Stochastic Optimization With Random Fields

Convergence in RKHS norms

Alois Pichler

Faculty of mathematics
TU Chemnitz
May 23, 2022

Mathematik!
TU Chemnitz

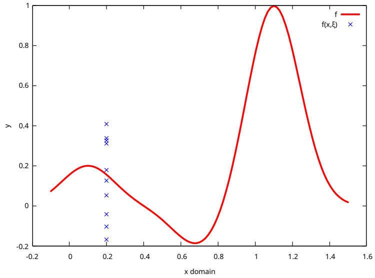
Motivation

Conditional expectation and stochastic optimization

Problem (Stochastic optimization)

Solve

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



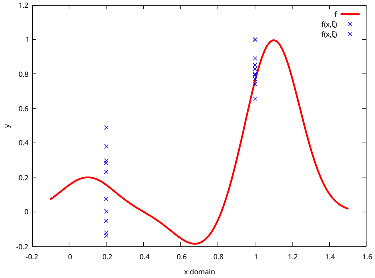
Motivation

Conditional expectation and stochastic optimization

Problem (Stochastic optimization)

Solve

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

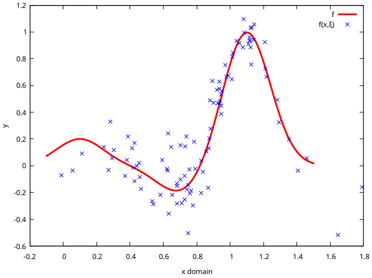


Motivation
Conditional expectation and stochastic optimization

Problem (Stochastic optimization)

Solve

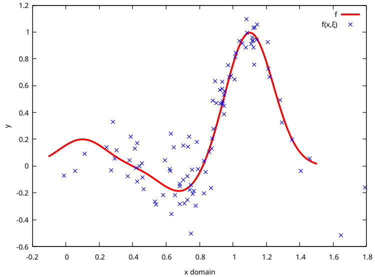
$$\min_{x\in\mathcal{X}} f_0(x) := \mathbb{E}\, f(x,Y).$$



Problem (Stochastic optimization)

Solve

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



Problem (Optimal control: Hamilton–Jacobi–Bellman)

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

Problem (Time series, learning)

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

Outline

1. **Deriving RKHS from stochastics**
   - Gaussian random fields
   - Traditional realization
   - Representation as RKHS function
2. **Predictions from Gaussian processes**
   - Conditional Gaussians
   - Conditional Gaussians, applied to RKHS
3. **Perspective from stochastic optimization**
   - Stochastic optimization problem
   - Denoising
   - Order of convergence

Outline

Gaussian random fields

Method I: Feature map

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

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

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

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

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

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

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

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

Hence,

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

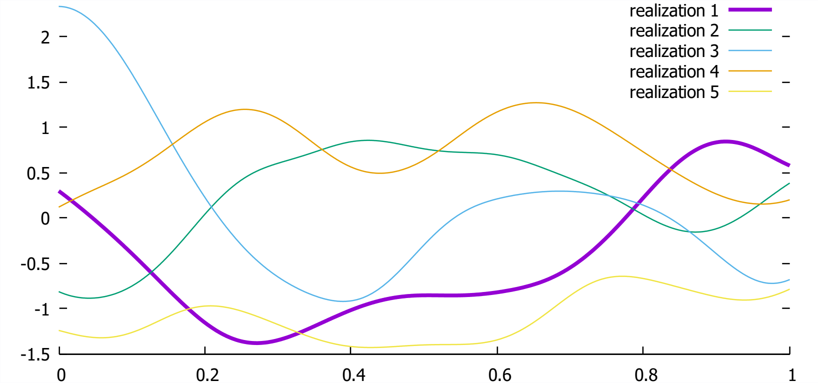
in particular, $f(x)\sim\mathcal{N}\big(0, \sum_{k=0}^{\infty}\sigma_k^2\,\varphi_k(x)^2\big)$.

Example

Gaussian like (polynomial, radial) feature map

Example (RBF)

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

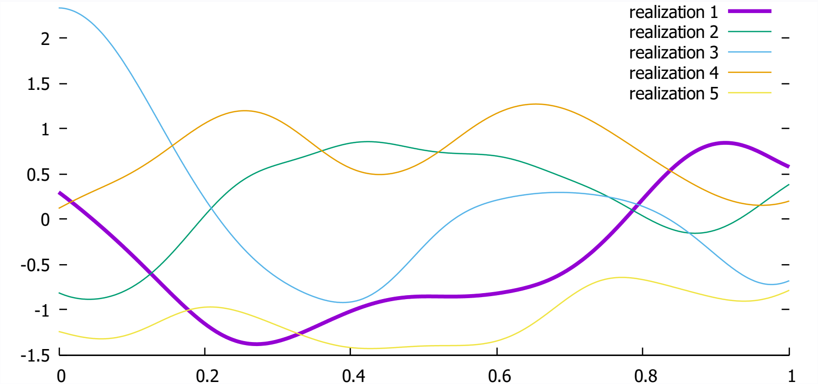


Example

Gaussian like (polynomial, radial) feature map

Example (RBF)

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



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

Example

Example

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

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

Example

Example

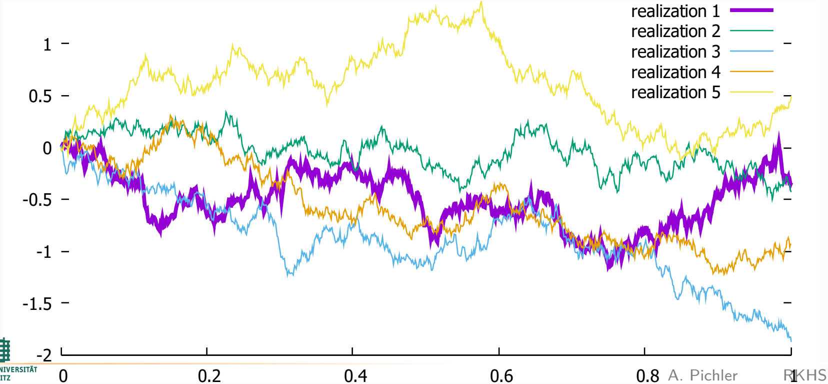
Feature map: $\varphi_k(x) := \sqrt{2}\sin\left((k-\frac{1}{2})\pi x\right)$, $\sigma_k := \frac{1}{(k-\frac{1}{2})\pi}$

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

Example

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

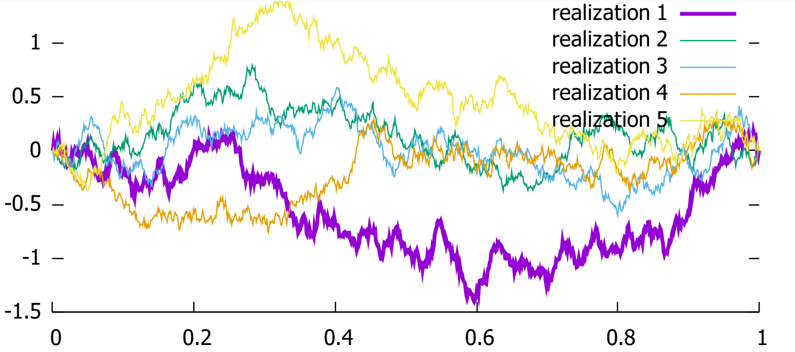
$$k(x,y) = \sum_{k=1} \sigma_k^2 \varphi_k(x)\varphi_k(y) = \min(x,y)$$



Example

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

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

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

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

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

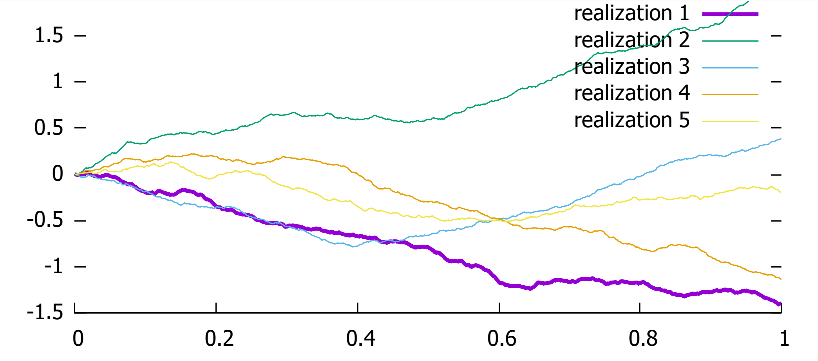
We find the realization

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

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

Example

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

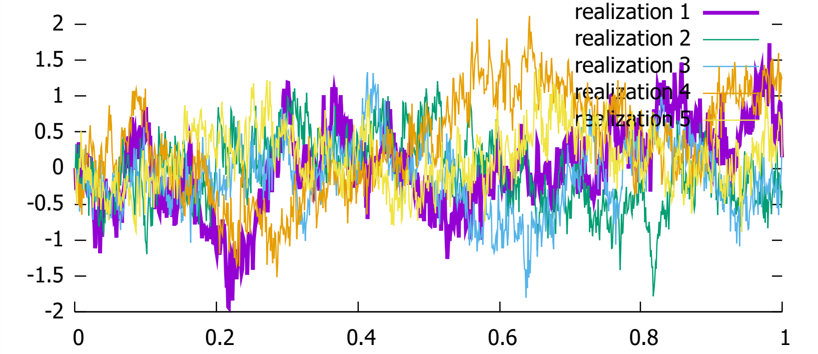


[a] The Wiener process has Hurst index $H = 1/2$.

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

Example

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



Outline

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[1]

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

and set

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

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

TECHNISCHE UNIVERSITÄT CHEMNITZ

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[1]

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

and set

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

[1] 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[1]

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

and set

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

Proposition

Then

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

[1] 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[1]

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

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

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

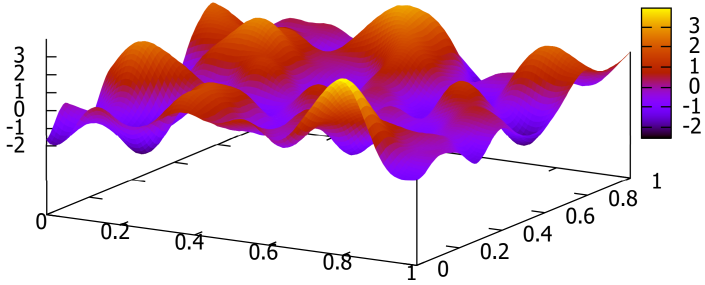
[1] In data science, the matrix K^{-1} is the *precision matrix*.

2D process visualizations

Example

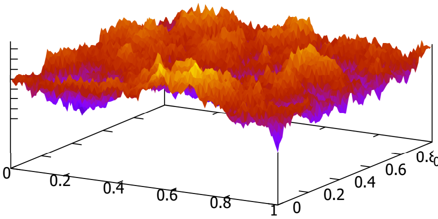
Choose the radial Gaussian kernel[a]

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

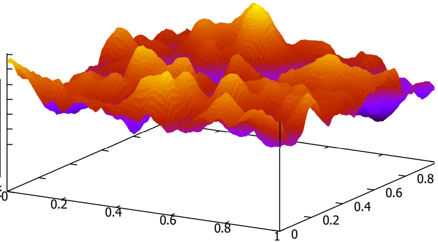


[a] This is a Matérn-∞ covariance kernel: all derivatives available everywhere a.s.

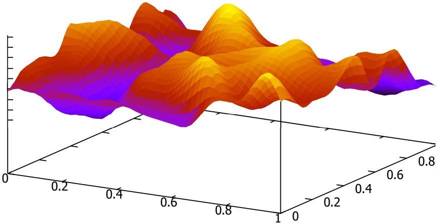
2D process visualizations



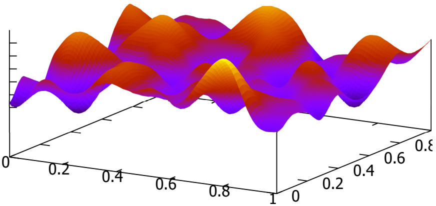
Laplace (Ornstein–Uhlenbeck)



Matérn



Sigmoid



Gauss

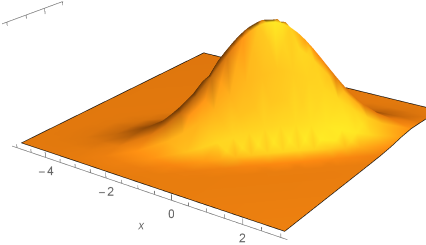
Outline

Conditional Gaussians are Gaussian

Theorem (Cf. [Bishop, 2006])

Suppose that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{pmatrix} \right),$$

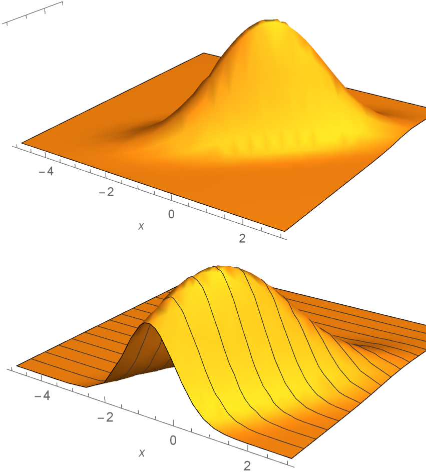


Conditional Gaussians are Gaussian

Theorem (Cf. [Bishop, 2006])

Suppose that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{pmatrix} \right),$$



Conditional Gaussians are Gaussian

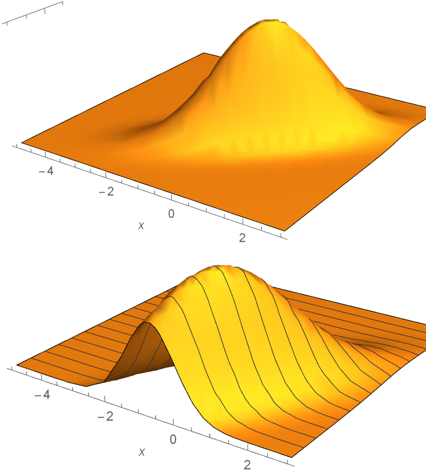
Theorem (Cf. [Bishop, 2006])

Suppose that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{pmatrix} \right),$$

then the conditional distribution is Gaussian as well:

$$X|\,Y \sim \mathcal{N}\left(\begin{matrix} \mu_X + K_{XY} K_{YY}^{-1} (Y - \mu_Y), \\ K_{XX} - K_{XY} K_{YY}^{-1} K_{YX} \end{matrix} \right)$$



Outline

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

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

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

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

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

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

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

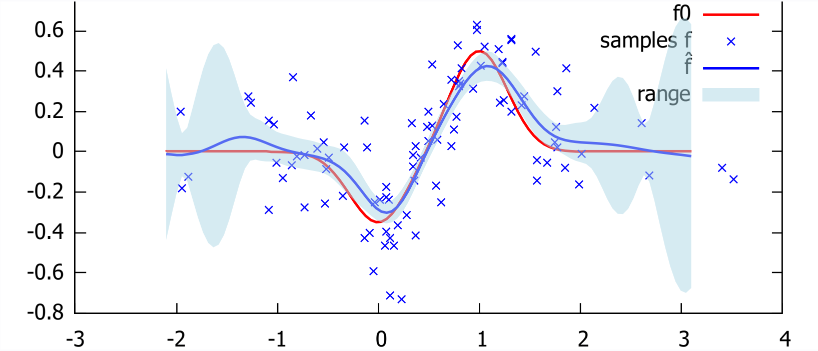
$$\hat{\mu} := k(\hat{X}, X) \left(k(X, X) + \Lambda \right)^{-1} f(X)$$

and

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

Quality of the predictor

Example



The local variance

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

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

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

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

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

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

Remark (Relation to kriging)

Kriging …

- … employs an unknown variogram instead of k,
  - … assumes a radial variogram,
  - … estimates the variogram, or the parameters in a parametric model;
  - typically, the error vanishes, $\Lambda = 0$.
- Design points X_i are known

Outline

Estimator

Problem

For $(X_i, f_i) \in \mathcal{X} \times \mathbb{R} \subset \mathbb{R}^d \times \mathbb{R}$ iid. observations with $X_i \sim P$ (the design measure) we study the estimator

$$\hat{f}_n(\cdot) := \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \, k(\cdot, X_i),$$

where

$$\lambda \, \hat{w}_i + \frac{1}{n} \sum_{j=1}^{n} k(X_i, X_j) \, \hat{w}_j = f_i,$$

$i = 1, \dots, n.$

Now RKHS

Worst case analysis: Generalization (learning) theory, cf. [Steinwart and Christmann, 2008]

Remark (Relation of norms)

$$\|g\|_2 \le \|g\|_\infty \le C_k \cdot \|g\|_k$$

More precisely,

$$\|g\|_2 \le \|K\|^{1/2} \cdot \|g\|_k \quad \text{and} \quad |g(x)| \le \sqrt{k(x,x)} \cdot \|g\|_k.$$

Remark (L^2-norm, $\|\cdot\|_k$ regularization)

Usual results consider the *expected risk*,
$\mathcal{E}(g(\cdot)) := \mathbb{E}\left(f - g(X)\right)^2 = \|f - g(X)\|^2$,

$$P\left(\mathcal{E}(f_z) - \mathcal{E}(f_{z;\mathcal{H}}) > \varepsilon\right) < \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{12M}\right) e^{-\frac{n\varepsilon}{300M^2}},$$

where $|f| < M$ and $\mathcal{N}$ balls, each radius $\frac{\varepsilon}{12M}$, cover $\mathcal{H}$.

Now RKHS

Worst case analysis: Generalization (learning) theory, cf. [Steinwart and Christmann, 2008]

Remark (Relation of norms)

$$\|g\|_2 \le \|g\|_\infty \le C_k \cdot \|g\|_k$$

More precisely,

$$\|g\|_2 \le \|K\|^{1/2} \cdot \|g\|_k \quad \text{and} \quad |g(x)| \le \sqrt{k(x,x)} \cdot \|g\|_k.$$

Remark (L^2-norm, $\|\cdot\|_k$ regularization)

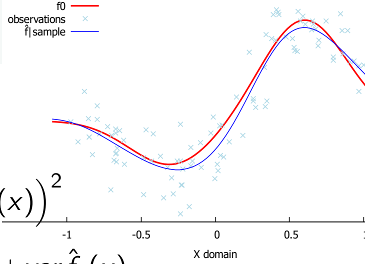
Usual results consider the *expected risk*,
$\mathcal{E}(g(\cdot)) := \mathbb{E}\left(f - g(X)\right)^2 = \|f - g(X)\|^2$,

$$P\left(\mathcal{E}(f_z) - \mathcal{E}(f_{z;\mathcal{H}}) > \varepsilon\right) < \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{12M}\right) e^{-\frac{n\varepsilon}{300M^2}},$$

where $|f| < M$ and $\mathcal{N}$ balls, each radius $\frac{\varepsilon}{12M}$, cover $\mathcal{H}$.

Mean (integrated) squared error

Density estimation, cf. [Tsybakov, 2008]



- Locally, at $x \in \mathcal{X}$,

$$\operatorname{mse} \hat{f}_n(x) := \mathbb{E}\left(\hat{f}_n(x) - f_0(x)\right)^2$$
$$= \left(\operatorname{bias} \hat{f}_n(x)\right)^2 + \operatorname{var} \hat{f}_n(x).$$

- Or globally (L^2 risk function),

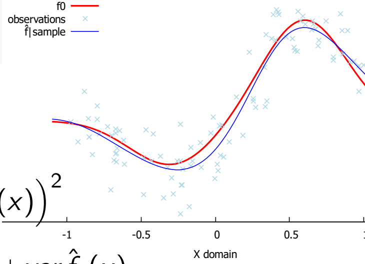
$$\operatorname{mise} \hat{f}_n := \mathbb{E} \int_{\mathbb{R}^d} \left(\hat{f}_n(x) - f_0(x)\right)^2 dx$$
$$\text{or } \int_{\mathbb{R}^d} \operatorname{mse}\left(\hat{f}_n(x)\right) p(x)\, dx = \mathbb{E}\, \|\hat{f}_n(\cdot) - f_0(\cdot)\|_2^2.$$

- For convergence in $(\mathcal{H}_k, \|\cdot\|_k)$ and thus uniform convergence,

$$\mathbb{E}\, \|\hat{f}_n(\cdot) - f_0(\cdot)\|_k^2.$$

Mean (integrated) squared error

Density estimation, cf. [Tsybakov, 2008]



- Locally, at $x \in \mathcal{X}$,

$$\begin{aligned}\operatorname{mse} \hat{f}_n(x) &:= \mathbb{E}\left(\hat{f}_n(x) - f_0(x)\right)^2 \\ &= \left(\operatorname{bias} \hat{f}_n(x)\right)^2 + \operatorname{var} \hat{f}_n(x).\end{aligned}$$

- Or globally (L^2 risk function),

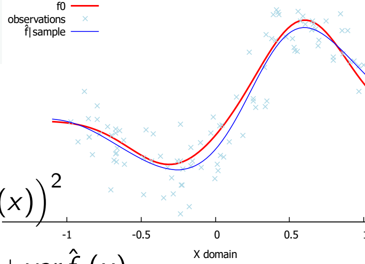
$$\begin{aligned}\operatorname{mise} \hat{f}_n &:= \mathbb{E}\int_{\mathbb{R}^d} \left(\hat{f}_n(x) - f_0(x)\right)^2 dx \\ \text{or } &\int_{\mathbb{R}^d} \operatorname{mse}\left(\hat{f}_n(x)\right) p(x)\, dx = \mathbb{E}\,\|\hat{f}_n(\cdot) - f_0(\cdot)\|_2^2.\end{aligned}$$

- For convergence in $(\mathcal{H}_k, \|\cdot\|_k)$ and thus uniform convergence,

$$\mathbb{E}\,\|\hat{f}_n(\cdot) - f_0(\cdot)\|_k^2.$$

Mean (integrated) squared error

Density estimation, cf. [Tsybakov, 2008]



- Locally, at $x \in \mathcal{X}$,

$$\begin{aligned} \operatorname{mse} \hat f_n(x) &:= \mathbb{E}\left(\hat f_n(x) - f_0(x)\right)^2 \\ &= \left(\operatorname{bias} \hat f_n(x)\right)^2 + \operatorname{var} \hat f_n(x). \end{aligned}$$

- Or globally (L^2 risk function),

$$\begin{aligned} \operatorname{mise} \hat f_n &:= \mathbb{E} \int_{\mathbb{R}^d} \left(\hat f_n(x) - f_0(x)\right)^2 dx \\ \text{or } &\int_{\mathbb{R}^d} \operatorname{mse}\left(\hat f_n(x)\right) p(x)\, dx = \mathbb{E}\, \|\hat f_n(\cdot) - f_0(\cdot)\|_2^2. \end{aligned}$$

- For convergence in $(\mathcal{H}_k, \|\cdot\|_k)$ and thus uniform convergence,

$$\mathbb{E}\, \|\hat f_n(\cdot) - f_0(\cdot)\|_k^2.$$

Theorem (Representer theorem [Schölkopf et al., 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

$$\hat{f}_n(\cdot) := \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \cdot k(\cdot, X_i).$$

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, cf. [Norkin et al., 1998])

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

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

Theorem (Representer theorem [Schölkopf et al., 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

$$\hat{f}_n(\cdot) := \frac{1}{n}\sum_{i=1}^n \hat{w}_i \cdot k(\cdot, X_i).$$

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

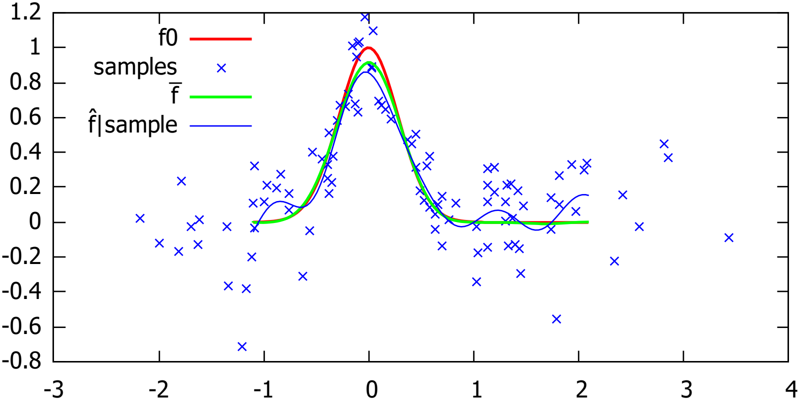
Proposition ($\hat{\vartheta}_n$ is downwards biased, cf. [Norkin et al., 1998])

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

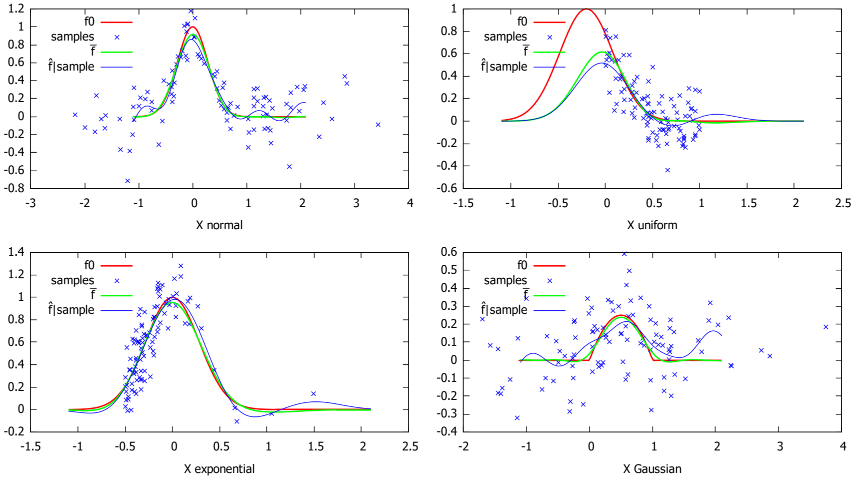
$$\mathbb{E}\,\hat{\vartheta}_n \le \mathbb{E}\,\hat{\vartheta}_{n+1} \le \vartheta^*.$$

The expectation of $\hat{f}_n$

BLU Predictions



Design measure, empirical
BLU Predictions



Remark

Consider the random variable $(X, f) \sim P$ and the problem

$$\vartheta^* := \min_{f_\lambda(\cdot)} \mathbb{E}\left(f - f_\lambda(X)\right)^2 + \lambda \|f_\lambda\|_k^2$$

and note that

$$\vartheta^* = \underbrace{\mathbb{E}\left(f - f_0(X)\right)^2} + \min_{f_\lambda(\cdot)} \mathbb{E}(f_0(X) - f_\lambda(X))^2 + \lambda \|f_\lambda\|_k^2 .$$

By Doob–Dynkin, $f_0(x) = \mathbb{E}(f \mid X = x)$.

Remark

Consider the random variable $(X, f) \sim P$ and the problem

$$\vartheta^* := \min_{f_\lambda(\cdot)} \mathbb{E}\left(f - f_\lambda(X)\right)^2 + \lambda \|f_\lambda\|_k^2$$

and note that

$$\vartheta^* = \underbrace{\mathbb{E}\left(f - f_0(X)\right)^2}_{\text{irreducible}} + \min_{f_\lambda(\cdot)} \mathbb{E}(f_0(X) - f_\lambda(X))^2 + \lambda \|f_\lambda\|_k^2 .$$

By Doob–Dynkin, $f_0(x) = \mathbb{E}(f \mid X = x)$.

Proposition

The solution of

$$\min_{f_\lambda(\cdot)} \mathbb{E}\left(f_0(X) - f_\lambda(X)\right)^2 + \lambda \|f_\lambda\|_k^2$$

is

$$f_\lambda = K\, w_\lambda, \text{ where } (\lambda I + K) w_\lambda = f_0,$$

where

$$K\, w(x) = \int_{\mathcal{X}} k(x,y)\, w(y)\, P(dy).$$

Proposition

It holds that $f_0 - f_\lambda = \lambda\, w_\lambda$ *and*

$$\|f_0 - f_\lambda\|_k \le C_0\, \lambda$$

Now RKHS

The limit: $f_i \longleftrightarrow f_0 \longleftrightarrow f_\lambda$

Proposition

The solution of

$$\min_{f_\lambda(\cdot)} \mathbb{E}\left(f_0(X) - f_\lambda(X)\right)^2 + \lambda \|f_\lambda\|_k^2$$

is

$$f_\lambda = K\, w_\lambda, \quad \text{where } (\lambda I + K) w_\lambda = f_0,$$

where

$$K\, w(x) = \int_{\mathcal{X}} k(x,y)\, w(y)\, P(dy).$$

Proposition

It holds that $f_0 - f_\lambda = \lambda\, w_\lambda$ *and*

$$\|f_0 - f_\lambda\|_k \leq C_0\, \lambda$$

Proposition

The solution of

$$\min_{f_\lambda(\cdot)} \mathbb{E}\left(f_0(X) - f_\lambda(X)\right)^2 + \lambda \|f_\lambda\|_k^2$$

is

$$f_\lambda = K\, w_\lambda, \quad \text{where } (\lambda I + K) w_\lambda = f_0,$$

where

$$K\, w(x) = \int_{\mathcal{X}} k(x,y)\, w(y)\, P(dy).$$

Proposition

It holds that $f_0 - f_\lambda = \lambda\, w_\lambda$ *and*

$$\|f_0 - f_\lambda\|_k \le C_0\, \lambda$$

Remark (Inhomogeneous Fredholm equation of the second kind)

Suppose that

$$\lambda\,\tilde{w}_\lambda(x)+p(x)\cdot\int_{\mathcal{X}}k(x,y)\,\tilde{w}_\lambda(y)\,dy=p(x)\cdot f_0(x),$$

then

$$f_\lambda(x):=\int_{\mathcal{X}}k(x,y)\,\tilde{w}_\lambda(y)\,dy$$

satisfies

$$(\lambda I+K)f_\lambda=K\,f_0.$$

Outline

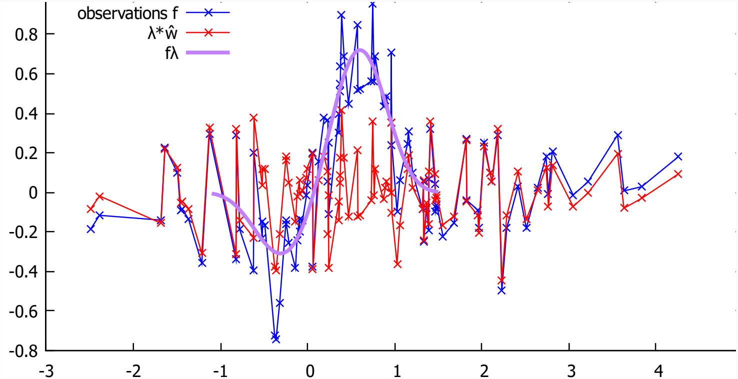
TECHNISCHE UNIVERSITÄT CHEMNITZ

Denoising

Tight relation between noise and weights

Conjecture

The noise f_i and the weights $\hat{w}_i$ are related/ correlated



$$\lambda\,\hat{w}_i + \underbrace{\frac{1}{n}\sum_{j=1}^{n} k(X_i, X_j)\,\hat{w}_j}_{\approx f_\lambda(X_i)} = f_i$$

TECHNISCHE UNIVERSITÄT CHEMNITZ

Denoising: the predictor $\tilde f_n(\cdot)$

$f_i \longleftrightarrow f_0 \longleftrightarrow f_\lambda \longleftrightarrow \tilde f_n \longleftrightarrow \hat f_n$

Definition

With

$$\tilde w_i = \frac{f_i - f_\lambda(X_i)}{\lambda}$$

set

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

Theorem (Unbiased)

Then

$$\operatorname{corr}(f_i, \tilde w_i \,|\, X = x) = 1$$

and, for every $x \in \mathcal{X}$,

$$\mathbb{E}\,\tilde f_n(x) = f_\lambda(x).$$

Denoising: the predictor $\tilde{f}_n(\cdot)$

Definition

With

$$\tilde{w}_i = \frac{f_i - f_\lambda(X_i)}{\lambda}$$

set

$$\tilde{f}_n(\cdot) := \frac{1}{n}\sum_{i=1}^n k(\cdot, X_i)\, \tilde{w}_i.$$

Theorem (Unbiased)

Then

$$\operatorname{corr}(f_i, \tilde{w}_i \,|\, X = x) = 1$$

and, for every $x \in \mathcal{X}$,

$$\mathbb{E}\, \tilde{f}_n(x) = f_\lambda(x).$$

Denoising: the predictor $\tilde f_n(\cdot)$

Definition

With

$$\tilde w_i = \frac{f_i - f_\lambda(X_i)}{\lambda}$$

set

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

Theorem (Unbiased)

Then

$$\operatorname{corr}\left(f_i, \tilde w_i \middle| X = x\right) = 1$$

and, for every $x \in \mathcal{X}$,

$$\mathbb{E}\,\tilde f_n(x) = f_\lambda(x).$$

$$\begin{aligned}
\mathbb{E}\,\frac{1}{n}\sum_{i=1}^n k(x, X_i)\,\tilde w_i &= \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n k(x, X_i)\,\mathbb{E}\left(\frac{f_i - f_\lambda(X_i)}{\lambda}\middle| X_i\right)\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,k(x, X_i)\,\frac{f_0(X_i) - f_\lambda(X_i)}{\lambda}\\
&= \mathbb{E}\,k(x, X_i)\,w_\lambda(X_i) = K\,w_\lambda(x) = f_\lambda(x)
\end{aligned}$$

Theorem (Consistency for heteroscedastic data)

Further,

$$\mathbb{E}\left\|f_\lambda(\cdot) - \tilde{f}_n(\cdot)\right\|_k^2 = \frac{C}{n},$$

where

$$C := \frac{1}{\lambda^2}\int_{\mathcal{X}} \Big(\big(\underbrace{f_0(x) - f_\lambda(x)}\big)^2 + \operatorname{var}(f\,|\,x)\Big) k(x,x) P(dx) - \|f_\lambda\|_k^2 .$$

Here, the data are possibly heteroscedastic,

$$\operatorname{var}(f\,|\,x) = \mathbb{E}\Big(\big(f - f_0(X)\big)^2 \,\big|\, X = x\Big).$$

Theorem (Consistency for heteroscedastic data)

Further,

$$\mathbb{E}\left\| f_\lambda(\cdot) - \tilde{f}_n(\cdot) \right\|_k^2 = \frac{C}{n},$$

where

$$C := \frac{1}{\lambda^2} \int_{\mathcal{X}} \Big(\big(\underbrace{f_0(x) - f_\lambda(x)}_{bias} \big)^2 + \mathrm{var}(f\,|\,x) \Big) k(x,x) P(dx) - \|f_\lambda\|_k^2 .$$

Here, the data are possibly heteroscedastic,

$$\mathrm{var}(f\,|\,x) = \mathbb{E}\Big(\big(f - f_0(X)\big)^2 \,\big|\, X = x \Big).$$

Proposition (Consistency)

- $$\tilde{f}_n(\cdot) - \hat{f}_n(\cdot) = \frac{1}{n}\sum_{j=1}^n \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)_j^{-1} k(\cdot, X_j),$$
- $$\|\tilde{f}_n(\cdot) - \hat{f}_n(\cdot)\|_k^2 = \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)^{-1} \frac{1}{n}K \left(\lambda + \frac{1}{n}K\right)^{-1} \tilde{r}_n,$$

where $\tilde{r}_n = \left(\tilde{f}_n(X_i) - \hat{f}_n(X_i)\right)_{i=1}^n$.

Now RKHS

$f_i \longleftrightarrow f_0 \longleftrightarrow f_\lambda \longleftrightarrow \tilde{f}_n \longleftrightarrow \hat{f}_n$

Proposition (Consistency)

- $$\tilde{f}_n(\cdot) - \hat{f}_n(\cdot) = \frac{1}{n}\sum_{j=1}^{n} \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)_j^{-1} k(\cdot, X_j),$$
- $$\|\tilde{f}_n(\cdot) - \hat{f}_n(\cdot)\|_k^2 = \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)^{-1} \frac{1}{n}K \left(\lambda + \frac{1}{n}K\right)^{-1} \tilde{r}_n,$$

where $\tilde{r}_n = \left(\tilde{f}_n(X_i) - \hat{f}_n(X_i)\right)_{i=1}^{n}$.

Theorem

$$\mathbb{E}\,\|\tilde{f}_n - \hat{f}_n\|_k^2 \le \frac{C_3}{\lambda^3 n},$$

in some cases even $\mathbb{E}\,\|\tilde{f}_n - \hat{f}_n\|_k^2 \le \frac{C_3}{\lambda^2 n}$.

Proof.

$$\left(\lambda + \tfrac{1}{n}K\right)^{-1} \tfrac{1}{n}K \left(\lambda + \tfrac{1}{n}K\right)^{-1} \le \frac{1}{4\lambda}. \qquad \square$$

Proposition (Consistency)

- $$\tilde{f}_n(\cdot) - \hat{f}_n(\cdot) = \frac{1}{n}\sum_{j=1}^n \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)_j^{-1} k(\cdot, X_j),$$
- $$\|\tilde{f}_n(\cdot) - \hat{f}_n(\cdot)\|_k^2 = \tilde{r}_n^\top \left(\lambda + \frac{1}{n}K\right)^{-1} \frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1} \tilde{r}_n,$$

where $\tilde{r}_n = \left(\tilde{f}_n(X_i) - \hat{f}_n(X_i)\right)_{i=1}^n$.

Theorem

$$\mathbb{E}\,\|\tilde{f}_n - \hat{f}_n\|_k^2 \leq \frac{C_3}{\lambda^3 n},$$

in some cases even $\mathbb{E}\,\|\tilde{f}_n - \hat{f}_n\|_k^2 \leq \frac{C_3}{\lambda^2 n}$.

Proof.

$$\left(\lambda + \tfrac{1}{n}K\right)^{-1} \tfrac{1}{n}K\left(\lambda + \tfrac{1}{n}K\right)^{-1} \leq \frac{1}{4\lambda}.$$ □

Outline

Order of convergence

$f_i \longleftrightarrow f_0 \longleftrightarrow f_\lambda \longleftrightarrow \tilde{f}_n \longleftrightarrow \hat{f}_n$

$\|f_i - f_0\|$	irreducible
$\|f_0 - f_\lambda\|_k^2$	$\leq C_0\,\lambda^2$
$\mathbb{E}\,\|f_\lambda - \tilde{f}_n\|_k^2$	$\leq \frac{C_1}{\lambda^2 n}$
$\mathbb{E}\,\|\tilde{f}_n - \hat{f}_n\|_k^2$	$\leq \frac{C_2}{\lambda^3 n}$, $\leq \frac{C_2}{\lambda^2 n}$

Theorem (Unbiased)

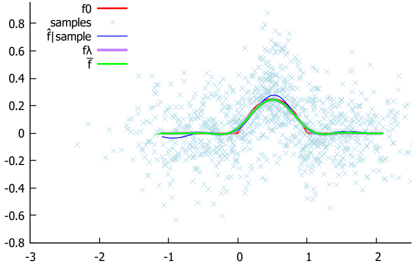
If $\lambda_n = \mathcal{O}\left(n^{-1/5}\right)$, *then*

$$\mathbb{E}\,\|f_0(\cdot) - \hat{f}_n(\cdot)\|_k^2 = \mathcal{O}\left(n^{-2/5}\right).$$

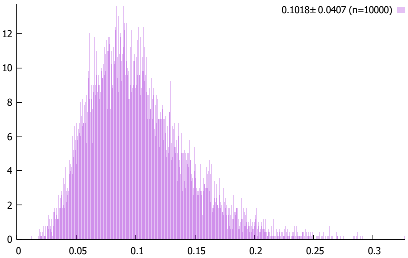
For the best constant, an oracle is needed.

Precision analysis: $f_0 \notin \mathcal{H}_k$

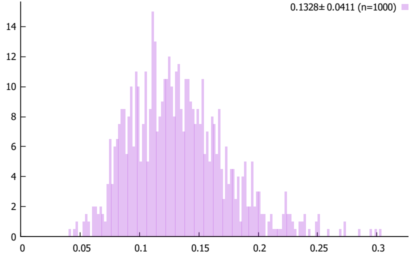
Histogram of $n\lambda \| f_\lambda(\cdot) - \hat{f}_n(\cdot) \|_k^2$



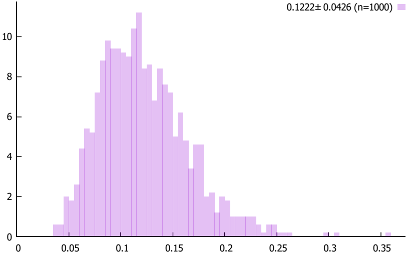
$n = 1000, \lambda = 0.01$



$n = 10000, \lambda = 0.1$



$n = 1000, \lambda = 0.01$

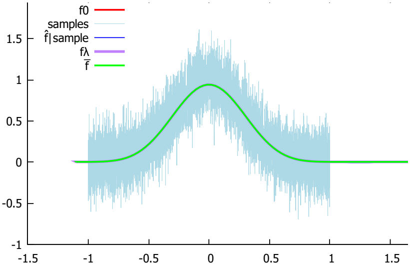


$n = 1000, \lambda = 0.001$

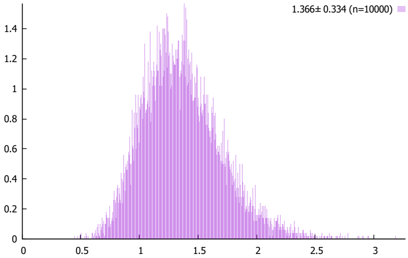
TECHNISCHE UNIVERSITÄT CHEMNITZ

Precision analysis (cont): $f_0 \in \mathcal{H}_k$

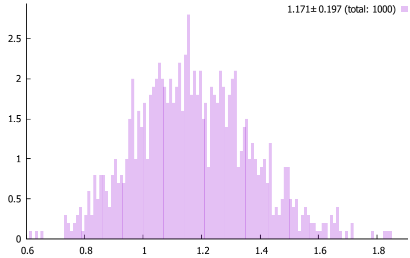
Histogram of $\sqrt{n}\,\|f_0(\cdot) - \hat{f}_{\lambda_n}(\cdot)\|_k^2$ for $\lambda_n = n^{-1/2}$



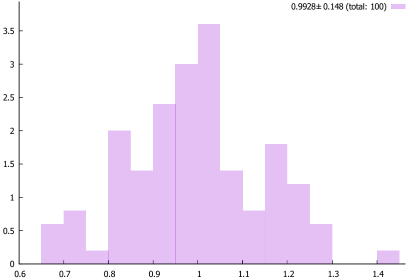
$n = 1000$



$n = 100$



$n = 1000$



$n = 10\,000$

Proposition (Weak consistency)

For $\varepsilon > 0$ it holds that $f_0(x) = \mathbb{E}[f \mid X = x]$

$$P\left(\|f_\lambda - \hat{f}_n\|_k \geq \varepsilon\right) \to 0$$

as $n \to \infty$ (convergence in probability).

Proposition

Consistency of $\hat{\vartheta}_n$: it holds that

$$P\left(|\vartheta^* - \hat{\vartheta}_n| \geq \varepsilon\right) \to 0$$

as $n \to \infty$.

Quantile estimation employs the loss function

$$\ell_\alpha(y) := \begin{cases} -(1-\alpha)\, y & \text{if } y \le 0,\\ \alpha \cdot y & \text{if } y \ge 0. \end{cases}$$

The expectile

$$e_\alpha(X) := \operatorname{argmin}_{x\in\mathbb{R}} \mathbb{E}\,\ell_\alpha(X-x),$$

the only *elicitable* risk functional, which is coherent – employs the loss function

$$\ell_\alpha(y) := \begin{cases} -(1-\alpha)\, y^2 & \text{if } y \le 0,\\ \alpha \cdot y^2 & \text{if } y \ge 0. \end{cases}$$

The conditional expectile is

$$e_\alpha(x) := \operatorname{argmin}_{f_\lambda(\cdot)} \mathbb{E}\,\ell_\alpha\big(f - f_\lambda(X)\big) + \lambda \|f_\lambda(\cdot)\|_k^2,$$

with discretized version

$$\hat{e}_\alpha(x) := \operatorname{argmin}_{f_\lambda(\cdot)} \frac{1}{n}\sum_{i=1}^n \ell_\alpha\big(f_i - f_\lambda(X_i)\big) + \lambda \|f_\lambda(\cdot)\|_k^2.$$

Quantile estimation employs the loss function

$$\ell_\alpha(y) := \begin{cases} -(1-\alpha)\,y & \text{if } y \le 0,\\ \alpha \cdot y & \text{if } y \ge 0. \end{cases}$$

The expectile

$$e_\alpha(X) := \operatorname{argmin}_{x\in\mathbb{R}} \mathbb{E}\,\ell_\alpha(X-x),$$

the only *elicitable* risk functional, which is coherent – employs the loss function

$$\ell_\alpha(y) := \begin{cases} -(1-\alpha)\,y^2 & \text{if } y \le 0,\\ \alpha \cdot y^2 & \text{if } y \ge 0. \end{cases}$$

The conditional expectile is

$$e_\alpha(x) := \operatorname{argmin}_{f_\lambda(\cdot)} \mathbb{E}\,\ell_\alpha\big(f - f_\lambda(X)\big) + \lambda\,\|f_\lambda(\cdot)\|_k^2,$$

with discretized version

$$\hat{e}_\alpha(x) := \operatorname{argmin}_{f_\lambda(\cdot)} \frac{1}{n}\sum_{i=1}^n \ell_\alpha\big(f_i - f_\lambda(X_i)\big) + \lambda\,\|f_\lambda(\cdot)\|_k^2.$$

Conditional improvements

Eigenvalues

Theorem

Assume the spectrum of the matrix K decays exponentially, i.e., there are constants α and β such that

$$\sigma_i \le \alpha\, e^{-\beta\, i}.$$

Then

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} w_i^N k(\cdot, X_i)\right\|_k^2 \le \sigma_{\max}^2 c_1 \frac{\log n}{p\, n\, \lambda} + c_2 \frac{\sigma_{\max}^2}{\lambda^2\, n^{\frac{1}{p}+1}}$$

holds for all $p \ge 1$. Moreover, for $\lambda_n = \frac{c}{\sqrt{n}}$ it holds that

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} w_i^N k(\cdot, X_i)\right\|_k^2 \le \sigma_{\max}^2 c_1 \frac{\log n}{\sqrt{n}} + c_2 \frac{\sigma_{\max}^2}{\sqrt{n}}$$

with the constants $c_1 = \frac{1}{4\beta c}$ and $c_2 = \frac{\alpha}{2c^2(1-e^{-\beta})}$.

Remarks and follow-up questions

Invitation for future work

- The results do *not* depend on the dimension.
- Risk: the expectile is an M-estimator and consistent with this type of optimization,
- cf. [Dentcheva and Lin, 2021]
- Further implications on machine learning: different loss functions ℓ
- Bandwidth selection
- What is the limiting distribution of $n \cdot \|f_n(\cdot) - f_\lambda(\cdot)\|^2$
- Correct order of convergence in special cases
- Implications on the stochastic optimization problem

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

  for smooth functions
- Implications on multistage programs and HJB
- time series analysis, machine learning: predict X_{t+1}, given the past observations $X_t, \dots, X_{t-\ell}$.
- ANOVA

Observations (X_i, f_i), where $f_i = f_0(X_{i1}, \dots, X_{id}) + \varepsilon$:

$$f_0(x_1, \dots, x_d) + \varepsilon = \underbrace{\mathbb{E} f}_{\hat{f}_0 \in \mathbb{R}} + \varepsilon_0$$

Observations (X_i, f_i), where $f_i = f_0(X_{i1}, \dots, X_{id}) + \varepsilon$:

$$f_0(x_1, \dots, x_d) + \varepsilon = \underbrace{\mathbb{E} f}_{\hat f_0 \in \mathbb{R}} + \underbrace{\mathbb{E}(f \mid X_1 = x_1)}_{\hat f_1(x_1)} + \dots + \underbrace{\mathbb{E}(f \mid X_d = x_d)}_{\hat f_d(x_d)} + \varepsilon_1$$

Here, $\hat f_i(\cdot) = \sum_{\ell=1}^n \hat w_i \, k(\cdot, X_i)$, where

$$\lambda \, \hat w_i + \sum_{\ell=1}^n k(X_i, X_\ell) \, \hat w_\ell = f_i - \sum_{j<i} \hat f_j(X_i).$$

Observations (X_i, f_i), where $f_i = f_0(X_{i1}, \dots, X_{id}) + \varepsilon$:

$$
\begin{aligned}
f_0(x_1, \dots, x_d) + \varepsilon = & \underbrace{\mathbb{E} f}_{\hat{f}_0 \in \mathbb{R}} \\
& + \underbrace{\mathbb{E}(f \mid X_1 = x_1)}_{\hat{f}_1(x_1)} + \dots + \underbrace{\mathbb{E}(f \mid X_d = x_d)}_{\hat{f}_d(x_d)} \\
& + \sum_{i<j} \underbrace{\mathbb{E}(f \mid X_i = x_i, X_j = x_j)}_{\hat{f}_{ij}(x_i, x_j)} + \varepsilon_2
\end{aligned}
$$

Here, $\hat{f}_i(\cdot) = \sum_{\ell=1}^{n} \hat{w}_i \, k(\cdot, X_i)$, where

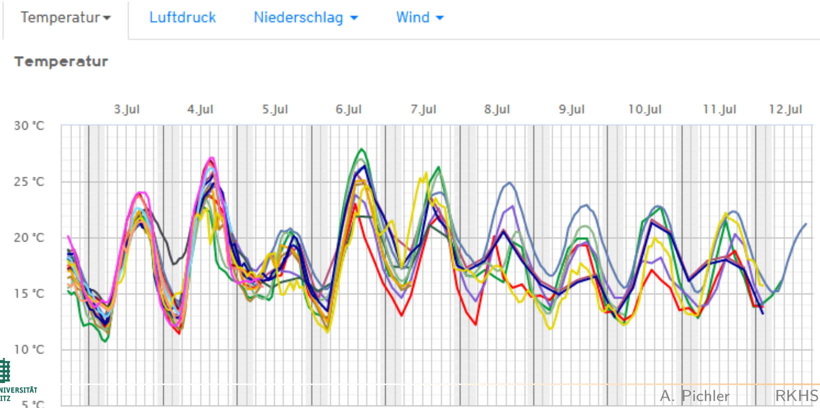
$$
\lambda \, \hat{w}_i + \sum_{\ell=1}^{n} k(X_i, X_\ell) \, \hat{w}_\ell = f_i - \sum_{j<i} \hat{f}_j(X_i).
$$

Observations

$$X_0,\dots,X_{t-\ell-1},\underbrace{X_{t-\ell},\dots X_t,X_{t+1}},X_{t+2},\dots,X_n,$$

where

$$X_{t+1}=f(X_{t-\ell},\dots,X_t)+\varepsilon.$$

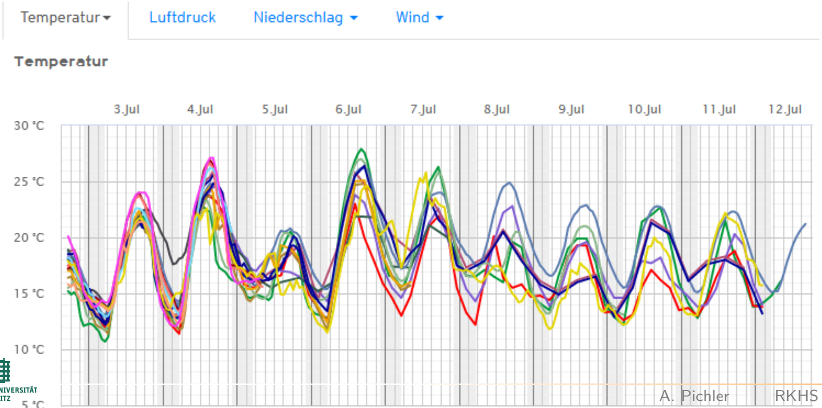


Observations

$$X_0, \dots, X_{t-\ell-1}, \underbrace{X_{t-\ell}, \dots X_t, X_{t+1}}, X_{t+2}, \dots, X_n,$$

where

$$X_{t+1} = f(X_{t-\ell}, \dots, X_t) + \varepsilon.$$



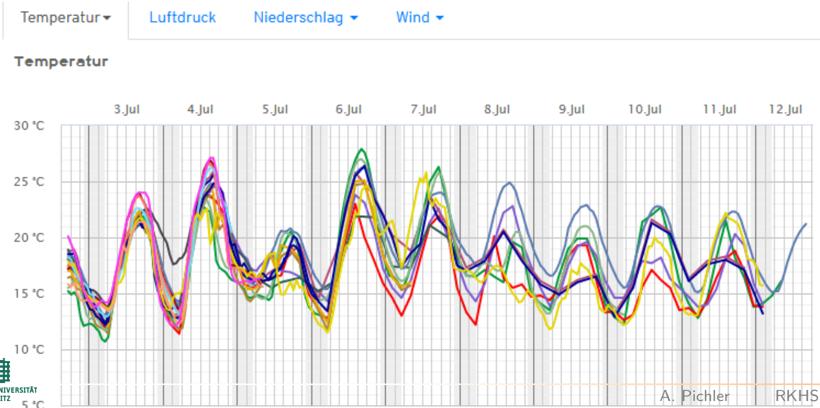
Nonlinear time series
Predictions based on temporal lag ℓ

Observations

$$X_0, \dots, X_{t-\ell-1}, \underbrace{X_{t-\ell}, \dots X_t, X_{t+1}}_{\text{training}}, X_{t+2}, \dots, X_n,$$

where

$$X_{t+1} = f(X_{t-\ell}, \dots, X_t) + \varepsilon.$$



Nonlinear time series
Predictions based on temporal lag ℓ

Observations

$$X_0, \dots, X_{t-\ell-1}, \underbrace{X_{t-\ell}, \dots X_t, X_{t+1}}_{\text{training}}, X_{t+2}, \dots, X_n,$$

where

$$X_{t+1} = f(X_{t-\ell}, \dots, X_t) + \varepsilon.$$

Here,

$$\hat{f}(x_{-\ell}, \dots, x_0) = \sum_{t=1}^{n} \hat{w}_t\, k\Big((x_{-\ell}, \dots, x_0), (X_{t-\ell}, \dots, X_t)\Big),$$

where

$$\lambda\, \hat{w}_t + \sum_{j=1}^{n} k\Big((X_{t-\ell}, \dots, X_t), (X_{j-\ell}, \dots, X_j)\, \hat{w}_j = X_{t+1}.$$

Thank you!
References

Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer-Verlag New York Inc.

Dentcheva, D. and Lin, Y. (2021).
Bias reduction in sample-based optimization.
SIAM Journal on Optimization, 32(1):130–151.

Dommel, P. and Pichler, A. (2021).
Uniform function estimators in reproducing kernel Hilbert spaces.

Norkin, V. I., Pflug, G. Ch., and Ruszczyński, A. (1998).
A branch and bound method for stochastic global optimization.
Mathematical Programming, 83(1-3):425–450.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In *Lecture Notes in Computer Science*, pages 416–426.
Springer Berlin Heidelberg.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer New York.

Tsybakov, A. B. (2008).
Introduction to Nonparametric Estimation.
Springer, New York.