigma_k^2 \varphi_k(x)^2\big)$.

Example
Gaussian like (polynomial, radial) feature map

Example (RBF)

Feature map: $\varphi_k(x) := \left({}^{x}/_{\ell}\right)^k \cdot e^{-x^2/2\ell^2}$, $\sigma_k^2 := \frac{1}{k!}$

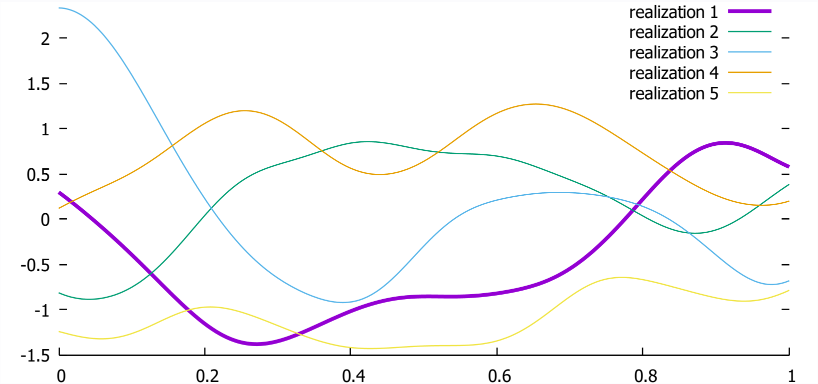


Example

Gaussian like (polynomial, radial) feature map

Example (RBF)

Feature map: $\varphi_k(x) := (x/\ell)^k \cdot e^{-x^2/2\ell^2}$, $\sigma_k^2 := \frac{1}{k!}$



$$k(x,y) = \sum_{k=0} \sigma_k^2 \varphi_k(x) \varphi_k(y) = \exp\left(-\frac{1}{2\ell^2}(x-y)^2\right)$$

TECHNISCHE UNIVERSITÄT CHEMNITZ

Example

Feature map: $\varphi_k(x) := \sqrt{2}\sin\left((k-\frac{1}{2})\pi x\right)$, $\sigma_k := \frac{1}{(k-\frac{1}{2})\pi}$

$$k(x,y) = \sum_{k=1} \sigma_k^2 \varphi_k(x)\varphi_k(y) = \min(x,y)$$

Example

Example

Feature map: $\varphi_k(x) := \sqrt{2}\sin\left((k-\frac{1}{2})\pi x\right)$, $\sigma_k := \frac{1}{(k-\frac{1}{2})\pi}$

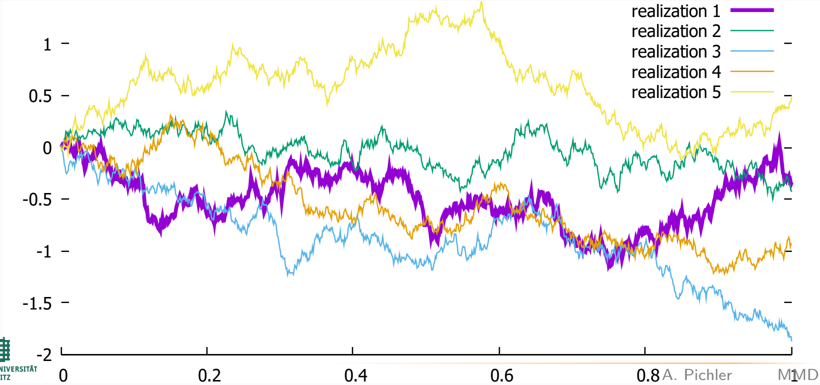
$$k(x,y) = \sum_{k=1} \sigma_k^2 \varphi_k(x)\varphi_k(y) = \min(x,y)$$

TECHNISCHE UNIVERSITÄT CHEMNITZ

Example

Feature map: $\varphi_k(x) := \sqrt{2}\sin\left((k-\frac{1}{2})\pi x\right)$, $\sigma_k := \frac{1}{(k-\frac{1}{2})\pi}$

$$k(x,y) = \sum_{k=1} \sigma_k^2 \varphi_k(x)\varphi_k(y) = \min(x,y)$$

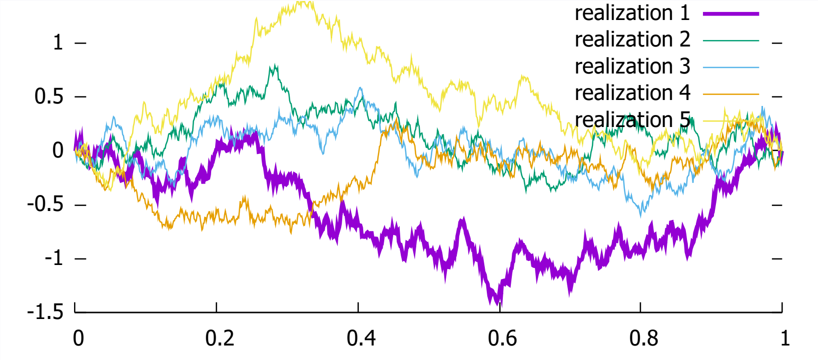


Example
Brownian bridge

Example

Choose $\varphi_k(x) := \sqrt{2}\sin(k\pi x)$, $\sigma_k := \frac{1}{k\pi}$

$$k(x,y) = \min(x,y) - x\,y = \sum_{k=1} \sigma_k^2 \varphi_k(x)\varphi_k(y)$$



Outline

TECHNISCHE UNIVERSITÄT CHEMNITZ

Gaussian random fields

Method II: Gramian

If $\xi_i \sim \mathcal{N}(0,1)$ are iid and

$$K = \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix} = \Phi\Phi^\top$$

(for example $\Phi = K^{1/2}$), then

$$X := \mu + \Phi \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \sim \mathcal{N}(\mu, K).$$

If $\xi_i \sim \mathcal{N}(0,1)$ are iid and

$$K = \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix} = \Phi\Phi^\top$$

(for example $\Phi = K^{1/2}$), then

$$X := \mu + \Phi \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \sim \mathcal{N}(\mu, K).$$

We find the realization

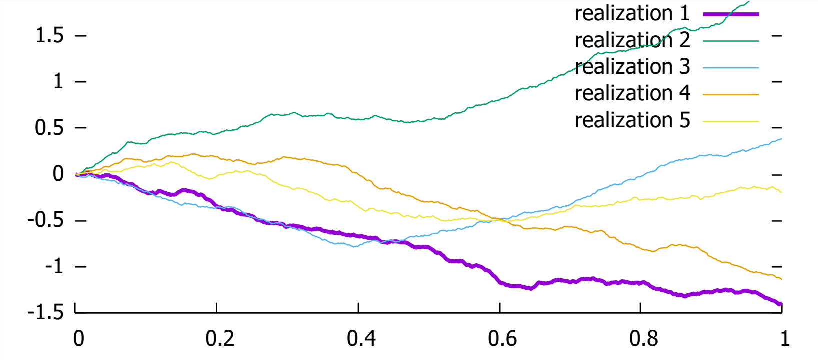
$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} := X \sim \mathcal{N}(0, K).$$

Example
Fractional Brownian motion

Choose $2k(x,y) = x^{2H} + y^{2H} - |x-y|^{2H}$

Example

Hurst index $H = 0.8$:[a] increments are positively correlated

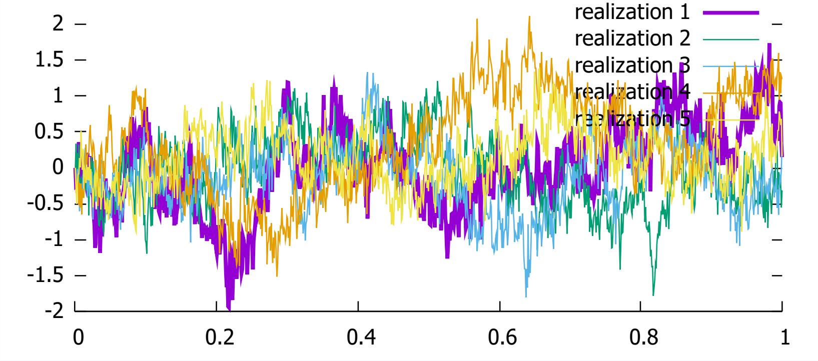


[a] The Wiener process has Hurst index $H = 1/2$.

Choose $2k(x,y) = x^{2H} + y^{2H} - |x-y|^{2H}$

Example

Hurst index $H = 0.2$: increments are negatively correlated



Outline

With Gramian $K := \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix}$, choose the weights[3]

$$w \sim \mathcal{N}(0, K^{-1})$$

and set

$$f(\cdot) := \sum_{i=1}^{n} w_i \cdot k(\cdot, x_i)$$

[3] In data science, the matrix K^{-1} is the *precision matrix*.

Gaussian random fields

Method III: RKHS representation

With Gramian $K := \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix}$, choose the weights[3]

$$w \sim \mathcal{N}(0, K^{-1})$$

and set

$$f(\cdot) := \sum_{i=1}^{n} w_i \cdot k(\cdot, x_i)$$

[3] In data science, the matrix K^{-1} is the *precision matrix*.

With Gramian $K := \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix}$, choose the weights[3]

$$w \sim \mathcal{N}(0, K^{-1})$$

and set

$$f(\cdot) := \sum_{i=1}^{n} w_i \cdot k(\cdot, x_i)$$

Proposition

Then

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}(0, K).$$

[3] In data science, the matrix K^{-1} is the *precision matrix*.

With Gramian $K := \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix}$, choose the weights[3]

$$w \sim \mathcal{N}(0, K^{-1})$$

and set

$$f(\cdot) := \sum_{i=1}^n w_i \cdot k(\cdot, x_i) \in \mathcal{H}_k: \ \text{RKHS}, \text{ with } \langle k(\cdot,x), k(\cdot,y)\rangle_k = k(x,y)$$

Proposition

Then

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}(0, K).$$

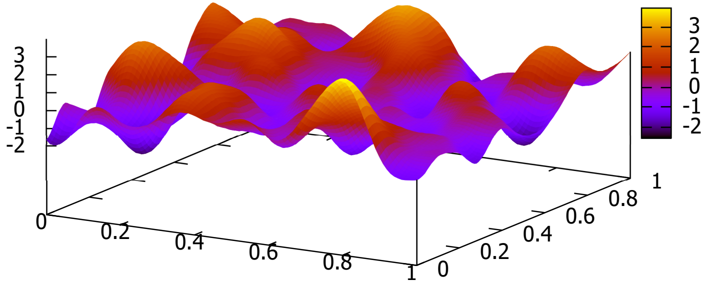
[3] In data science, the matrix K^{-1} is the *precision matrix*.

2D process visualizations

Example

Choose the radial Gaussian kernel[a]

$$k(x,y) = \sigma_f^2 \cdot \exp\left(-\|x-y\|^2/\ell^2\right)$$



[a] This is a Matérn-∞ covariance kernel: all derivatives available everywhere a.s.

2D process visualizations

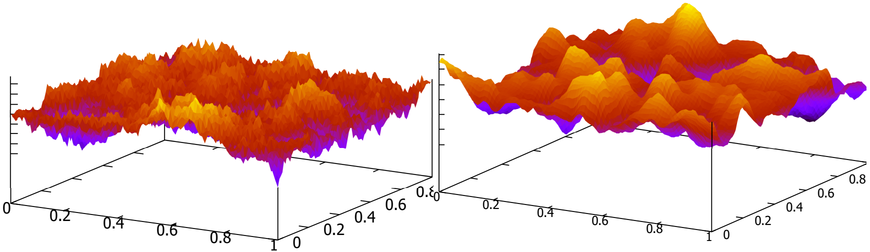


Figure: Laplace (Ornstein–Uhlenbeck) Matérn

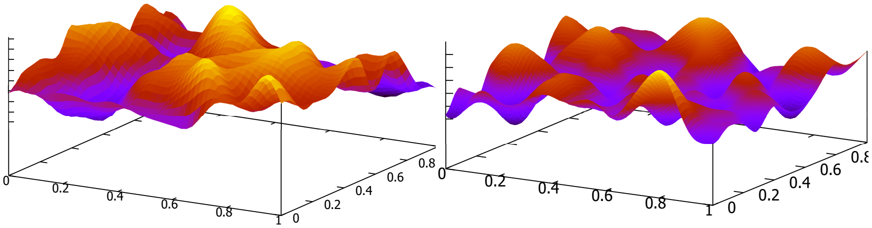


Figure: Sigmoid Gauss

Outline

TECHNISCHE UNIVERSITÄT CHEMNITZ

Suppose that

$$f_i = f_0(\hat{x}_i) + \varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0, \Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X}, \hat{X}) & k(\hat{X}, X) \\ k(X, \hat{X}) & k(X, X) + \Lambda \end{pmatrix} \right).$$

Now RKHS

Signal + noise: predictions

Suppose that

$$f_i = f_0(\hat{x}_i) + \varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0, \Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X}, \hat{X}) & k(\hat{X}, X) \\ k(X, \hat{X}) & k(X, X) + \Lambda \end{pmatrix} \right).$$

Now RKHS

Signal + noise: predictions

Suppose that

$$f_i = f_0(\hat{x}_i) + \varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0, \Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X}, \hat{X}) & k(\hat{X}, X) \\ k(X, \hat{X}) & k(X, X) + \Lambda \end{pmatrix} \right).$$

TECHNISCHE UNIVERSITÄT CHEMNITZ

Suppose that

$$f_i = f_0(\hat{x}_i) + \varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0, \Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X}, \hat{X}) & k(\hat{X}, X) \\ k(X, \hat{X}) & k(X, X) + \Lambda \end{pmatrix} \right).$$

It follows that

$$f_0(\hat{X}) \mid f(X) \sim \mathcal{N}(\hat{\mu}, \hat{K}),$$

where

$$\hat{\mu} := k(\hat{X}, X) \big(k(X, X) + \Lambda\big)^{-1} f(X)$$

and

$$\hat{K} := k(\hat{X}, \hat{X}) - k(\hat{X}, X) \big(k(X, X) + \Lambda\big)^{-1} k(X, \hat{X}).$$

Now RKHS

Signal + noise: predictions

Suppose that

$$f_i = f_0(\hat{x}_i)+\varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0,\Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X},\hat{X}) & k(\hat{X},X) \\ k(X,\hat{X}) & k(X,X)+\Lambda \end{pmatrix}\right).$$

It follows that

$$f_0(\hat{X}) \mid f(X) \sim \mathcal{N}(\hat{\mu}, \hat{K}),$$

where

$$\hat{\mu} := k(\hat{X},X)\big(k(X,X)+\Lambda\big)^{-1} f(X)$$

and

$$\hat{K} := k(\hat{X},\hat{X}) - k(\hat{X},X)\big(k(X,X)+\Lambda\big)^{-1} k(X,\hat{X}).$$

Now RKHS
Signal + noise: predictions

Suppose that

$$f_i = f_0(\hat{x}_i) + \varepsilon.$$

Let $\hat{X} := (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ be sequences of points and $\varepsilon \sim \mathcal{N}(0, \Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X}, \hat{X}) & k(\hat{X}, X) \\ k(X, \hat{X}) & k(X, X) + \Lambda \end{pmatrix} \right).$$

It follows that

$$f_0(\hat{X}) \mid f(X) \sim \mathcal{N}(\hat{\mu}, \hat{K}),$$

where

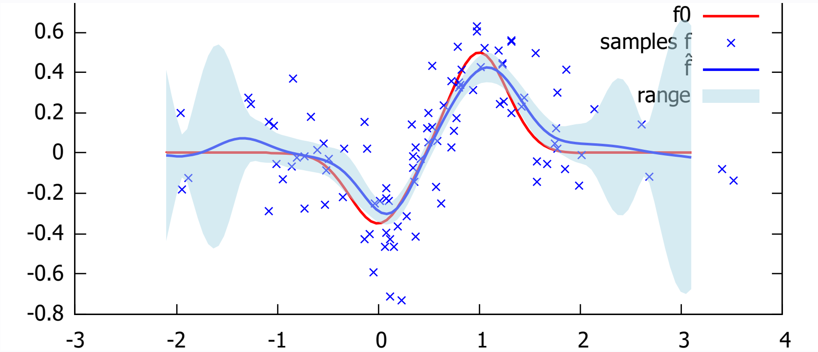
$$\hat{\mu} := k(\hat{X}, X)\big(k(X, X) + \Lambda\big)^{-1} f(X)$$

and

$$\hat{K} := k(\hat{X}, \hat{X}) - k(\hat{X}, X)\big(k(X, X) + \Lambda\big)^{-1} k(X, \hat{X}).$$

Quality of the predictor

Example



The local variance

$$\operatorname{var}\left(f_0(x) \middle| f(X_1) = f_1, \ldots, f(X_n) = f_n\right)$$
$$= k(x,x) - k(x,X)\left(k(X,X) + \Lambda\right)^{-1} k(X,x).$$

does *not* depend on the samples f_i!

In other words, the prediction for a single new point x is

$$\mathbb{E}\big(f_0(\cdot)\big|\, f(x_1)=f_1,\dots,f(x_n)=f_n\big)=\sum_{i=1}^n \hat{w}_i\cdot k(\cdot,x_i),$$

where $\hat{w}$ solves the linear system of equations

$$\sum_{j=1}^n \big(k(x_i,x_j)+\Lambda_{ij}\big)\,\hat{w}_j=f_i,\quad i=1,\dots,n.$$

In other words, the prediction for a single new point x is

$$\mathbb{E}\big(f_0(\cdot)\big|\, f(x_1)=f_1,\dots,f(x_n)=f_n\big)=\sum_{i=1}^n \hat{w}_i\cdot k(\cdot,x_i),$$

where $\hat{w}$ solves the linear system of equations

$$\sum_{j=1}^n \big(k(x_i,x_j)+\Lambda_{ij}\big)\,\hat{w}_j=f_i,\quad i=1,\dots,n.$$

TECHNISCHE UNIVERSITÄT CHEMNITZ

In other words, the prediction for a single new point x is

$$\mathbb{E}\big(f_0(\cdot)\big|\, f(x_1)=f_1,\dots,f(x_n)=f_n\big)=\sum_{i=1}^n \hat{w}_i\cdot k(\cdot,x_i),$$

where $\hat{w}$ solves the linear system of equations

$$\sum_{j=1}^n \big(k(x_i,x_j)+\Lambda_{ij}\big)\,\hat{w}_j=f_i,\quad i=1,\dots,n.$$

The variance is

$$\begin{aligned}&\operatorname{var}\big(f_0(x)\big|\, f(X_1)=f_1,\dots,f(X_n)=f_n\big)\\&\qquad=k(x,x)-k(x,X)\big(k(X,X)+\Lambda\big)^{-1}k(X,x).\end{aligned}$$

If $\Lambda=0$, then $\operatorname{var}\big(f_0(X_i)\big|\, f(X_1)=f_1,\dots,f(X_n)=f_n\big)=0$.